Replik til Svend Kreiners notat "Om statistiske analyser af resultater fra pædagogiske test"

August 2023

Af Simon Calmar Andersen, Peter Fallesen, Eskil Heinesen, Mathilde Lund Holm, Maria Humlum, Rasmus Landersø, Anne Brink Nandrup og Helena Skyt Nielsen

Svend Kreiner (SK) beskriver i sit notat "Om statistiske analyser af resultater fra pædagogiske test", hvad han fejlagtigt opfatter som problemer i fire artikler fra hhv. Aarhus Universitet og ROCKWOOL Fonden.

SKs kritik bygger på mangelfuld læsning af de studier, han kritiserer, og en sammenblanding af, hvilke udfordringer testene kan have, når det kommer til at måle den enkelte elevs udvikling fra ét klassetrin til det næste, og hvilke udfordringer der kan være, når man sammenholder to grupper i eksempelvis et lodtrækningsforsøg (som i et af de kritiserede studier).

Der bliver allerede taget højde for samtlige af hans kritikpunkter, og dette fremgår ved læsning af studierne. Omhyggelig analyse af data er essentielt for at nå frem til retvisende resultater. Det er grundlaget for arbejdet både på TrygFondens Børneforskningscenter på Aarhus Universitet, VIVE og i ROCKWOOL Fonden, og vores studier er derfor også baseret på de bedste metoder til at analysere de emner, vi arbejder med.

Vi byder debat af forskningsmetode og målemetoder velkommen. Det er et vigtigt grundlag for forskning, at den efterprøves af fagfæller. Og vi indbyder gerne SK til dialog om de undersøgelser, vi arbejder med. Vi er dog uforstående over, hvorfor SK – på baggrund af en mangelfuld og misforstået kritik – vælger at drage tvivl om vigtige forskningsresultater omkring børns udvikling og trivsel; fx at to-lærer-ordninger i klasser specielt hjælper børn med faglige udfordringer, og at forældres skilsmisse har negative konsekvenser for børns indlæring i skolen.

I teksten nedenfor gennemgår vi SKs kritikpunkter og redegør for, hvorfor de ikke giver anledning til bekymring for validiteten af analyserne.

Simon Calmar Andersen, Peter Fallesen, Eskil Heinesen, Mathilde Lund Holm, Maria Humlum, Rasmus Landersø, Anne Brink Nandrup og Helena Skyt Nielsen

SKs notat berører flere emner. De kan placeres i fire kategorier.

1. Problemer med brug af de nationale tests til kausale analyser.
2. Problemer med brug af lineære regressionsmodeller.
3. Transformation af de nationale testresultater fra en skala til en anden.
4. Brug af et samlet gennemsnit på tværs af testenes tre profilområder.

1. Problemer med kausale analyser

En kausal analyse henviser til en undersøgelse af, om en ting påvirker en anden. Det kunne være, om antallet af lærere i klassen påvirker elevernes læsefærdigheder.

Hvis man ser på klasser i Danmark generelt, er det ikke tilfældigt, hvor der er mere end én lærer i klassen. Flere lærere (eller andre voksne) vil ofte skyldes, at ét eller flere børn i klassen har brug for ekstra støtte. Derfor kan man ikke uden videre konkludere, at en sammenhæng mellem antallet af lærere og elevernes læsefærdigheder skyldes antallet af lærere.

For at finde den kausale sammenhæng mellem antallet af lærere og elevernes læsefærdigheder kan man fx udføre et forsøg, hvor nogle klasser – ved lodtrækning – tildeles en ekstra lærer. Det er netop, hvad der er blevet undersøgt i et af vores studier: En kontrolgruppe beholdt én lærer som hidtil, og en interventionsgruppe blev tildelt en ekstra lærer.

Målefejl i test af læsefærdigheder kan have to udfordringer for en analyse som denne. Den første er, hvis der er en systematisk sammenhæng mellem målefejlen, og hvilke klasser der ved lodtrækning blev tildelt to lærere. Studiet viser imidlertid, at der ikke er nogen sammenhænge mellem, hvem der får en ekstra lærer, og andre bagvedliggende faktorer, netop fordi det var tilfældigt, hvem der fik en ekstra lærer. Således forventes, at målefejl også er ens fordelt i kontrol- og indsatsgrupperne, og eventuelle målefejl vil dermed ikke påvirke vores resultater.

Selvom de øvrige studier undersøger andre spørgsmål, er ræsonnementet for to af de andre studier, SK diskuterer, det samme som beskrevet ovenfor. Studierne påviser, at der ikke er nogen systematisk sammenhæng mellem, hvem der er i interventionsgruppen og i kontrolgruppen. Derfor er SKs kritik på dette punkt ubegrundet.

Den anden udfordring ved målefejl er, at resultaterne kan blive så upræcise, at resultaterne ikke er tilstrækkeligt statistisk sikre. Dette er ikke tilfældet i de pågældende studier. Det kan eksempelvis med stor sikkerhed konkluderes, at en ekstra lærer har en gavnlig effekt på elevernes læsefærdigheder. Derfor er SKs kritik også her ubegrundet.

2. Problemer med lineære regressionsmodeller

Dette punkt indeholder to kritikpunkter, som vi besvarer hver for sig nedenfor.

2.a. Lineær regression og linearitet mellem variabler

SK skriver:

> *"Den lineære regressionsmodel bygger på to fundamentale forudsætninger. For det første at alle sammenhænge mellem læsefærdigheden Y og baggrunds-variablene er lineære."*

I tre af artiklerne undersøger vi effekten af en "intervention" (fx en ekstra lærer), som har påvirket én gruppe og ikke en anden. Analyserne undersøger, om der er statistisk signifikant forskel mellem de to gruppers gennemsnitsscorer. Der er ikke særlige krav om linearitet, når man skal sammenligne to grupper på denne måde, og man vil få samme resultat, hvis man brugte andre statistiske modeller end en lineær regression (hvilket en af artiklerne faktisk også viser). Derfor er SKs kritik ubegrundet.

2.b. Varianshomogenitet

SK skriver videre om lineære regressionsmodellers forudsætninger:

> *"Og for det andet, at fejlledet E er normalfordelt med middelværdi 0 og den samme varians σ^2 for alle personer uanset værdierne af baggrundsvariablene".*

Dette tager alle de kritiserede studier højde for ved at benytte såkaldt robuste og hvor relevant klynge-korrigerede standardfejl. Det har været standard i forskningen i mange år, og det fremgår ved læsning af studierne. Derfor er SKs kritik af vores studier på denne baggrund ubegrundet.

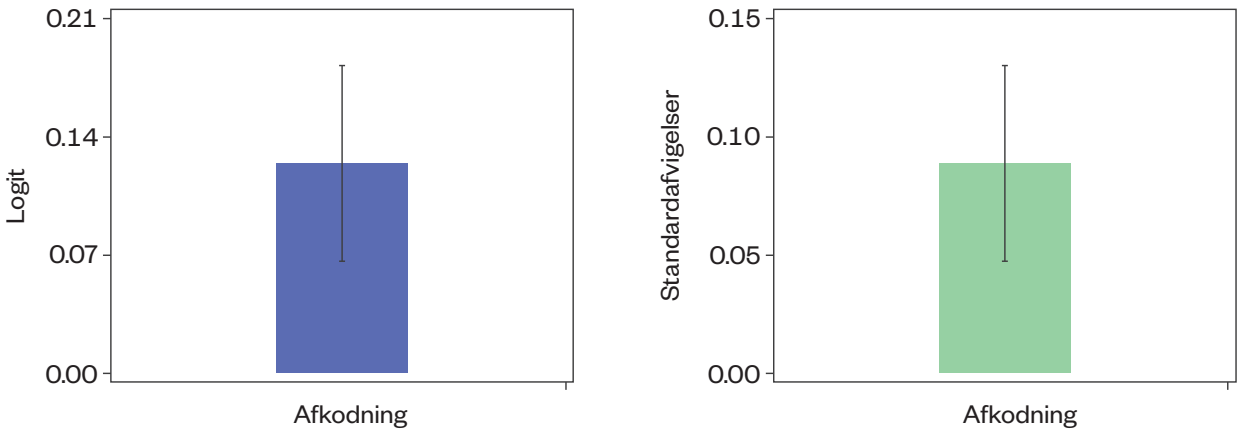
3. Transformation af DNT-resultaterne

Artiklerne omregner testscorerne fra de nationale test fra en Rasch-skala til en standardiseret score, som har standardafvigelse 1 og gennemsnit på 0. SK skriver, at:

> *"Transformationen af DNT-resultaterne som gør det umuligt at tolke analyseresultaterne på en måde, der fortæller noget konkret om de færdigheder, som DNT forsøger at måle. (...) Hvis man havde været så venlige at oplyse om de middelværdier og standardafvigelser, som de benyttede til omregningerne, havde vi kunnet regne tilbage til resultater på logit niveau".*

Den transformation mellem skalaerne, som SK henviser til, ændrer ikke på artiklernes konklusioner i forhold til, om effekterne af de interventioner, artiklerne undersøger, har en statistisk signifikant effekt på elevernes nationale testresultater. Resultaterne rapporteres på en anden måde, end SK foretrækker, da den valgte skala gør det lettere at sammenligne de målte effekter med andre studier, som også undersøger lignende spørgsmål.

Figur 1. Det samme resultat vist på standardiseret skala (grøn søjle) og på logit-skalaen (blå søjle): De lodrette sorte linjer viser den usikkerhed, der er på beregningen (såkaldte 95%-konfidensintervaller).



Figur 1 viser et eksempel fra én af de artikler, SK kritiserer. Den grønne figur viser effekten af et nedbrud i de nationale test på elevernes færdigheder i afkodning målt på den standardiserede skala. Det er det resultat, vi har publiceret i artiklen (her illustreret grafisk). I den blå søjle har vi regnet resultatet tilbage til logit-skalaen. Tallene er forskellige, men det er den samme effekt og den samme usikkerhed (de sorte linjer), uanset om det opgøres på den ene eller anden skala.

4. Profilområder og deres gennemsnit

SK skriver:

> *"Det tredje problem er, at man i artiklerne erstatter målinger af DNTs tre såkaldte profilområder med vægtede gennemsnitsværdier, uden belæg for at disse værdier kan betragtes som valide mål for det, som DNT forsøgte at måle."*

Der er ikke nogen fejl i at præsentere resultater som et samlet gennemsnit inden for et område – det er endog en ofte anvendt metode til at reducere antallet af resultater, der rapporteres samtidigt – men SK kritiserer resultater baseret på det samlede gennemsnit for at være uinteressante. Det vil vi lade læserne af de pågældende studier vurdere.

Vi ser som regel både på effekter på hvert profilområde for sig og på et gennemsnit af hver af de tre profilområder, der vægter resultaterne fra hvert af de tre profilområder lige højt. En fordel ved at se på det samlede gennemsnit er, at noget af den tilfældige usikkerhed, der er ved måling af et enkelt profilområde, bliver reduceret, hvis man ser på et samlet mål. I studiet, hvor vi undersøgte effekten af en ekstra lærer, havde vi ikke hypoteser om, at den ekstra lærer skulle være særlig god for afkodning eller tekstforståelse, og foretrak derfor at anvende gennemsnittet.

I et andet af de studier, som SK kritiserer for ikke at præsentere resultater opdelt på profilområder, nemlig "The Positive Effects of Nationwide Testing on Student Achievement in a Low-Stakes System"[1], vil det fremgå ved læsning af studiet, at resultaterne faktisk også præsenteres opdelt på profilområder i studiets appendiks.

1 Arbejdspapiret er senere publiceret som "Learning from Performance Information", Journal of Public Administration Research and Theory, 2020, pp. 415-431. Også i appendiks til den publicerede artikel viser vi resultater opdelt på profilområder.

Variable med usikkerhed og fejl og de genberegnede nationale test

Udover de fire ovenstående problemstillinger vil vi afslutningsvist adressere en anden problemstilling, som SK også berører i sit notat, nemlig at der er usikkerheder i målingen af de nationale test (som i alle test) og muligvis også systematiske fejl.

Det har været anført af SK og andre, at der var systematiske fejl i beregningen af elevernes testscores – specielt for elever, der klarede sig godt i testene. Undervisningsministeriet har derfor for nylig anvendt en ny måde at beregne elevernes testresultater på. SK skriver:

> *"Hvor vidt de [analyser baseret på de oprindelige testscores] er misvisende i et omfang, der giver problemer for de endelige konklusioner, kan jeg naturligvis ikke sige noget om her, men jeg vil forvente, at de forskere, der har foretaget analyserne, er i stand til at forsvare deres konklusioner med andre argumenter end, at de tror, at fejlene ikke betyder noget, fordi der er en høj grad af korrelation mellem de sande og fejlbehæftede DNT-resultater".*

Til det er der flere ting at bemærke. For det første så er det faktisk meget relevant, at korrelationen mellem de nye og de gamle beregnede tests er meget høj (for eksempel 97,1% i et af vores datasæt, hvor vi har mulighed for at sammenligne de nye og gamle testresultater). For langt de fleste er elevernes resultater tæt på identiske, uanset om man bruger den nye eller gamle beregning.

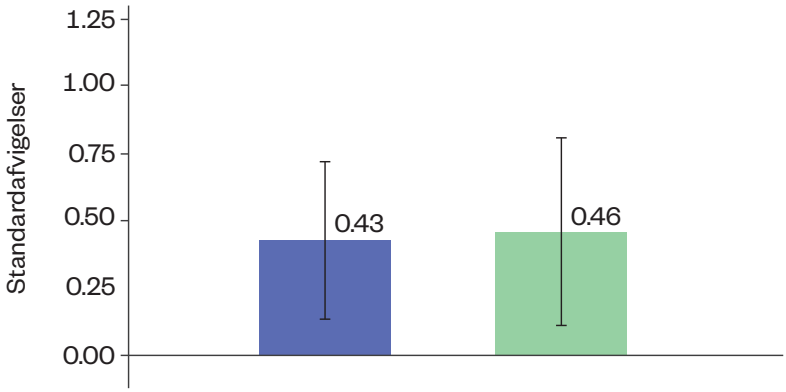
For det andet har de små forskelle mellem de nye og gamle beregninger ikke nogen sammenhæng med de interventioner, som vi undersøger i artiklerne. Som beskrevet ovenfor vil det kun være tilfældet, hvis der er en sammenhæng mellem forskelle i beregning, og hvordan kontrolgruppe og interventionsgruppe opdeles – fx ved lodtrækning. Der er ikke nogen grund til at tro, at forskellen i de to beregningsmetoder skulle være større for skoler, som tilfældigt fik tildelt ekstra lærere, end dem som tilfældigvis ikke fik det. Denne lodtrækning har ikke noget med den nye og den gamle beregningsmetode at gøre.

For det tredje har vi i et nyere studie, som er baseret på data fra efter 2014, mulighed for at sammenligne vores resultater baseret på de gamle beregninger med resultater baseret på de nye beregninger, fordi ministeriet har genberegnet testscores tilbage til 2014. Resultatet kan ses i figur 2. Den grønne søjle viser det resultat, vi har publiceret i artiklen.[2] Den blå søjle viser resultatet for den gennemsnitlige læsetestscore baseret på de nye beregninger. Figuren viser, at vi estimerede effekten til 0,46 standardafvigelser med de gamle beregninger og 0,43 med de nye beregninger. Denne minimale forskel på 0,03 standardafvigelser ligger klart inden for den usikkerhed (det såkaldte 95%-konfidensinterval), som er markeret med de lodrette streger.

Denne direkte sammenligning af de nye og de gamle beregnede testscores bekræfter således også, at eventuelle systematiske fejl ikke har reel betydning for resultater og konklusioner i de typer undersøgelser, vi foretager.

2 Andersen, S. C., Guul, T. S., & Humlum, M. K. (2022). How first-language instruction transfers to majority-language skills. Nature Human Behavior, 6(2), 229–235.

Figur 2. Effekt af modersmålsundervisning målt med de nyberegnede nationale testscores (blå søjle) og gamle beregnede testscores (grøn søjle). Kilde til resultatet baseret på de gamle beregnede test-scores: Andersen, S. C., Guul, T. S., & Humlum, M. K. (2022). "How first-language instruction transfers to majority-language skills". Nature Human Behavior, 6(2), 229–235.



Konklusion

SKs kritik giver ikke anledning til at revidere resultater eller konklusioner fra de fire studier, der kritiseres.

Kritikken er dels baseret på mangelfuld læsning af studierne, og dels baseret på mangelfuld forståelse eller sammenblanding af, hvilke forhold der skal til for at måle enkelte elevers faglige udvikling fra ét år til et andet, og hvilke forhold der skal til for at undersøge effekterne af et forsøg som to-lærer-ordningen eller en begivenhed som skilsmisse.

SKs kritik af tests fra tidligere år er ydermere ikke relevante. Alle konklusioner er uændrede, når de opdaterede testresultater anvendes.