Om statistiske analyser af resultater fra pædagogiske test

1. Indledning

Dette notat knytter kommentarer til den sejlivede diskussion af mulighederne for at anvende testresultater fra de såkaldte nationale test (DNT) inklusiv de problemer, som fejlberegningerne af elevernes dygtighed kan have haft og fortsat vil kunne have, hvis undervisningsministeriet ikke fjerner alle de fejlbehæftede resultater.

1.2 Om fejl i DNT

Dokumentationen af fejl i DNT blev første gang af publiceret af Bundesen & Kreiner (2019), hvor vi påviste, at de estimater af itemparametre og sværhedsgrader i læsning i 8. klasse, som DNT benytter når den adaptive algoritme vælger opgaver til eleverne, afveg signifikant fra estimaterne af sværhedsgraderne estimeret i data fra de obligatoriske test i 2017. Disse fejl havde to konsekvenser fordi DNT var en adaptiv test. For det første, at DNTs målinger af læsefærdighederne i 8. klasser vil være behæftet med en større grad af usystematiske fejl end forventet. Og for det andet – og måske endnu værre – at der kunne være tale om *systematiske* fejl i DNTs målingerne af læsefærdighed.

Konklusionerne i rapporten stillede naturligvis en række spørgsmål. Først og fremmest om det kun var et problem for test af læsning i 8. klasse og for det andet, om det var et aktuelt problem, der havde eksisteret i flere år. Disse spørgsmål kunne vores analyser ikke besvare. Vi kunne kun konkludere, at det var vigtigt, at det blev undersøgt, fordi fejl i beregningerne af elevernes dygtighed betød, at testresultater fra DNT hverken kunne benyttes pædagogisk i klassen eller til den form for statistiske opgørelser, som ministeriet har brug for, medmindre fejlene blev rettet. Og naturligvis at resultater af statistiske analyser af DNT-resultater kan være tvivlsomme, hvis det viser sig, at der er fejl i data. Vi kunne også konkludere, at forskere, der havde anvendt testresultater fra DNT, burde overveje om- og på hvilken måde resultaterne af analyser baseret på rettede DNT-resultater ville afvige fra deres forskningsresultater.

Disse betragtninger gav anledning til en del postyr, fordi en række forskere mente, at det var forkert af os at stille spørgsmål til resultaterne af deres forskning. Børn- og Undervisningsministeriet fulgte vores opfordring til at kontrollere, at de sværhedsgrader, som DNT benyttede, svarede til de virkelige sværhedsgrader. STIL (2020) påviste, at de problemer, som vi havde fundet i et enkelt fag og på et enkelt klassetrin, rent faktisk havde eksisteret for alle fag og alle klassetrin siden 2010. Og at fejlene var så store, at ministeriet fandt det nødvendigt at genberegne testresultaterne for alle fag fra 2014[1].

Derved burde det problem være løst. Forskere, der ønsker at bruge DNT-resultater fra og med 2014, kan nu gøre det med forventning om, at tallene er uden systematiske fejl og uden usystematiske fejl udover dem, der altid er forbundet med pædagogiske test. Hvis ministeriet samtidig havde besluttet også at rette testresultaterne fra 2010 til 2014 eller helt havde slettet resultaterne fra før 2014, ville vi have været fri for den sidste tids diskussion af resultaterne fra Holm et.al. (2023). Det er desværre ikke tilfældet.

2. Fire rapporter om DNT resultater.

Nu er diskussionen dukket op igen, og jeg vil tillade mig at trække tre konklusioner op fra gennemlæsning af fire artikler, der har fællestræk, som de deler med de fleste af de artikler, som jeg har haft anledning til at se på, og som er udtryk for forskningsmæssige problemer.

Det første problem er, at de alle benytter sig af relativt enkle lineære regressionsmodeller, der ikke kan garantere at være robuste over for den type af fejl i data, som vi og ministeriet fandt i DNT-resultaterne. Regressionsmodellerne baseres i øvrigt på nogle forudsætninger, der hverken omtales eller afprøves.

Det andet problem er, at man i artiklerne transformerer DNT-resultaterne, på en måde, der udover at reducere objektiviteten af testresultaterne, skjuler den usikkerhed, der er forbundet med målingerne af færdighederne. Transformationen af DNT-resultaterne som gør det umuligt at

[1] STIL (2020) rapport er hensynsløst ærlig gennemgang af alle de fejl, som ministeriet afslørede og af det, som det betød for DNT-resultaterne. Rapporten bør være tvangslæsning for alle, der ønsker at benytte de DNT-resultater, som ministeriet endnu ikke har rettet.

tolke analyseresultaterne på en måde, der fortæller noget konkret om de færdigheder, som DNT forsøger at måle.

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.

Det første problem er et relativt elementært statistisk problem. Det drejer sig om forudsætninger om lineære strukturer i testresultater fra pædagogiske test og om antagelser om såkaldt varianshomogenitet. Lineære relationer mellem variable er naturligvis ikke umulige, men det er noget, jeg altid har belært mine studerende om skal kontrolleres. Antagelsen om varianshomogenitet kan på forhånd siges at være urealistisk, fordi DNT-resultaterne er behæftet med usikkerhed, der varierer meget mellem eleverne. Ud fra det vi får at vide i artiklerne, er det ikke muligt at se, hvor stort problemet er for artiklernes resultater, men problemerne er velkendte, og der er flere forskellige måder, de kan løses på, hvis man er opmærksom på problemerne.

De to andre problemer er fundamentale. Det er disse to problemer, der er den væsentligste grund til, at jeg har brugt tid på at skrive dette notat. Tre af de fire artikler stiller interessante spørgsmål om faktorer, der antages at kunne have betydning for udvikling af elevers færdigheder og kompetencer, men de håndterer testresultaterne på en sådan måde, at den eneste konklusion, der i givet fald kan drages er, at ”ja, det ser ud til, at den nævnte faktor har betydning”, uden at sige noget konkret om på hvilken måde, med hvilken styrke og under hvilke betingelser de nævnte faktorer har en relevant effekt på elevernes færdigheder og kompetencer. Og da det er det, som pædagogiske forskning har brug for, kan resultaterne kun opleves som uinteressante og irrelevante. Og det er for tre af artiklerne en stor skam.

De næste afsnit knytter nogle kommentarer til disse problemer.

3. Om målinger af færdigheder vha. pædagogiske test og Rasch modeller.

Den måde de fire artikler håndterer DNT-resultaterne på rejser en mistanke om, at forfatterne ikke forstår, hvad de har at gøre med, når de analyserer resultater fra pædagogiske test. Jeg har i kapitel 3 i Bendixen og Kreiner (2009) forklaret, hvad man får ud af at måle færdigheder ved hjælp

af såkaldte Rasch modeller, og jeg er nødt til at nøjes med at referere til dette kapitel. Der er imidlertid fire ting, der er vigtige for diskussionen.

For det første at hensigten med testresultater fra Rasch modeller er at konstruere målinger af enkeltstående færdigheder, der kan sammenlignes med målinger i natur- og sundhedsvidenskaberne. For at værdsætte dette er man nødt til at læse, hvad Rasch har skrevet om det. I Rasch (1968), som jeg har fornøjelsen at genlæse lige nu, formulerer han det på følgende måde:

> The models are referred to as ”models for measuring” because of the close relationship between measurement by Rasch models and measurement as understood in physics.

Målinger vha. Rasch modeller (og i øvrigt alle andre IRT-modeller) skal med andre ord være valide, præcise og objektive uden systematisk bias og med så lille grad af usikkerhed som muligt. Og de skal sige noget konkret om det fænomen, som man forsøger at måle. Da et DNT-resultat for et bestemt fag på et bestemt klassetrin består derfor af tre valide, objektive og konkrete målinger, som passer til tre *forskellige* Rasch modeller. Et DNT-resultat betragter med andre ord testresultatet som målinger af tre forskellige færdigheder. Og den afprøvning, som i sin tid blev foretaget, giver ikke belæg for at opfatte det på anden måde, selvom man kan have en mistanke om, at der i virkeligheden kun er tale om en eller to forskellige færdigheder[2].

For det andet er det vigtigt for diskussionen at forstå, at målinger vha. Rasch modeller består af statistiske estimater af person- eller elevparametre, der placerer elevernes færdigheder på en abstrakt intervalskala, som teorien for Rasch modeller omtaler som logit-skalaer. DNT leverer tre sådanne mål for hvert fag og hvert klassetrin svarende til tre forskellige profilområder, og afprøvningen af DNT bekræftede, at opgaverne inden for hvert profilområde levede op til det, som Rasch modellen kræver.

For det tredje er det vigtigt, at usikkerheden på målingerne angives ved standardfejlen af dette estimat. Psykometrien får det til at lyde som noget specielt, ved at omtale det som ”standard error

[2] Beslutningen om, at der skulle være netop tre profilområder på hver klassetrin og hvert fag, var en embedsmandsbeslutning, som jeg aldrig har hørt forsvaret ud fra faglige argumenter. Man kan synes, at det ikke giver mening, men afprøvningen vha. Rasch modeller understøtter ikke påstande om, at eventuelle gennemsnits vurderinger er udtryk for valide målinger.

of measurement" (SEM), men det skal ikke misforstås. Der er tale om den samme form for standardfejl, som beregnes i alle elementære statistiske analyser. Psykometrisk teori insisterer på, at testresultater, der skal bruges på elev- og klasseniveau, skal have SEM værdier mindre end eller lig 0,30. Hvis DNT havde fungeret som forventet, ville det store flertal af eleverne have SEM værdier for alle profilområder på dette niveau. Som bekendt lykkedes det ikke, hvilket betyder at elever måles med meget forskellige SEM værdier og at disse værdier som regel er større end 0,30.

For det fjerde er det vigtigt at nævne, at logit-skalaen er en abstrakt skala med værdier, der kan være vanskelige at tolke for alle andre end erfarne psykometrikere. I forsøget på at konstruere målinger, der er sammenlignelige med målinger fra natur- og sundhedsvidenskaberne, er vi nødt til at indrømme, at en elev læsefærdighed målt på Rasch modellens logit-skala ikke fortæller noget konkret om, hvorledes eleven læser. Det skal ikke betragtes som et principielt problem for Rasch modellen. Et udsagn om at temperaturen er lig med 42 grader giver først konkret mening, når vi får at vide, om der er tale om målinger på Celsius eller Fahrenheit skalaen. Problemet er det samme for logit-skalaen. Der skal noget mere til at give det mening. Men det er et praktisk problem, som man med rette kan kritisere os for ikke at have løst ordentligt. Af den grund præsenteres testresultater på flere forskellige måder, hvor der kan stilles spørgsmål til både objektivitet og i hvor høj grad testresultaterne giver mening.

Jeg vil bruge resultater fra en analyse af PIRLS 2016 i Danmark til at illustrere disse problemer. PIRLS indeholder en række forskellige hæfter med skønlitterære og faglige tekster og tilhørende opgaver. De fleste hæfter er ikke offentlig tilgængelige, men et af dem (Booklet16) er offentliggjort, og den danske version kan ses i Meiding et.al. (2017b). Selvom det aldrig har været hensigten at anvende opgaverne i PIRLS på elev og/eller klasseniveau, kan vi bruge Booklet16 til at illustrere, hvordan PIRLS testen ville have fungeret på elev niveau, hvis man havde fået den ide at bruge PIRLS i stedet for DNT.

Opgaverne i Booklet 16 passer faktisk til en generaliseret Rasch model[3]. PIRLS leverer altså valide målinger af læsefærdighed i fjerde klasse, som rasch ville påstå var objektive. Tabel 1 viser fire forskellige måder man kan opgøre resultaterne på. De er alle valide, men det er kun to af dem, der

[3] Dokumentationen for denne påstand vil blive offentliggjort i anden sammenhæng.

er objektive, og det er kun to af dem, som i begrænset omfang siger noget konkret og meningsfuldt om elevernes læsefærdigheder.

Det første, der altid sker, når man opgør testresultatet, er, at man optæller antallet af point eller antallet af rigtigt besvarede opgaver. Hvis opgaverne er kendt, og hvis læreren har erfaringer med testen og har sat sig ind i baggrunden for opgaverne, vil denne score give en begrænset grad af mening for læreren. Antallet af point i PIRLS kunne derfor give en vis mening for læreren, men det ville være vanskeligt at sige meget konkret om, hvad eleven kan, eller at beskrive, hvad forskelle i testresultater betyder.

I forbindelse med de adaptive test i DNT blev antallet af korrekt besvarede opgaver naturligvis også beregnet, men da opgaverne var hemmelige, og da alle elever i princippet besvarer forskellige opgaver, giver det ingen mening at rapportere dette resultat. DNT måtte derfor bruge en af de tre andre muligheder, som man kan se i Tabel 1.

Den næste søjle med logit-værdier viser estimaterne af elevparameteren i Rasch modellen. Disse tal lever op til de krav, som man stiller til kvantitative målinger på intervalskalaniveau, og det er derfor logit-værdierne, der giver den form for målinger, som Rasch efterlyste. Og det er standard-fejlen (SEM) på disse estimater, som psykometrikere bruger til at vurdere, om målingerne fungerer sikkert nok til, at man kan bruge testen på elev-niveau. Midt på skalaen er SEM tæt på dette mål, men blandt elever med lave eller høje scores er sikkerheden ikke helt god nok. Sammenlignet med DNT må man derfor konkludere, at PIRLS 2016 leverede bedre målinger end DNT. Men logit-værdierne fra PIRLS har præcise de samme problemer som DNT havde. Logit-værdierne siger intet konkret om, hvordan og hvor godt eleverne læser. En logit-værdi på -0,101 svarer til en samlet score på 22 ud af 40 point. Det lyder måske ikke så dårligt, men hvorvidt det er udtryk for en god læser afhænger af sværhedsgraden af opgaverne. Lærerens vurdering vil formodentlig afhænge af, om der er tale om mange lette eller mange vanskelige opgaver, men logit-værdien på -0.101 siger intet i sig selv.

Tabel 1 Fire forskellige måder at score resultaterne i Booklet 16 i PIRLS 2016

Score	Logit	SEM	Percentil	Z_logit
0	-5,042	,651	,0000	-5,9761
1	-3,877	,767	,0000	-4,6501
2	-3,301	,704	,0000	-3,9945

3	-2,902	,623	,0046	-3,5403
4	-2,592	,557	,0068	-3,1875
5	-2,335	,507	,0091	-2,8949
6	-2,114	,469	,0137	-2,6434
7	-1,920	,440	,0251	-2,4226
8	-1,745	,417	,0274	-2,2234
9	-1,585	,399	,0388	-2,0413
10	-1,438	,383	,0502	-1,8739
11	-1,301	,370	,0616	-1,7180
12	-1,173	,360	,0731	-1,5723
13	-1,051	,351	,0890	-1,4335
14	-,934	,343	,1005	-1,3003
15	-,823	,337	,1301	-1,1739
16	-,714	,332	,1644	-1,0499
17	-,609	,328	,1872	-,9304
18	-,505	,324	,2032	-,8120
19	-,403	,322	,2397	-,6959
20	-,302	,320	,2763	-,5809
21	-,201	,319	,3014	-,4660
22	-,101	,319	,3607	-,3521
23	-,001	,320	,4018	-,2383
24	,099	,321	,4429	-,1245
25	,199	,324	,4840	-,0107
26	,301	,328	,5228	,1054
27	,405	,334	,5731	,2238
28	,511	,341	,6233	,3444
29	,622	,350	,6895	,4708
30	,740	,361	,7534	,6051
31	,867	,376	,8014	,7497
32	1,007	,394	,8562	,9090
33	1,161	,418	,8950	1,0843
34	1,336	,448	,9384	1,2835
35	1,536	,487	,9680	1,5111
36	1,772	,538	,9795	1,7797
37	2,060	,605	1,0000	2,1076
38	2,436	,686	1,0000	2,5355
39	2,990	,751	1,0000	3,1661
40	4,138	,643	1,0000	4,4728

Det er for at give testresultaterne en vis grad af mening, som de fleste ville kunne forholde sig til, at man anvender såkaldte norm-baserede skalaer, hvor en elevs resultat sammenlignes med testresultaterne i en – forhåbentlig relevant – elevpopulation. Hvis vi kan gå ud fra, at de elever, der svarede på Booklet 16 i 2016, er et repræsentativt sample af 4. klasses elever i Danmark i 2016, kan man bruge såkaldte percentil-scores, der angiver hvor stor en del af eleverne i fjerde klasse, der er lige så svage læsere som den elev, man er interesseret i. En elev, der har scoret 22 point hører til de 36.1 % svageste elever i fjerde klasse i 2016. Selvom eleven svarede korrekt på

mere end halvdelen af opgaverne, må læreren konkludere, at det var et relativt dårligt resultat. Om niveauet er kritisk er et andet spørgsmål. I en population, der kun bestod af super-læsere, vil der naturligvis være nogle, der hører til de 10 % svageste. Udsagn fra percentil- scores giver naturligvis mening, men meningen er begrænset og udsagnet er ikke objektivt. Meningen er begrænset, fordi det kun er et udsagn om en rangordning af elever på fjerde klassetrin. Et sådant tal kan måske være interessant for elevernes forældre, men en percentil-score fortæller intet konkret, som læreren har brug for, hvis hun ønsker at hjælpe eleven. Og udsagnet er ikke objektivt, fordi det afhænger af andre elevers resultater. En percentil-score på 36.1 % i 2016 vil for eksempel næppe svare til en percentil-score på 36.1 % i den seneste PIRLS, hvor resultaterne, antyder, at danske elever som helhed er dårligere, end de var i 2016. En læser fra 2016 med en percentil score på 36,1 % i 2016 vil læse bedre målt på logit-skalaen end en læser fra 2021 med en percentil-score på 36,1 % i forhold til eleverne i 2021.

Med andre ord: percentil-scores har mening, men meningen er begrænset, og percentil-scores kan ikke opfattes som målinger på samme niveau som målinger fra natur- og sundhedsvidenskab.

Det sidste forslag, som er med i Tabel 1 er de såkaldte standardiserede z-scores. Man beregner en z-score ved først at beregne middelværdien og standardafvigelsen af logit værdierne i hel population, hvorefter man trækker middelværdien fra elevens logit-score og dividerer resultatet standardafvigelsen. Resultatet er en z-score med middelværdi lig nul standardafvigelse lig med en i populationen af elever.

For mere end 50 år siden brugte man standardiserede scores til beregning af percentil-scores, hvis man kun havde mulighed for at indsamle begrænsede data fra den relevante elev-population. Man håbede – ofte uden at have grund til det og uden at have kontrolleret det – at fordelingen af eleverne var normal, således at man kunne bruge percentilerne fra den standardiserede normal til at beregne standardiserede percentil-scores. Men ikke som testresultater på elevniveau. Standardiserede z-scores er hverken meningsfulde eller objektive, og det giver ingen mening at begynde at sammenligne standardiserede z-scores på tværs af data indsamlet på forskellige tidspunkter. Den samme standardiserede z-værdi kan fortælle meget forskellige historier om læsefærdigheden, hvis fordelingerne er meget forskellige.

Det er klart at disse betragtninger skaber problemer. Er det overhovedet muligt at tolke testresultater på en meningsfuld måde, der fortæller noget om, hvad et testresultat fortæller om enkelte elevers læsefærdigheder, og hvad forskelle på testresultater og effekten af forskellige forsøg på at forbedre læsefærdighederne betyder?

Svaret er, at det er muligt. Man skal holde sig til logit-scoren og bruge den til at beregne såkaldte skala-forankrede fordelinger med sandsynligheder for svar på spørgsmål og opgaver for alle elever på præcis det samme logit-niveau. Det er klart, at det kræver et relativt stort og omhyggeligt arbejde at sammenfatte fortolkningen af resultaterne på en måde, der giver mening for lærere og andre brugere af testene. Og at det kræver samarbejde med fag-didaktikere. Det har jeg ikke mulighed at gør noget ved her, men jeg kan illustrere det ved at se på resultaterne fra vores analyse af PIRLS data.

PIRLS 2016 stillede krav til fire læse- og forståelses processer:

1) Elever skal være i stand til at finde og uddrage informationer i teksten. Opgave T08C er den vanskeligste af de opgaver, der stiller dette krav.
2) Elever skal være i stand til at drage direkte følgeslutninger, ud fra det man har læst. Opgave T03C er den vanskeligste af de opgaver, der stiller dette krav.
3) Elever skal kunne fortolke og samordne centrale ideer og informationer i teksten. H04C er den vanskeligste af de opgaver, der stiller dette krav.
4) Elever skal være i stand til at vurdere og tage kritisk stilling til indhold og tekstuelle elementer. Opgave H16c er den vanskeligste af de opgaver, der stiller dette krav.

Hvis man vil forstå, hvad en logit-værdi fortæller om elevens læsefærdigheder, er det en god ide at se på sandsynlighederne for korrekte svar på de vanskeligste opgaver inden for de fire processkrav.

Tabel 2 viser disse sandsynligheder for elever med logit-værdier lig med -0.101.

Tabel 2. Sandsynligheder for svar på fire vanskelige opgaver[1]. Logit = -0,101.

		Antal point		
Opgave	**Process**	**0**	**1**	**2**
T08C	**Informationer**	**38,1 %**	**61,9 %**	
T03C	**Følgeslutninger**	**35,7 %**	**35,0 %**	**29,3 %**
H04C	**Fortolkninger**	**87,8 %**	**12,2 %**	

H16C	**Vurdering og kritik**	**82,5 %**	**17,5 %**	

[1]: Opgave T03C er en såkaldt polytom opgave, hvor eleverne kan score 0, 1 eller 2 point.

Ifølge Tabel 2 kan man se tegn på en positiv udvikling af færdigheder i at søge informationer og i at drage følgeslutninger blandt elever med logit-værdier tæt på -0,101, men at færdigheder i informationssøgning og evner i at drage følgeslutninger er relativt usikre med en ikke ubetydelig risiko for at begå fejl. Til gengæld har disse elever, så store problemer med fortolkninger og kritiske vurderinger, at det er tvivlsomt om udviklingen af disse færdigheder for alvor er startet.

Hvis man antager at et forsøg med intensiveret læseundervisning kan påvise en logit-effekt på 0,4 skal man se på en tabel med skalaforankrede fordelinger for elever med en logit-score på 0.301. Tabel 3 viser betydningen af den intensiverede læseundervisning for eleverne i Tabel 2. Søgning efter informationer er blevet så meget bedre, at vi kan karakterisere informationssøgning som forholdsvis sikker. Evner til at drage følgeslutninger er også bedre med en begrænset risiko for slet ikke at score noget på opgave T03C. Til gengæld er der stadig store problemer med tolkninger af teksten, og der kan kun ses svage tegn på en begyndende udvikling i færdigheder i vurdering og kritik.

Tabel 3. Sandsynligheder for svar på fire vanskelige opgaver[1]. Logit = 0,301.

		Antal point		
Opgave	**Process**	**0**	**1**	**2**
T08C	**Informationer**	**24,2 %**	**75,8 %**	
T03C	**Følgeslutninger**	**19,5 %**	**32,3 %**	**48,2 %**
H04C	**Fortolkninger**	**82,6 %**	**17,4 %**	
H16C	**Vurdering og kritik**	**73,7 %**	**26,3 %**	

I hvor høj grad man synes, at udbyttet af den intensiverede læseundervisning er tilstrækkelig, og naturligvis om mine tolkninger giver mening, kan naturligvis diskuteres. Jeg er ikke læse-didaktiker, og jeg lader mig gerne belære om, hvordan tallene i Tabel 2 og 3 skal fortolkes. Pointen er imidlertid, at det er muligt at beskrive læsefærdighederne i termer, der fortæller, hvor eleven er i sin udvikling og i givet fald at vurdere effekten af interventionen i fagdidaktiske termer. Men kun hvis man måler læsefærdigheden på Rasch modellens logit-skala. Hvis man havde målt effekten af

forsøget i forhold til antallet af point eller i forhold til percentilerne eller de standardiserede z-scores, ville man naturligvis kunne påvise en signifikant effekt[4], men man ville ikke kunne sige noget som helst om effekten af forsøget og derfor heller ikke, om det ud fra faglige synsvinkler havde været besværet værd.

4. De fire artikler.

Kæderne ryger af på flere forskellige måder i forbindelse med de analyser, vi kan læse om i de fire artikler.

4.1 The Danish National Tests at a Glance

Artiklen af Beuchert & Nandrup (2018) indeholder en kort præsentation af DNT og illustrerer, hvorledes B&N mener, at man kan anvende DNT resultater i forbindelse med statistiske analyser af den måde færdigheder påvirker senere oplysninger som resultater af afgangsprøver eller efterfølgende tilgang til uddannelser efter folkeskolen.

I forbindelse med sådanne analyser sammenfatter B&N testresultaterne inden for et fag som en vægtet sum af resultaterne fra de tre profilområder ved først at standardisere logit-værdierne fra de enkelte profilområder og derefter standardisere summen af de standardiserede profiler til et samlet mål for det faglige niveau. Det er tilsyneladende denne vægtede sum af profilresultaterne, som B&N mener, man skal bruge til i en regressionsanalyse, der skal vise effekten af færdigheder målt af DNT på det senere resultat af folkeskolens afgangsprøver. De konkluderer, at der er tale om en signifikant effekt, og at den forventede forskel på karaktererne for to elever, hvor den ene ligger 1 point højere på deres samlede færdighedsskala end den anden vil være ca. to point.

Jeg kan desværre ikke sige det på en mere diplomatisk måde. Det er ikke første gang jeg ser analyser af sammenhængen mellem DNT-resultater og senere karakterer i afgangsprøverne fra folkeskolen. Disse analyser viser altid det samme. At der er en signifikant effekt. Det er af den grund, at jeg til at begynde med havde lidt svært ved at finde noget interessant i artiklen. Men lad

[4] Hvis man kun er interesseret i p-værdier og signifikans, er det naturligvis ikke forkert at standardisere testresultaterne. Men det er fuldstændig overflødigt, i forbindelse med de analyser vi bliver præsenteret for i artiklerne. P-værdierne er de samme for analyser baseret på logit-scores og analyser baseret på z-scores.

gå med det. Jeg er til gengæld nødt til at sige, at det er umuligt for mig at finde noget positivt at sige om den måde de har analyseret deres data på. Og at det vil være ulykkeligt, hvis deres artikel har inspireret andre til at håndtere DNT-data på samme måde.

En ting er at standardisere de enkelte profil-resultater. Det har jeg allerede diskuteret. Det er en enkel lineær transformation, som gør det umuligt at tolke resultaterne. 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 og forsøge at vurdere, hvad en ændring på logit-niveau ville betyde for de efterfølgende afgangskarakterer.

Standardiseringen af testresultater er en alvorlig fejl. Men det værste er, at de efterfølgende først beregner summen af de standardiserede profil-scores og derefter standardiserer denne sum. For så vidt angår læsning betyder det, at den score, som benyttes til at vurdere effekten af DNT-resultatet i læsning i 6. klasse er en vægtet sum af logit-værdierne for afkodning, sprogforståelse og tekstforståelse uden oplysninger (og formodentlig også uden viden) om hvilke af de tre profilområder, der vejer tungest. De Rasch analyser, der validerede de enkelte profilområder kan ikke bruges som argument for, at de har et validt og objektivt mål for læsefærdigheder i 6. klasse, og det er ganske enkelt umuligt at vurdere den usikkerhed, der er forbundet med deres læse-score. Og det er fuldstændig umuligt at vurdere, hvad et enkelt point på deres score fortæller om forskelle i læsefærdigheder.

Processen er den samme for resultater andre fag inklusiv læsning i 8.klasse. Det eneste, man kan sige om det, er, at deres læse-score i 8. klasse formodentlig vægter de tre profilområder på forskellig vis. Det vil sige, at man ikke en gang kan sammenligne et point på deres læse-score i 6. klasse med et point på læsescoren i 8. klasse.

Der er mange andre ting at grave sig ned i, hvis man vi se, hvad der kunne gøres bedre:

1) En regressionsanalyse med de tre profilområder hver for sig kunne have været interessant.
2) En konfirmatorisk faktoranalyse, der kunne underbygge ideen om at samle de treprofilområder i en samlet score, og som kunne have fortalt, hvorledes man skulle vægte de tre profilområder, ville have været meget interessant.
3) En regressionsanalyse, der behandlede afgangskarakteren som en ordinal kategorial variabel burde have været en selvfølge. Der er mange måder at foretage sådanne på. De er

lidt mere besværlige end de lineære regressioner, som artiklen benytter. Men så er det heller ikke værre.

4.2 The Positive Effects of Nationwide Testing on Student Achievement in a Low-Stakes System.

Artiklen af Andersen & Nielsen (2016) stiller et interessant spørgsmål, som fortjener et klart svar. På grund af et crash af DNTs programmer I 2010 var der en del elever, der ikke fik mulighed for at tage de obligatoriske test. A&N benyttede dette uheld til at indsamle data, der kunne belyse, om der de næste år var forskel i DNT resultaterne blandt de elever, der tog testene i 2010 og de elever, der ikke kunne tage testene, for således at undersøge, om der kunne påvises en positiv effekt af det at tage en test i et specifikt år på udviklingen af færdigheder de senere år.

Desværre benytter denne artikel næsten den samme måde at håndtere testresultaterne på som B&N. Det var elever fra 2., 4. og 6. klasse, der blev testet i læsning i 2010 og som blev testet i læsning to år senere, og elever fra 3. og 6. klasse, der blev testet i matematik i 2010 og gentestet i 2013. For at definere en samlet variabel, der kunne bruges som afhængig variable, standardiserede A&N profilresultaterne fra hvert enkelt fag og hvert enkelt klassetrin og beregnede gennemsnittet af disse som et samlet mål for færdigheder i henholdsvis læsning og matematik i henholdsvis 2012 og 2013 uden skelnen mellem testresultater i det samme fag. Analyserne kunne påvise en signifikant effekt af at være blevet testet i 2010 på læsefærdigheden i 2012, men ingen effekt af at være blevet testet i 2010 på færdighed i matematik i 2013. Forskellen i læsefærdigheder i 2012 på dem, der blev testet, og dem, der ikke blev testet i 2010, er lig med 0,0918 på den standardiserede skala forlæsefærdighed i 2012. Hvor vidt det er en stærk effekt, der skal tages alvorligt eller en svag effekt, der kan ignoreres, er umuligt at sige ud fra det artiklen oplyser, fordi testresultaterne i læsning i 2012 er vægtede summer af afkodning, sprogforståelse og tekstlæsning med ukendte vægte der formodentlig er forskellige for 4., 6. og 8. klasse. Interessen for disse resultater er derfor lige så begrænset som interessen for resultaterne i B & N.

4.3 The Effect of Teacher's Aides in the Classroom: Evidence from a Randomized Trial.

Andersen et.al. (2020) stiller interessante spørgsmål vedrørende effekten af hjælpelærere i dansk og matematik og tager hensyn til testresultater i matematik og læsning taget, før der blev knyttet

hjælpelærer til klassen. Resultaterne er de samme som i forrige artikel. Der er en signifikant effekt i forbindelse med læsning, men ingen i forbindelse med matematik. Desværre præsenteres udelukkende resultater for testscores der er standardiseret på samme måde som i B&N. Spørgsmål vedrørende karakteren og styrken af effekten på læsefærdigheder er derfor et også et ubesvaret spørgsmål i denne artikel.

4.4 The Effect of parental Union Dissolution on Children's Test Scores.

Artiklen af Holm, Fallesen & Heinesen (2023) afviger på flere punkter fra de tre foregående. Artiklen, der forsøger at måle den kausale effekt af skilsmisser på elevernes færdigheder, ser kun på et enkelt færdighedsområder H, F & M ser kun på tekstforståelse og undlader at beregne vægtede summer af de tre profilområderne knyttet til læsefærdighederne. De bruger lineære regressionsmodeller, men de kausale analyser, de foretager, er af en helt anden karakter end de elementære regressionsanalyser, som vi fandt i de forrige artikler. Så vidt, så godt, men målingerne af tekstforståelsen standardiseres på samme måde som i de forrige artikler

Data til de analyser, de fortager, består af samtlige målinger af tekstforståelse i 2., 4., 6. og 8. klasse fra 2009/2010 til 2017/8, inklusiv oplysninger om tidspunktet for opløsning af forældrenes parforhold og en række relevante baggrundsvariable som forældrenes indtægt og uddannelse.

De analyser, der foretages omtales som "Difference-in-Differences" analyser. Det vil føre for vidt at beskrive detaljerne i sådanne analyser, så jeg må nøjes med at henvise til den på alle måder fine artikel af Callaway & Sant'Anna (2021), som forfatterne læner sig op ad, og hvor man også kan læse om de forudsætninger, som analyserne bygger på.

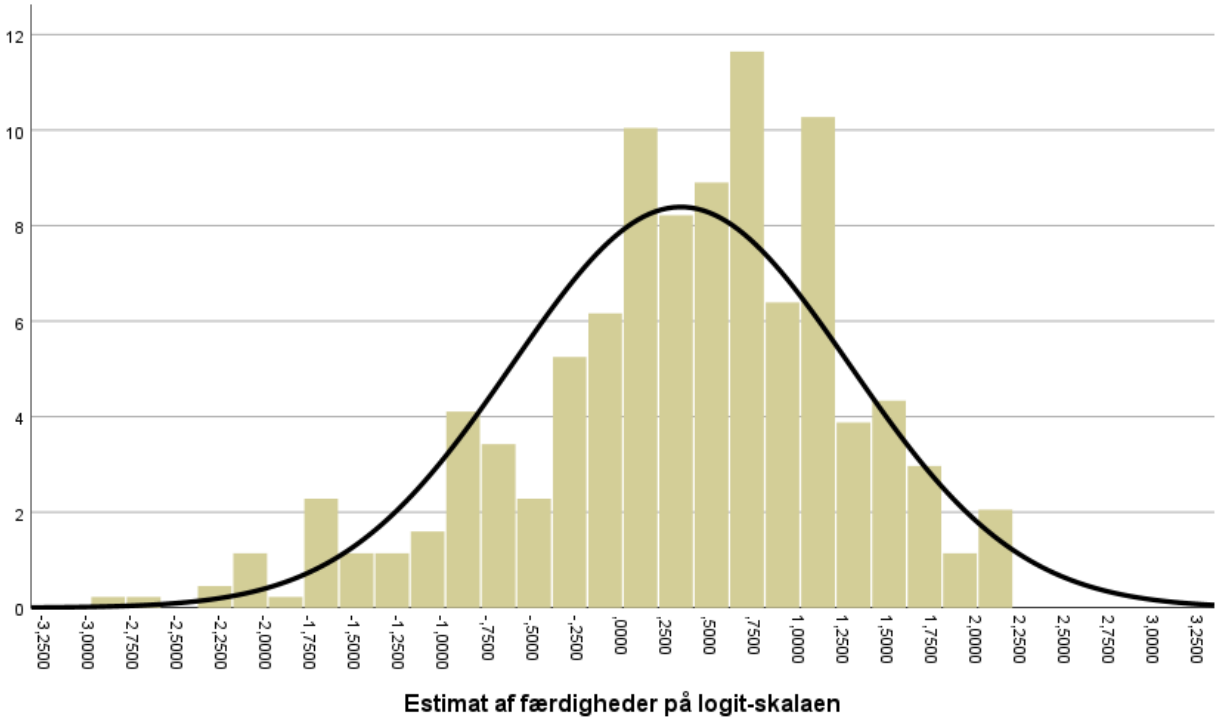
Jeg kan imidlertid illustrere, hvad moderne kausale analyser forsøger at gøre og de konsekvenser, som standardiseringer af testresultater har for resultaterne af analysen.

Jeg vil endnu engang benytte resultater fra PIRLS 2016, hvor svarene på opgaverne i Booklet 16 passer til en generaliseret udgave af Rasch modellen Figur 1 viser fordelinger af disse målinger for de elever, der svarede på samtlige opgaver.

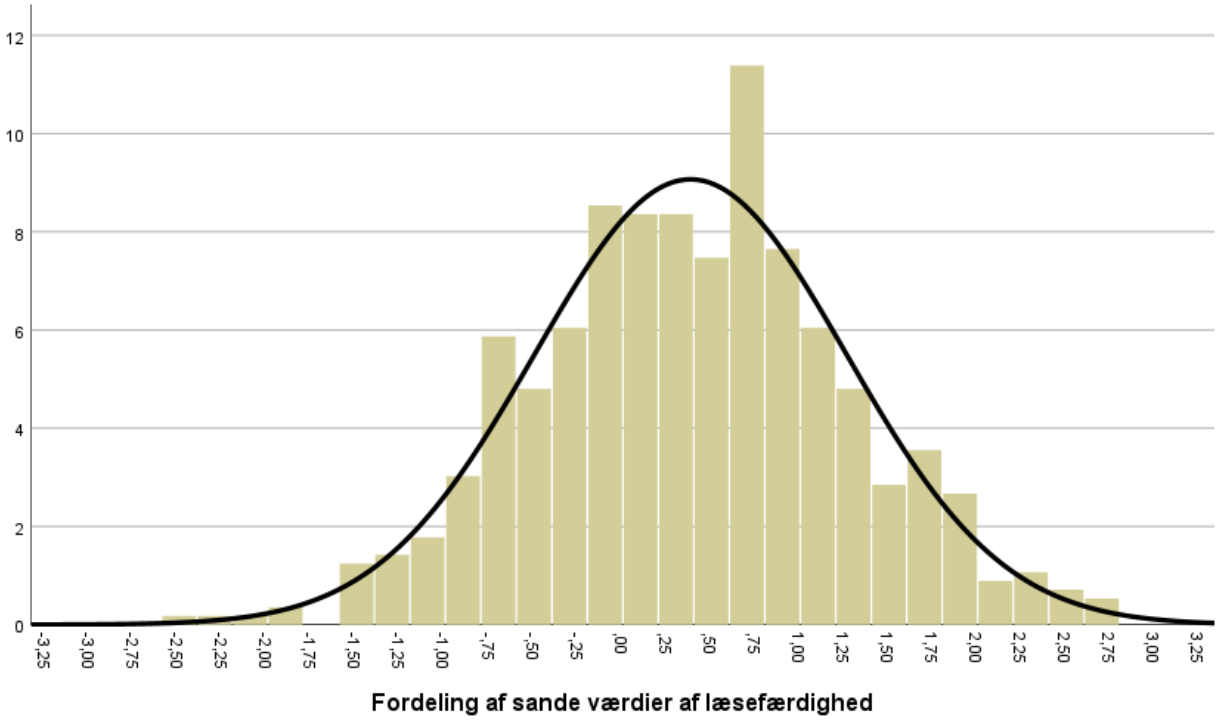
Målingerne er foretaget med målefejl, og der er problemer med systematisk bias for elever på et så højt niveau at de kan svare korrekt på alle opgaver. Fordelingen af de estimerede logit-værdier ligner en normalfordeling, og det er muligt at estimere middelværdien af målinger af færdigheden uden fejl, hvis vi antager at den sande fordeling rent faktisk er normal. Middelværdien af de sande værdier for læsefærdigheden er lig med 0,37 og standardafvigelsen er lig med 0,91. Figur 2 viser fordelingen af 561 elever trukket tilfældigt fra denne fordeling.

Hvis man havde standardiseret værdierne ud fra fordelingen af de sande værdier for læsefærdigheden, på samme måde som det sker i de fire artikler, ville fordelingerne se ud på fuldstændig samme måde, men middelværdien ville være lig med nul og standardafvigelsen lig med en i figur 2, mens middelværdien ville være lig med 0,03 og standardafvigelsen lig med 1,04 i figur 1. Dvs. lidt forskellige på grund af usikkerheden i målingerne.

Som beskrevet tidligere er forskellen på målingerne på logit-skalaen og målingerne på den standardiserede z-skala er, at det kun er logit-værdierne, der fortæller os noget om, hvor mange vanskelige og hvor mange lette opgaver elever på forskellige niveauer har. Da vi kender sværhedsgraderne på opgaverne, kan vi regne os frem til, at der ingen lette opgaver er for elever med en logit-værdier, der er mindre end -1,5, mens mere end halvdelen af opgaverne vil være vanskelige med sandsynligheder der er større end 75 % for forkerte svar. Og vi kan regne os frem til, at der ikke er nogen vanskelige opgaver for elever med logit-værdier på mere end 1,5, og at næsten alle opgaverne er lette for disse elever. Disse to værdier på logit skalaen definerer altså elever med læsevanskeligheder og elever uden læsevanskeligheder ud fra faglige kriterier. Hvis der kun bruges standardiserede resultater uden oplysninger om middelværdier og standardafvigelser på de oprindelige logit-værdier (værdier, der altid bør rapporteres, men som ikke kan findes nogen stedet i de fire artikler), er det umuligt at uddrage nogen informationer om læsefærdighederne ud fra resultaterne af analyserne.



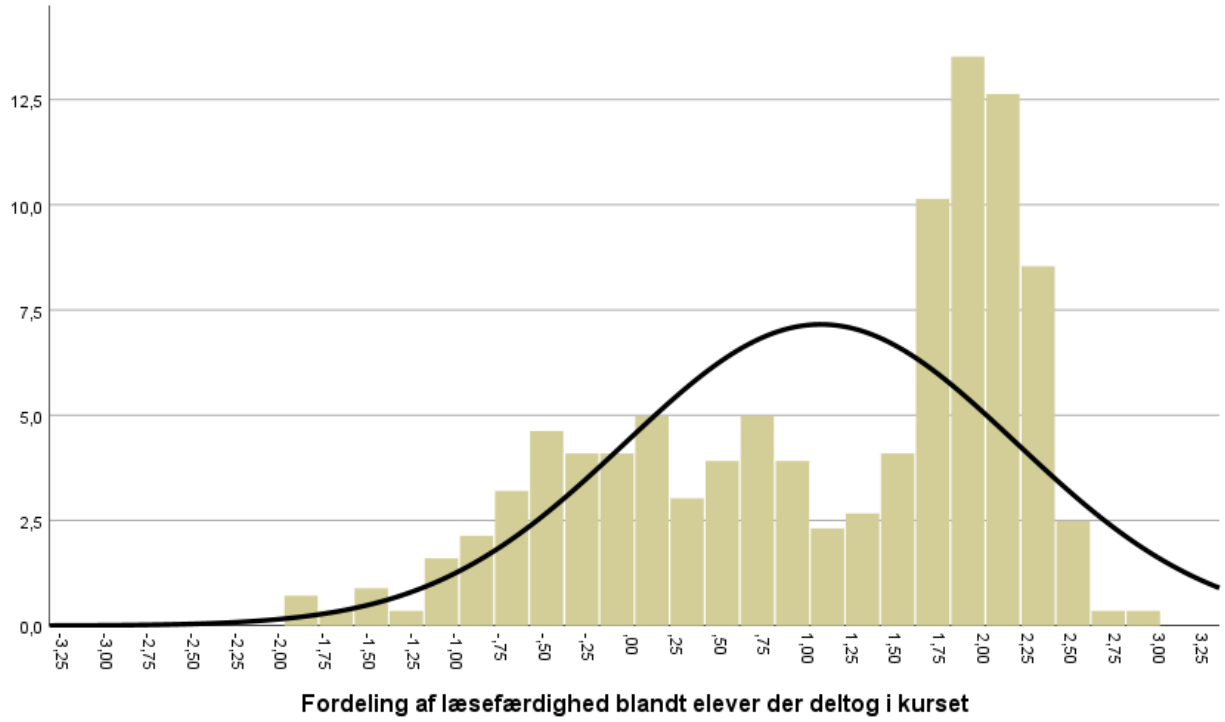
Figur 1. Fordeling af målinger af læsefærdigheden



Figur 2. Fordeling af den sande læsefærdighed for blandt 561 tilfældigt udvalgte elever

Antag igen, at der har været tale om et forsøg med intensiv læseundervisning på et eller andet tidspunkt før læseprøven blev taget. For at måle den kausale effekt af et sådant forsøg vil man i forbindelse med moderne kausale analyser først forsøge at estimere, hvorledes læseresultaterne ville have fordelt sig, hvis eleverne ikke havde fået lov til at deltage i kurset, hvorefter man vil beregne et mål for den kausale effekt som forskellen på middelværdien af de observerede testresultater og middelværdien af de kontrafaktiske resultater.

I vores tænkte eksempel er situationen den, at det kun er en del af eleverne, der får mulighed for at deltage i kurset og at Figur 2 viser den kontrafaktiske fordeling således som fordelingen af læsefærdigheden ville have set ud, hvis der ikke var nogen, der havde deltaget i kurset, mens figur 3 viser fordelingen således som den kom til at se ud pga. forsøget.



Figur 3. Læsefærdighed for elever der deltog i kurset.

Da gennemsnittet af de kontrafaktiske færdigheder i Figur 2 er lig 0,37, og gennemsnittet af de observerede færdigheder er lig med 1.06, vil konklusionen være, at den kausale effekt er lig med 0.69 på logit-skalaen.

I forbindelse med sådanne analyser er vurderingen af den kausale effekt uproblematisk, så længe man holder sig til logit-værdierne og en logit forskel på 0,69 kan (med nød) tolkes som udsagn om forskelle i sandsynligheder for korrekte svar på de enkelte opgaver. Hvis man insisterer på at analyserne skal foretages ved hjælp af standardiserede scores, har man et problem. Hvilken fordeling skal testresultaterne standardiseres i forhold til. I forhold til en hypotetisk fordeling, i forhold til den kontrafaktiske fordeling som man ville have forventet, hvis forsøget ikke havde været gennemført, eller den faktiske fordeling (Figur 3) efter forsøget?

Omregnet til den kausale effekt af de standardiserede scores er effekten lig med 0.76. Tæt på logit-effekten for standardafvigelsen i Figur 2, der er er lig med 0.91, men hvis man kun får resultatet på den standardiserede skala uden oplysninger om standardafvigelsen på logit-skalaen, er der ingen mulighed for at tolke tallet 0,76 på en måde, som siger noget om styrken af den kausale effekt på læsefærdigheden.

Svaret i artiklen er tilsyneladende, at man skal anvende testresultater fra et tidligere tidspunkt og antage, at alle elever (bortset fra rent tilfældige variationer) ville have udviklet sig parallelt, fra før forsøget gik i gang og indtil der skulle testes efter forsøget. At den standardiserede kontrafaktiske fordeling i Figur 2 er den samme som den standardiserede fordeling før fordelingen gik i gang. Og at elevernes percentil-scores i den kontrafaktiske fordeling er de samme (på nær lidt tilfældig variation) som elevernes percentil scores i den tidligere fordeling.

Metoder der bygger på forudsætninger om parallel udvikling omtales som difference in difference metoder. Forudsætningerne kan være korrekte, men behøver ikke at være det. Hvis de ikke er det, vil relationerne mellem de standardiserede værdier og percentilerne ændres, og det bliver vanskeligere og vanskeligere at tillægge standardiserede scores nogen som helst konkret betydning.

Hvis man fx antager. at læseudviklingen målt på logit-niveau foregår hurtigere eller langsommere på et bestemt klassetrin blandt stærke og svage elever. vil enkle difference-in-difference metoder ikke give meningsfulde målinger af kausal effekt, men analyser på det oprindelige logit-niveau vil stadig kunne give meningsfulde resultater. hvis man brugte andet end enkle lineære modeller til at beskrive sammenhængen mellem tidligere og senere testresultater, og hvis man udover kun at

interessere sig for den gennemsnitlige kausale effekt også interesserede sig for fordelingerne af de observerede og kontrafaktiske fordelinger.

5. Regressionsanalyser med variable der er målt med usikkerhed og fejl

Det er tilstrækkeligt at se på tre typer af regressionsanalyser for at illustrere de problemer, som man har i forbindelse med statistiske analyser af testresultater fra pædagogiske test.

5.1 Testresultater som afhængige variable

Formålet med den første type er at undersøge, om læsefærdigheden Y på et bestemt tidspunkt afhænger af en eller flere interessante baggrundsvariabler $X_1,\ldots,X_k$.

En lineær regressionsanalyse antager at

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + E \qquad (1)$$

Model (1) antager, at Y er den sande læsefærdighed målt uden fejl. Det er derfor, at det er værdierne af beta parametrene ($\beta_1,\ldots, \beta_k$), der måler effekten af de forskellige baggrundsvariable, som vi ønsker at få noget at vide noget om. I forbindelse med analysen vil vi derfor have statistiske test, der kan understøtter påstande om effekt, og vi vil have estimater af parametrene uden systematiske fejl, så vi kan udtale os om styrken af, den effekt som baggrundsvariablene har for læsefærdigheden.

Variablen E i model (1) omtales som et fejlled, der skyldes, at en elevs læsefærdighed afhænger af andet og mere end de baggrundsvariable, der er inkluderet i modellen. Da disse forhold ikke inkluderes i analyserne, opfatter modellen E som en stokastisk (tilfældig) variabel.

Den lineære regressionsmodel bygger på to fundamentale forudsætninger. For det første at alle sammenhænge mellem læsefærdigheden Y og baggrundsvariablene er lineære. 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[5]. At der med andre ord kun er tale om usystematiske afvigelser mellem, det modellen forudsiger om læsefærdigheden, og det der observeres, og at

[5] Denne antagelse, der omtales som en antagelse af varianshomogenitet.

omfanget af afvigelserne er de samme for alle personer uanset værdierne af baggrundsvariablene. Udover at teste og estimere effektparametrene bør en lineær regressionsanalyse altid undersøge, om disse antagelser er rigtige. I den sammenhæng er kontrollen af antagelserne om linearitet afgørende, fordi det aldrig er nok at få at vide, at en faktor har betydning, og hvor stærk den er. Vi vil også vide noget om *på hvilken måde*, baggrundsvariablene påvirker læsefærdigheden. Antagelser om, at de er lineære, er valgt, fordi analyserne er enkle, når relationerne er lineære, men der er sjælden saglige argumenter for, at de nødvendigvis skal være lineære. Antagelser om linearitet er en bekvemmelighedsantagelse.

Udover problemerne med de to antagelserne har analyser af DNT-resultater to problemer. Det første er, at pædagogiske testresultater altid måles med en vis grad af usikkerhed. Det andet er at testresultater fra DNT tidligere blev fejlberegnet. Det betyder at DNT-resultaterne var behæftet med flere fejl end dem, der altid følger med pædagogiske test. Disse ekstra fejl kan være usystematiske, hvilket betyder at beregningerne af de såkaldte SEM værdier overvurderer sikkerheden af målingerne, men der kan også være tale om systematiske fejl, hvor dygtigheden undervurderes i visse områder og overvurderes i andre. Systematiske fejl, som vi vil omtale som testbias.

Det betyder, at regressionsanalysen er nødt til at benytte en model med flere typer fejl, fordi man i stedet for at bruge de sande værdier af læsefærdigheden er nødt til at bruge DNT-resultater, der udover den sande læsefærdighed også afhænger af de tre typer fejl.

$$Y_{DNT} = Y + Bias(Y) + E_{DNT}(Y) + F(Y) \tag{2}$$

I formel (2) er Y_{DNT} testresultatet, Y er den sande værdi, $E_{DNT}(Y)$ er den fejl, som altid følger med pædagogiske test (og som afhænger af værdien af Y), Bias(Y) er den systematiske fejl, der skyldes fejlberegningerne af estimatet af Y, og F er den usystematiske fejl, der også skyldes fejlberegningerne, og som vi vil antage har en middelværdi lig med nul.

For at se, hvad det betyder for regresionsanalyser med testresultater i stedet for sande værdier af færdighederne, indsætter vi regressionsmodellen påstande om Y fra formel (1) i formel (2).

$$Y_{DNT} = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + E + Bias(Y) + E_{DNT}(Y) + F(Y) \tag{3}$$

I denne regressionsmodel er fejlledet en sum af tre fejl, hvoraf to af dem afhænger af den sande værdi af færdigheden. Alene af den grund lever denne model ikke op til den lineære regressions-analyses forudsætninger om varianshomogenitet. Konsekvensen er, at den samlede sikkerhed, med hensyn til hvor godt baggrundsvariablene kan prædiktere testresultatet, er mindre end analysen med de sande værdier for læsefærdigheden ville have været.

Hvis der ikke havde været tale om bias på grund af regnefejlene, ville det betyde meget lidt for estimationen af de parametre, som man er interesseret i. Men beregninger af de p-værdier, der skal underbygge påstandene om, at baggrundsvariablene har en effekt, vil have problemer. De vil være forkerte, fordi antagelsen om varianshomogenitet ikke holder, og risikoen for at analysen ender med at konkludere, at der ikke er signifikant evidens for effekterne, vil være større, end hvis man havde målinger uden fejl.

Og for at gøre det værre: Hvis der er tale om systematisk bias, således at sammenhængen mellem de sande værdier og de estimerede værdier ikke er lineære, er antagelserne om linearitet heller ikke opfyldt. Den regressionsmodel, som man har tænkt sig at anvende til analyserne, er vist i formel (4). Når vi allerede ved, at dette er en forkert model, fordi antagelsen om varianshomoge-nitet ikke holder, skal man være optimistisk for at tro på, at estimaterne af parametrene i model (4) er estimater af parametrene i model (1), hvis antagelsen om linearitet heller ikke holder på grund af testbias, der skyldes regnefejlene i DNT. Estimaterne af β^*-estimaterne vil ikke være estimater af de sande parametre, som vi er interesserede i.

$$Y_{DNT} = \alpha^* + \beta_1^* X_1 + \beta_2^* X_2 + ... + \beta_k^* X_k + E^* \quad (4)$$

Ifølge A & N (2016) er styrken og karakteren af den eventuelle effekt af tidligere testning på senere testresultater signifikant, men tilsyneladende ret svag for læsning, og der kan ikke påvises nogen overbevisende effekt for matematik. Brugen af vægtede z-scores over flere profilområder, gør det i forvejen ret umuligt at vurdere styrken og relevansen af effekten på læsefærdighederne. Det hjælper derfor ikke, at fejlene i testresultaterne er væsentlig større, end de i forvejen for usikre DNT-resultater ville have været uden regnefejl. Da styrken af de statistiske test afhænger af fejlledende i regressionsmodellen, kan det ikke afvises, at resultaterne ville have været tydeligere for læsning og signifikante for matematik, hvis der ikke havde være fejl i DNTs beregninger.

5.2 Testresultater som uafhængige variable

I det næste eksempel vil vi se på, hvad usikre og fejlbehæftede målinger kan betyde, hvis læsefærdigheden Y benyttes som en uafhængig variabel i en lineær regressionsanalyse.

Vi antager derfor, at der er en kvantitativ variabel Z, der afhænger af læsefærdigheden Y samt en række andre baggrundsvariable, som det er vist i formel (5). Udgangspunktet for analyserne er altså en påstand om, at relationen mellem Z og Y er lineær, og at afvigelsen mellem de observerede og prædikterede værdier af Z er den samme for alle personer uanset værdierne af den afhængige og de uafhængige variable.

$$Z = \alpha + \beta_Y Y + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + E \tag{5}$$

Da Y er målt med usikkerhed og fejl er vi nødt til at erstatte Y med Y_{DNT} i formel (5) og derefter indsætte formlen, der viser hvorledes Y_{DNT} afhænger af den sande Y-værdi og de forskellige fejl, som målingen af Y kan være behæftet med. Resultatet kan ses i formel (6)

$$\begin{gathered} Z = \alpha + \beta_Y \left[Y + Bias(Y) + E_{DNT}(Y) + F(Y) \right] + \beta_1 X_1 + ... + \beta_k X_k + E \\ = \\ Z = \alpha + \beta_Y \left[Bias(Y) + E_{DNT}(Y) + F(Y) \right] + \beta_Y Y + \beta_1 X_1 + ... + \beta_k X_k + E \end{gathered} \tag{6}$$

Problemet med denne model er, at den antager, at alle de faktorer, der har med målesikkerheden at gøre (den usystematiske usikkerhed i forbindelse med selve testsituationen og usikkerheden og bias pga. regnefejlene) har den samme (kausale) effekt, som den sande værdi af Y har.

I sig selv er det en så meningsløs tanke, og det er klart, at det kommer til at påvirke resultaterne, uanset om der er eller ikke er regnefejl. Restleddet E_{DNT} er jo ikke til at komme uden om. Det vil desværre føre for langt at forsøge med yderligere elaborering af disse problemer i dette notat. Jeg må derfor nøjes med at referere til litteraturen om "error in measurement models" og fortælle, at resultatet af en analyse med testresultatet Y_{DNT} i stedet for den sande Y værdi vil undervurdere effekten af Y på Z med en faktor, der afhænger af reliabiliteten af Y_{DNT}, $\beta_Y^* = (\text{reliabiliteten af Y}) \times \beta_Y$, *hvis der ikke havde været nogle regnefejl*. Og da regnefejlene fører til mere usikkerhed og måske også bias, kan man ikke engang stole på det. Effekten af Y på Z vil blive endnu mere undervurderet. Analyser med testresultater som uafhængige variable skal man altid være meget

varsomme med, og man bør kun foretage dem, hvis reliabiliteten er høj. Når der oven i den almindelige DNT-usikkerhed også er egentlige regnefejl, eller hvis man transformerer testresultater på en måde, hvor man ikke har styr på reliabiliteten, må det anbefales at søge efter metoder, der tager hånd om usikkerheden. Sådanne metoder findes faktisk.

5.3 Testresultater som både afhængige og uafhængige variable

Hvis både den afhængige variabel Z og den uafhængige variabel Y i formel (6) er målt med fejl, får man både de problemer, der blev beskrevet i 5.1, og de problemer, der blev beskrevet i afsnit 5.2 og hvor den lineære regressionsmodel hverken lever op til kravene om linearitet eller kravene om varianshomogenitet. Estimaterne af effektparametrene vil ikke være estimater af de parametre man er interesserede i, og beregninger af de p-værdier, der skal underbygge påstande om signifikans, vil være forkerte. Da det på forhånd kan forudsiges, at testresultater er relativt stærkt korrelerede, vil effekten af Y_{DNT} på Z_{DNT} altid være statistisk signifikant, medmindre der er tale om undersøgelser med meget få elever. Men analyserne kan hverken give realistiske bud på, hvor stærk effekten eller *hvordan* det tidlige testresultat påvirker det senere. Og det er det, sådanne regressionsanalyser skal bidrage med.

5.4 Kausale analyse ved hjælp af difference-in-differences metoder.

Difference-in-differences metoder er en elegant måde at estimere den *kausale* effekt af interventioner og hændelser ved hjælp af lineære modeller, der svarer til dem, vi lige har beskrevet. Metoderne har imidlertid de samme problemer som de øvrige, og målinger med fejl og bias pga regnefejl vil fordreje resultaterne på en vanskeligt gennemskuelig måde. Jeg vil nøjes med at referere til Callaway & Sant'Anna (2021) for at forklare, hvorfor disse analyser også har problemer. På side 8 i artiklen nævnes fire forskellige krav, der skal være opfyldt, for at man kan bruge metoderne.

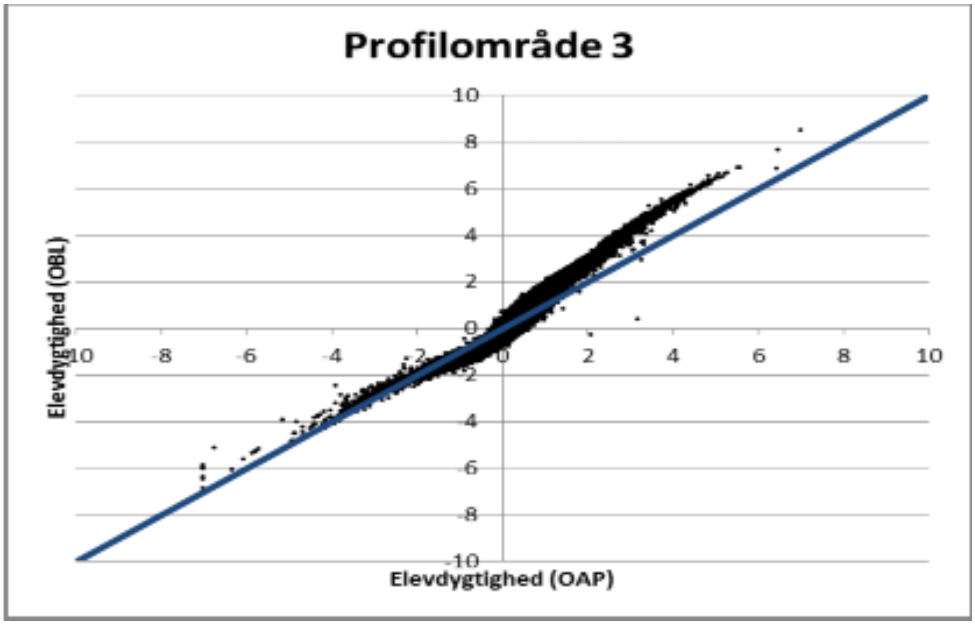
Den første er, at alle individer, der inddrages i analyserne, skal være uafhængige og identisk fordelte mht. alle undersøgelsens testresultaterne. I og med at undersøgelsen inddrager testresultater fra før 2014, hvor der stadig er regnefejl og testresultater fra efter 2014, hvor UVM har rettet fejlene, må vi konstatere, at denne forudsætning ikke er opfyldt.

Den anden er, at tendenserne i udviklingen af testresultaterne skal være den samme i samtlige perioder mellem to test vha. DNT. Da disse perioder involverer test fra både før 2014 og efter 2014, og hvor der er bias før, men ikke efter, kan DNT resultaterne ikke leve op til det krav. Det er blevet påstået, at det elegante difference-in-differences design garanterer, at resultaterne af analyserne er robuste over for regnefejlene i DNT før 2014. De forudsætninger, som analyserne bygger på, understreger, at det er de ikke.

5.5 Bias i fejlberegnede DNT-resultater.

Problemet med risikoen for bias i de fejlberegnede DNT-resultater er vigtigt, fordi bias betyder, at relationerne mellem testresultater indbyrdes og i forhold til andre variable ikke vil være lineære, selvom kravet om linearitet er opfyldt for de sande værdier af færdighederne.

STIL (2020) giver ikke et fuldstændigt svar på dette spørgsmål men viser dog, at der er tale om bias i visse, lad os sige, ikke-ligegyldige situationer. Figur 4 findes på side 61 i rapporten. Den viser sammenhængen mellem estimaterne af dygtigheden i tekstforståelse i 8. klasse 2018, sådan som DNT beregnede det ud fra de forkerte sværhedsgrader (OAP), og dygtigheden beregnet ud fra estimaterne af sværhedsgraderne i de obligatoriske test i 2018 (OBL).



Figur 4. Sammenhænge mellem mål for læsefærdighed (Tekstforståelse) i 8. klasse i 2018 målt sådan som DNT beregnede det (OAP) og sådan som man ville beregne det, hvis man havde benyttet de rigtige sværhedsgrader (OBL)

Man skal være svagsynet, hvis man ikke kan se, at sammenhængen mellem de to målinger ikke er lineær. Korrelationen mellem de forkerte og korrekte mål for læsefærdigheden er naturlig vis høj, men det er ikke korrelationerne, der er det vigtige. Det interessante er hvor store forskellene på de to målinger er, og om forskellene afhænger af, om der er tale om mere eller mindre dygtige læsere.

Det er tydeligvis tilfældet i dette eksempel. Omkring logit-værdier på 0 er der ingen systematiske forskelle på de to målinger, men blandt de svageste og især blandt de bedste læsere kan man se en systematisk tendens til, at DNT undervurderede elevernes læsefærdigheder. Der er naturligvis grænser for, hvor meget man kan aflæse af sådanne figurer, men så vidt jeg kan se, ville DNT undervurdere de dygtigste læsere med omkring en hel logit. Og en hel logit er udtryk for en meget stor forskel på to logit-værdier der burde være lige store.

Ud fra disse resultater og ud fra det det, som jeg i øvrigt kan læse i STIL (2020), er der stor grund til at forventer at DNT-resultater havde problemer med bias før fejlene blev rettet, og at analyser baseret på lineære modeller derfor kan føre til misvisende resultater. Hvor vidt de 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.

6. Afsluttende kommentar

Elementære regressionsanalyser af resultater fra pædagogiske test kan have problemer, fordi der er tale om testresultater med indbyggede (men forhåbentlig usystematiske) fejl. Der er imidlertid flere forskellige måder, man kan håndtere sådanne problemer på. En af dem – men langt fra den eneste – ville være at benytte den form for plausible værdier, som PISA, TIMMS og PIRLS benytter sig af, og som i øvrigt synes at være standarden i seriøse analyser af pædagogiske testresultater. Disse metoder ville ikke afhjælpe problemerne med regnefejlene i DNT, men når disse fejl er rettet, for alle de år hvor der foreligger resultater, er det jo ikke længere et problem, vi behøver at tale om.

Det undrer mig derfor, at de fire artikler, som jeg har læst, helt ignorerer, de problemer man har med usikre testresultater, og at de ikke engang har forholdt sig til de metoder til analyse af pædagogiske test som for eksempel PISA, TIMMS og PIRLS anvender. Men det er måske for meget at forlange. Analyserne bliver jo noget vanskeligere end analyser ved hjælp af lineære regressionsmodeller.

Det mindste, jeg som afslutning vil udtrykke et håb om, er, at man kan forstå, at analyser af standardiserede summer af standardiserede testresultater er den sikreste måde at analysere resultater fra pædagogiske test på, hvis man ønsker, at der er ikke er nogen, der får noget interessant og relevant at vide. Hvis forfatterne til de fire artikler havde undladt standardiseringerne og havde undladt at beregne vægtede gennemsnit af profilscores uden at gøre rede for vægtene og uden at kontrollere, at de forskellige profiler målte én og samme færdighed, ville resultaterne af deres analyser have vagt min interesse på trods af mine forbehold over for deres metoder.

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.

Bendixen, C & Kreiner, S. (red) (2009) Test i Folkeskolen. København, *Hans Reitzels Forlag*

Bundsgaard J & Kreiner S. (2019) *Undersøgelse af De nationale Tests måleegenskaber*. Aarhus Universitet, DpU.

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

Callaway, B. & Sant'Anna, P.H.C (2021) Difference-in Differences with multiple Time Periods. *Journal of Economics* 225, 200-230

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

Meiding J, Neubert K & Larsen R (2017a) *PIRLS 2016. Rapport*. Aarhus Universitetsforlag

Meiding J, Neubert K & Larsen R (2017b) *PIRLS 2016. Bilag*. Aarhus Universitetsforlag

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

Oplysninger og mange referencer om regressionsanalyser med variable, der er målt med fejl kan man finde på WIKIPEDIA under overskriften ”Errors-in-variables models”.