Hatties brug af effect size som grundlag for rangordning af pædagogiske indsatser

Peter Allerup
Professor ved Aarhus Universitet

Hattie anvender i sin store metaanalyse af metaanalyser størrelsen effect size til vurdering af, hvor effektiv en konkret pædagogisk indsats er i sammenligning med andre typer af indsatser. Denne artikel ser på effect size-beregningen gennem statistiske briller og svarer på, hvilke tekniske forudsætninger der skal være opfyldt, for at beregningen giver statistisk mening som et mål, der kan benyttes ved sammenligninger mellem pædagogiske indsatser.

Ved hjælp af effect size rangordner Hattie (2009) 138 typer af pædagogiske indsatser, som alle har til formål at øge elevernes præstationsniveauer. Rangordningen er velkendt og antages af mange som en praktisk reference for prioritering af mulige pædagogiske indsatser i situationer, hvor man som lærer eller skoleleder planlægger at foretage pædagogiske indsatser i undervisningen. Dermed kommer toppen af listen til at dominere diskussioner i ledelses- og læremiljøer, når snakken drejer sig om, "hvad der virker", og "hvad der ikke virker".

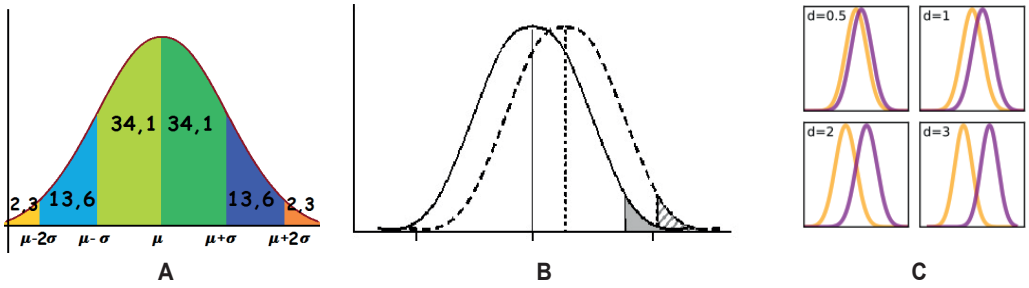
Der er ingen tvivl om, at Hatties rangliste, set gennem pædagogiske briller, rejser en lang række problemer i forhold til at forstå, hvad der egentlig ligger bag udsagnet om, at noget "virker". Problemet vil ikke blive diskuteret i denne sammenhæng, ligesom Hatties egne analyser, der konkret har ført ham til at lave listen, heller ikke er genstand for diskussion i denne artikel. Den kritik, som rejses nedenfor, er derfor ikke en kritik af listen eller den pædagogisk analytiske baggrund for den. Hensigten med artiklen er som nævnt at se på effect size-beregningen gennem statistiske briller og klargøre, hvilke tekniske forudsætninger der skal være opfyldt, for at beregningen giver statistisk mening som et mål, der kan benyttes ved sammenligninger mellem pædagogiske indsatser.

Diskussionen kommer til at foregå med inddragelse af elementære statistiske begreber som "middelværdi" og "spredning" samt grundbegreber fra statistisk testteori gennemført på verbalt niveau. Kendskab til matematiske formler er således ikke nødvendigt for at kunne læse artiklen. Artiklen henvender sig til de læsere, der er fascinerede af Hatties rangliste og måske spørger sig selv, om det nu også er den endelige sandhed, som kommer frem ved at rangordne "initiativer til fremme af elevpræstationer" ved hjælp af en effect size-beregning.

Hvordan beregner man effect size?

Når effect size udregnes, er der altid tale om at sammenligne tal fra før og efter en pædagogisk indsats – altså sammenligninger mellem *to* fordelinger.

Problemet er illustreret i figur 1B, hvor nogle elevers præstationer før og efter en pædagogisk indsats er illustrerede. Der er tale om præstationer fra elever, som før indsatsen ligger og svinger omkring μ_1, og som efter indsatsen er flyttet til at svinge omkring μ_2. De to fordelinger har samme spredning – "bredde" – og er placeret lidt forskudt i forhold til hinanden.



Figur 1: Figur A viser sammenhæng mellem procentafskæringer og spredningen σ i en normalfordeling, figur B antyder en forskydning på μ_1-μ_2 mellem to fordelinger, og endelig viser figur C forskellige forskydninger med dertil beregnede effect size-mål.

Umiddelbart er forskellen mellem før og efter altså afstanden mellem μ_1 og μ_2, det vil sige μ_1-μ_2, og det kunne være fristende at lade denne differens være et simpelt udtryk for, hvor effektiv indsatsen har været. Det er dog intuitivt klart, at "bredden" af fordelingerne, den såkaldte spredning, ikke er uden betydning for vurderingen af, om μ_1-μ_2 er stor eller lille. Med meget brede fordelinger, det vil sige store spredninger, udgør differensen μ_1-μ_2 jo kun en lille del i forhold til de tilfældige udsving, som er målt ved hjælp af spredningerne. Samme størrelse kan derimod se stor ud, hvis de to fordelinger er meget smalle, det vil sige har små spredninger. Ved at sammenligne med figur 1A kan man få en fornemmelse af, hvor stor forskellen μ_1-μ_2 egentlig er, udmålt i "spredningsenheder" (σ/sigma). Forskellen kan man skønne til at ligge på lidt under 1σ ved at se på figur 1A og 1B. I fortsættelse af den intuitive fornemmelse af, at spredningen bør inddrages i vurderingen af forskellen μ_1-μ_2, er det derfor ikke overraskende, at man som effektivitetsmål faktisk definerer effect size som forskellen μ_1-μ_2, udregnet i relation til den grundlæggende spredning (σ): Effect size = $(\mu_1 - \mu_2)/\sigma$.

I figur 1C antydes en forskel mellem μ_1 og μ_2 som – ved at sammenligne med figur 1A – er af størrelsesordenen ¾ σ. Det betyder, at effect size i dettte tilfælde er lig med $(\mu_1-\mu_2)/\sigma$ = ¾ σ, eller σ = 0,75. Udregnede effect size-værdier, som overstiger 5,00, er sjældne.

Vurdering af effect size som stor eller lille

Den numeriske værdi af effect size bruges til at afgøre, om interventionen har "virket". Det er klart, at jo større effect size er, jo større er forskellen mellem de to fordelinger, jævnfør figur 1C. Men er den kritiske grænse for, "at det virker" lig med 0,2, 0,4, 2,3, eller skal man anvende en anden værdi som skæringsværdi mellem "ikke virkningsfulde" og "virkningsfulde" interventioner? Jacob Cohen (1992) var en amerikansk statistiker, der introducerede begrebet "effect size", som efterfølgende fik betegnelsen "Cohens d". Han anvendte i praksis en tredeling af værdiområdet for effect size: "none", "small", "medium" og "large" ud fra værdierne 0,2, 0,3, 0,5 og 0,8. Effect sizes under 0,2 er derfor ubetydelige, mens effect sizes over 0,8 anses for "store".

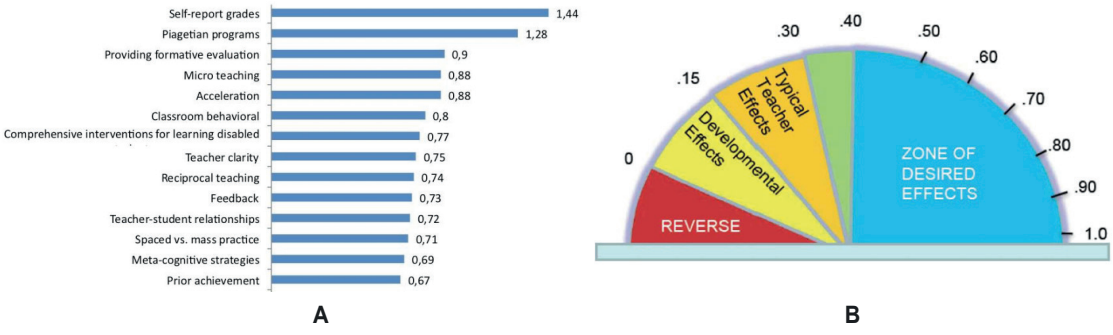
Illustrationen i figur 1C angiver, hvordan de to fordelinger ligger forskudt fra hinanden ved værdier af effect size på 0,5, 1,00, 2,00 og 3,00. At grænserne

ikke skal opfattes alt for rigide, ses af Cohens medfølgende bemærkninger:

> "The terms 'small,' 'medium,' and 'large' are relative, not only to each other, but to the area of behavioral science or even more particularly to the specific content and research method being employed in any given investigation. (…) In the face of this relativity, there is a certain risk inherent in offering conventional operational definitions for these terms for use in power analysis in as diverse a field of inquiry as behavioral science. This risk is nevertheless accepted in the belief that more is to be gained than lost by supplying a common conventional frame of reference which is recommended for use only when no better basis for estimating the Effect Size index is available." (Cohen 1992)

Mange anvender effect size som kriterium for, om den analyse, hvori beregningen indgår, er værd at inkludere for eksempel i en metaanalyse. I *Visible Learning* (Hattie 2009) er der medtaget studier med stærkt varierende værdier. Hattie vurderer selv i sit kompaslignende instrument, som er gengivet i figur 2B, at værdier over 0,4 er "ønskelige", mens værdier under 0,15 tilskrives betegnelsen "effekter som kan udvikles".

Cohen, Hattie og mange andre har et ønske om at tildele udregnede værdier af effect size nogle fortolkninger, som kan bruges direkte ved vurderingen af "om interventionen har virket", det vil sige er signifikant. Umiddelbart er dette *ikke* muligt, fordi formlen effect size = $(\mu_1-\mu_2)/\sigma$, set gennem statistiske briller, ikke genkendes som en statistisk størrelse, der kan underkastes sædvanlige signifikansvurderinger. Den præcise grund til, at der overhovedet er plads til en diskussion for eksempel om grænsen for, "det virker helt sikkert" skal ligge på 0,5 (Cohen), eller 0,4 (Hattie) er, at hvis man multiplicerer effect size = $(\mu_1-\mu_2)/\sigma$ med kvadratroden af antal observationer ($t= d\sqrt{n}$), får man faktisk en egentlig statistisk teststørrelse for hypotesen $H:\mu_1=\mu_2$. Altså en test for, at de to fordelinger har samme middelværdi. Afhængigt af antallet af observationer (n) slår man op i en t-tabel og kan her finde de kritiske værdier, hvor man forkaster hypote-



Figur 2: Toppen af Hatties liste med pædagogiske indsatser, rangordnet efter størrelsen af effect size. Til højre Hatties markering af betydningen af en udregnet effect size.

sen H. Dette sker, når t overstiger en vis værdi – altså bliver for stor. Hvis vi for eksempel har 25 observationer bag udregningerne af μ1, μ2 og σ, kan det vises, at grænsen ligger på t=1,708 eller t=2,060 afhængigt af, om der kan argumenteres eller ej for, at den implementerede intervention under ingen omstændigheder kan gøre tingene værre – kun bedre. Bruger vi værdien 2,060 (svarende til, at man er åben over for, at interventionen eventuelt kan vise sig at være "negativ"), fører et lille regnestykke os til:

$$t = d\sqrt{n} = 2{,}060 \Leftrightarrow d = 2{,}060/\sqrt{25}$$
eller
$$d = 0{,}4120$$

Dette passer meget godt med Hatties brug af 0,40 som nedre grænse for "det ønskværdige".

Set gennem rent statistiske briller undersøges alene, om t-værdien er større end 2,060 eller ej, og konklusionen lyder: "der er statistisk signifikant forskel" eller "det kan ikke afvises, at de to middelværdier er ens" afhængigt af om t er større end 2,060 eller ej. Der foretages ikke yderligere vurderinger som "meget signifikant" eller "meget lidt signifikant". Det ligger uden for den statistiske terminologi.

Det sker dog oftere og oftere, at denne klassiske afgørelse mellem "at afvise" eller "acceptere" en hypotese erstattes af udregningen af en såkaldt *signifikanssandsynlighed* (p), der konkret vurderer, hvor stor sandsynligheden er for at få en "endnu mere afvigende størrelse på t", som data aktuelt giver anledning til. Det er efterhånden blevet kutyme at lade denne p-værdi være selve kernen i resultatformidlingen af den statistiske analyse af forskellen mellem de to fordelinger. En lille udregning ved hjælp af en tabel over t-værdier og Hatties grænser på barometret viser, i det konkrete eksempel med n=25-observationer, følgende sammenhæng (se tabel 1) mellem p-værdier og Hatties grænser i barometret:

Effect size er altså ikke en statistisk størrelse, som kan underkastes statistisk vurdering med hensyn til størrelse. At være "lille" eller "stor" er derfor en subjektiv vurdering. Men den ligger tæt på en gængs statistisk t-test-størrelse til vurdering af forskellen mellem to fordelinger, *hvis man inddrager antallet af observationer (n)*.

Flere statistiske forudsætninger bag effect size

I figur 1 og efterfølgende fortolkninger af "stor" versus "lille" effect size (tabel 1) er der stiltiende gået ud fra, at en række statistiske forudsætninger er opfyldt. Det gælder eksempelvis, at de underliggende fordelinger enten helt eller i det mindste approksimativt kan beskrives ved hjælp af pæne symmetriske normalfordelinger. Hvad sker der, hvis denne forudsætning ikke er opfyldt? I figur 3 er medtaget mere eller mindre "skæve" fordelinger som eksempler på ikke-normale fordelinger. I disse fordelinger kan man i princippet

Hatties grænser	0,00	0,15	0,30	0,40	0,50	0,60	0,70	0,80	0,90	1,00
p-værdi	p=1,00	p=0,44	p=0,16	p=0,05	p=0,02	P=0,01	p=0,001	p<0,001	p<0,001	p<0,001

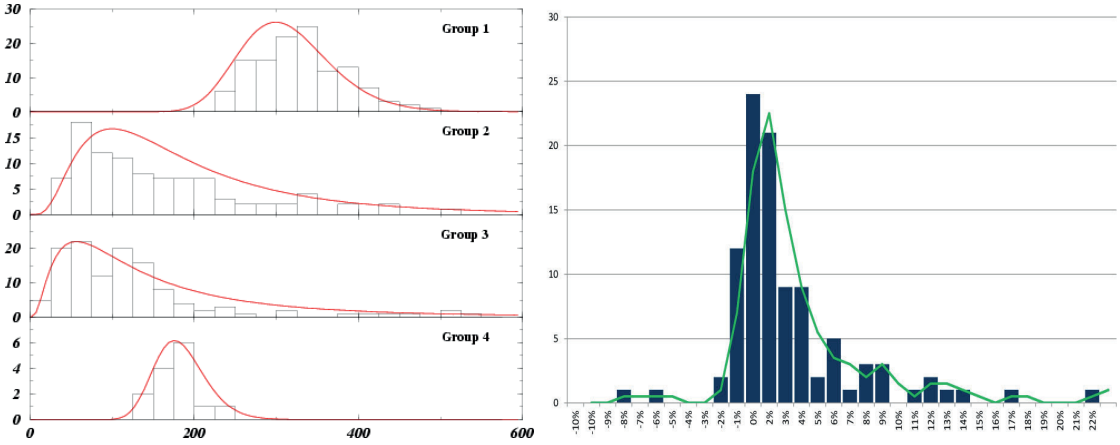
Tabel 1: Hatties grænseværdier for effect size sammenholdt med traditionelle p-værdier fra t-test-sammenligninger af to fordelinger med n=25-observationer.

regne effect size ud med den formel, som er anført ovenfor. Men dels betyder spredningen ikke det samme som tidligere beskrevet, og dels får man store problemer med at lade forskellen mellem udregnede gennemsnit fortolke som en "forskydning", blandt andet fordi det kan vises, at i "højreskæve" fordelinger ligger middelværdien μ ikke "i midten" af fordelingen, men ude til højre. Skæve fordelinger er enten højreskæve typer, som er gengivet i figur 3, eller venstreskæve, hvor "puklen" står til højre. I begge tilfælde er det et problem for både beregning og fortolkning af de udregnede effect size-værdier. Skævheden lægger hindringer i vejen for en simpel fortolkning af effect size, og man ved ikke, hvordan en given udregning af effect size skal forklares.

I det følgende gennemgås forudsætninger om ikkeskæve fordelinger, konstant spredning på elevmålene og brug af normalfordelingen. Hvilken betydning har det for validiteten af effect size som mål for effektivitet?

Et kik på figur 1 afslører en bestemt baggrund for fortolkningen af effect size, idet det stiltiende forudsætter, at de to fordelinger har *samme* værdi af *standardafvigelsen* σ. De to fordelinger er lige brede. Hvis de to fordelinger faktisk har *forskellige* standardafvigelser, opstår der direkte og indirekte problemer. Ét af de direkte problemer bliver tydelige, hvis den ene fordeling har meget lille spredning i forhold til den anden og i øvrigt ligger helt inden for den anden fordelings ydergrænser. Så vil det være vanskeligt eller nærmest umuligt at tale om en "forskydning" målt ved differensen μ_1-μ_2, og derfor mister udregningen af $(\mu_1$-$\mu_2)/\sigma$ selvfølgelig også mening. Det kan i denne forbindelse nævnes, at den viste korrespondance til t-testet tillige falder væk, fordi en statistisk forudsætning for t-testet er, at de to spredninger er lige store.[1] I hvilket omfang information om ens eller uens standardafvigelser har været tilgængelig for Hattie ved udarbejdelsen af hans analyser fra grundstudierne til *Visible Learning*, er ikke klart, men det er

1 Problemet går inden for matematisk statistik under navnet Fisher Berens problem.



Figur 3: Fem eksempler på "skæve" fordelinger.

ikke sjældent, at man observerer en større spredning af resultaterne *efter* en intervention i forhold til *før* interventionen, fordi selve interventionen har haft en ikke-ensartet indflydelse på elevernes præstationer. Interventionen har givet anledning til en større spredning mellem eleverne *efter* interventionen i forhold til, hvad den var *før* interventionen. For at effect size-udregningen skal kunne forstås som en simpel forskydning, skal standardafvigelserne i før- og efterfordelingerne derfor være ens.

En vigtig forudsætning for traditionel statistisk signifikansanalyse[2] er, at man kan benytte normalfordelingen som statistisk model for observationerne. Ved de internationale evalueringer under OECD (PISA) og IEA (TIMSS og PIRLS[3]) er de grundlæggende observationer ikke selv normalfordelte, og det gælder heller ikke antallet af rigtige besvarelser, elevscorerne. Efter transformation til de såkaldte Rasch scores (Allerup 1994, 2012), placeres værdierne derimod på en fælles skala. Det er denne skala med værdien 500 i midten, som offentliggøres og er genstand for vurdering i offentligheden og blandt politikere. Skalaen i Hatties rangordning af effect sizes kan i princippet sammenlignes med den velkendte internationale PISA-skala. Blot er PISA's Rasch scores for forskellige lande skiftet ud med størrelsen af effect sizes for forskellige pægagogiske indsatser. Det er ukendt, i hvor høj grad Hatties præsentationer af resultater fra metaanalyserne har kunnet anvende normalfordelingen som et relevant grundlag, men fordelene ved at kunne gøre det er mange. Én af de mere praktiske er den måde, hvorpå PISA for eksempel præsenterer sine sammenligninger landene imellem. Som det kan ses af figur 4A's blå streger med sort midte (se side 48) er PISA's rangordningsværdier forsynet med vandrette usikkerhedsmål[4] på den værdi, som karakteriserer landet. Det er normalfordelingen, som er grundlag for disse usikkerhedsberegninger. I den tilsvarende rangtabel over Hatties effect sizes (figur 4B, side 48) er det alene værdierne, som er tilgængelige, og det er ukendt, hvor sikkert effect size-størrelsen er bestemt. Man har derfor ikke som ved PISA-rangordningen adgang til en erkendelse af, at to eller flere indsatser/lande ikke adskiller sig *signifikant* fra hinanden – selv om de numerisk set er forskellige. Når man betænker politikeres omgang med numerisk set forskellige resultater fra lande, som i virkeligheden ikke adskiller sig statistisk, ligger der en latent fare for at komme til at fremhæve nogle pædagogiske indsatser frem for andre på et forkert grundlag. Er der for eksempel statistisk signifikant forskel mellerm to naboværdier "0,71 feedback" og "0,72 teacher student relationship"?

Som man kan se, har det stor positiv betydning for forståelsen af effect size-udregningen, inklusive dens delkomponenter i form af gennemsnit og standardafvigelse, hvis man kan benytte normalfordelingen som reference for før og efter interventionsfordelingerne. Der findes en vej til at opnå "tilladelse" til at benytte normalfordelingen direkte i forbindelse med beregninger af effect size. Denne vej er, at man kan støtte sig til ét af to matematisk statistiske argumenter. Det ene argument er, at hvis de grundlæggende observationer $x_1,\ldots, x_n$ er normalfordelte, så er den udregnede effect size:

$$d = \frac{X_1 - X_2}{s}$$

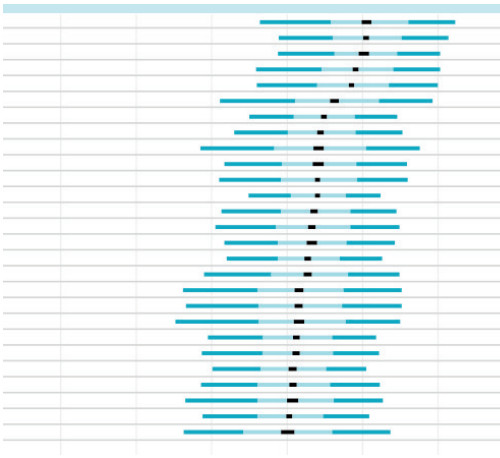
selv normalfordelt.[5] Det andet argument udnytter den såkaldte centrale grænseværdisætning, som siger, at selv om grundobservationerne $x_1,\ldots, x_n$ ikke selv

2 Altså analyser, hvor man kan udlede *signifikante* versus *ikke-signifikante* forskelle.

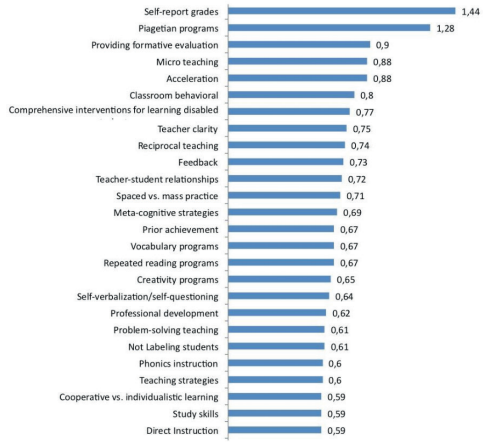
3 OECD's PISA tester 15-årige elever i matematik, naturfag og læsning. IEA's TIMSS tester 4. og 8. klasses elever i matematik, naturfag, og IEA's PIRLS tester 4. og 8. klasses elever i læsning.

4 Konfidensgrænser.

5 Tælleren er beregnede værdier af μ1-μ2, og nævneren (S) en beregnet værdi for σ.



A



B

Figur 4: Rangordning af normalfordelte gennemsnit (A) med usikkerheder indlagt som "boxplots" og toppen af Hatties liste (B) med rangordninger af pædagogiske indsatser ud fra størrelsen af effect size.

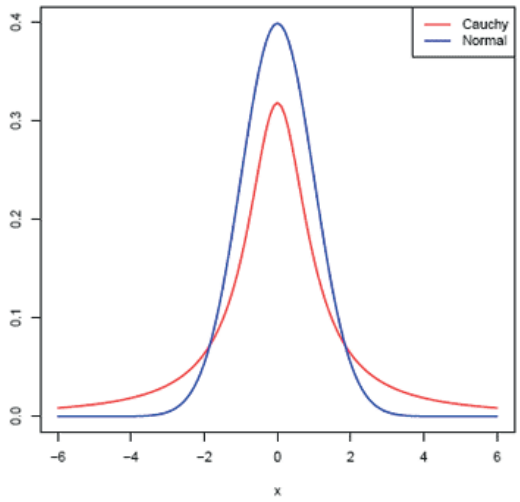
er normalfordelte, så er gennemsnittene i tælleren af det alligevel tilnærmelsesvist. Den samlede størrelse d bliver altså alligevel "næsten" normalfordelt i dette tilfælde. I begge tilfælde kan man med baggrund i normalfordelingen tegne de vandrette usikkerhedsmål (konfidensgrænser), som er illustreret i figur 4A. Hattie har undladt at forsyne sine effect size-angivelser (figur 4B) med sådanne vandrette usikkerhedsmål – måske fordi det ikke har været muligt at anvende ét af de to argumenter ovenfor?

Men der kunne være en helt specifik grund til ikke at kunne anvende et af argumenterne ovenfor. Det kunne være, at de grundlæggende observationer $x_1,\ldots,$ x_n følger en bestemt *ikke*-normalfordelingstype-fordeling – for eksempel den såkaldte Cauchy-fordeling.

Cauchy-fordelingen (se figur 5) ligner normalfordelingen i udseende (Poisson 1824). Men Cauchy-fordelingen har lidt tykkere "haler" end normalfordelingen, hvilket betyder, at afvigende observationer ses hyppigere under Cauchy end ved normalfordelingen. Det afgørende i forhold til argumenterne ovenfor består i, at Cauchy *ikke* har en middelværdi og *ikke* har en varians, sådan som normalfordelingens μ og σ har. Det har som konsekvens, at udregnede gennemsnit *ikke* adlyder den centrale grænseværdisætning som nævnt ovenfor, hvilket i praksis vil medføre, at udregnede gennemsnitsværdier vil springe temmeligt vildt frem og tilbage over store afstande på talaksen! Derfor falder de *ikke* "til ro" omkring en værdi, sådan som den centrale grænseværdisætning ellers udsiger.

Når Hattie bemærker, at han engang imellem selv har undret sig over meget varierende effect size-værdier, kan det måske skyldes, at nogle af de grundlæggende fordelinger ligner Cauchy?[6] Det er bemærkelsesværdigt, at den undervisningsministerielle, autoriserede karakterfordeling fra 7-trins- og ECTS-skalaen: 2 (10%), 4 (25%), 7 (30%), 10 (25%) og 12 (10%) netop udtrykker en fordeling, som er hø-

6 Bemærkningen faldt ved et personligt møde i Horsens i maj 2014.



Figur 5: Cauchy-fordeling (rød) og normalfordeling (blå) til illustration af Cauchy-fordelingens bredere haler.

jere i halerne end forventet med en normalfordeling. Et godt spørgsmål er derfor, om der i Hatties lister og rangordninger indgår udregninger af effect sizes, som helt eller delvist er baseret på anvendelsen af præstationsberegninger fra ECTS-skalaen.

Marginale og multivariate analyser med effect size

Hatties rangordninger er resultatet af *marginale* analyser ud fra forskellige pædagogiske indsatser. Altså analyser og vurderinger, som ikke tager hensyn til andre indsatser end netop den indsats, som lige nu betragtes. Det er imidlertid velkendt, at analyser i den pædagogiske verden ofte kræver inddragelse af *fler*dimensionale (multivariate) analyser, altså analyser, hvor flere variable analyseres på én gang. Kravet er aktuelt for at få et retvisende billede af, hvad der faktisk "styrer" matematikscorens variation, og for at beskrivelsen bliver så præcis som muligt, det vil sige med så lille usikkerhed som muligt. Ved at inddrage flere variable sammen med en bedre statistisk model, opnås en mindre usikkerhed, målt ved σ. Denne reduktion af usikkerheden påvirker effect size i opadgående retning, fordi σ indgår i nævneren. Marginale analyser produceret *efter* inddragelse af flere variable, hvor man i lys af den flerdimensionale beskrivelse bevidst vælger *alene* at se på én af variablene, eksempelvis matematik-scores, vil derfor ofte fremstå med mindre varians (standardafvigelse) sammenlignet med marginale analyser, hvor man *ikke ser* på andre variable samtidig. Analyser med én eller flere variable har altså stor indflydelse på, hvor stor effect size bliver.

Det er ikke klart ud fra Hatties analyser, hvilke af dem der er produkter af marginale analyser, og hvilke der er fremkommet i lys af statistiske analyser med mange variable. Et konkret eksempel på, at effect size udregnes forskelligt med og uden brug af flere variable fremkommer ved at se på virkningen af lærervariablen "linjefagskompetence (ja/nej)".

I matematik viste TIMSS 2011 for 4. klasses elever (se tabel 2) en (signifikant) forskel på μ_1-μ_2 = 543-533 = 10 skalapoint. Med en standardafvigelse på σ=67 kan en effect size beregnes til 0.15.

Dette passer godt med andres beregninger, som fremviser værdier fra 0,05 til 0,16[7]:

TIMSS-beregningerne er udført med cirka 1.000 observationer, mens de sidste beregninger (se figur 6) skyldes Ahn og Choi (2004) og Hattie (2009) med et antal observationer, som ikke er kendte.

Som det antydes i TIMSS-tabellen (tabel 2) underviser linjefagslærere socioøkonomisk bedre stillede elever end ikke-linjefagslærere. Det hænger måske sammen med en mere fleksibel anvendelse af lærere med linjefagskompetence på store skoler sammenlig-

7 Øjensynligt passer tabel 1's effect size-værdier versus p-værdier ikke helt her. Det skyldes, at der i dette TIMSS-tilfælde er tale om cirka 4.000 observationer.

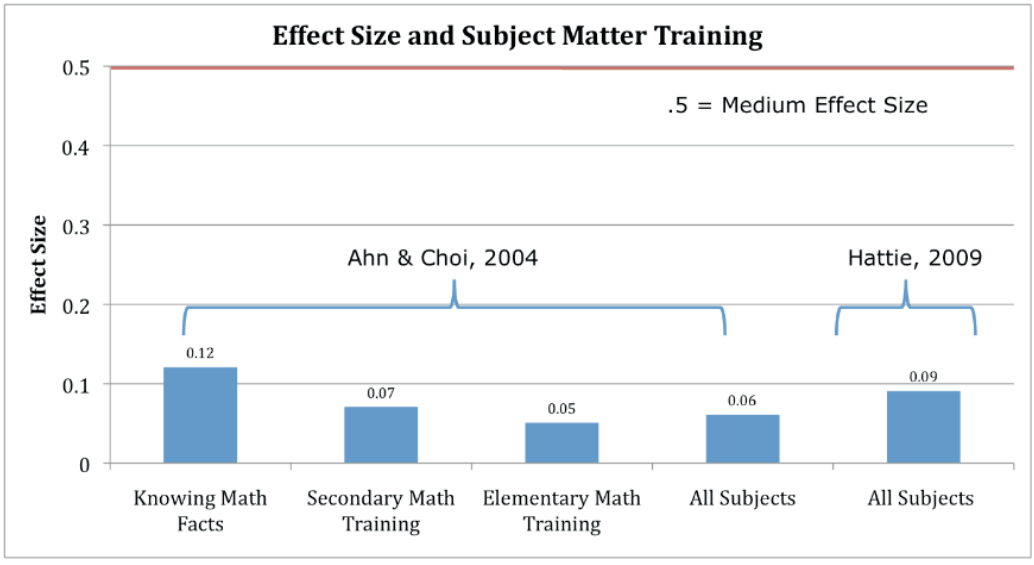
	Type lærer	TIMSS score (gennemsnit)	Socioøkonomisk score (gennemsnit)
Matematik	Linjefagslærer	543	2.99
	Ej linjefagslærer	533	2.91
Natur/teknik	Linjefagslærer	534	3.05
	Ej linjefagslærer	527	2.89

Tabel 2: Gennemsnitlige elevpræstationer fra TIMMS 2011 i matematik og natur/teknik samt deres socioøkonomiske gennemsnitstal.

net med små skoler, og fordi store skoler ligger i byer, hvor det socioøkonomiske niveau er højt. Det er derfor oplagt at benytte elevernes socioøkonomiske indeks som skjult tredjevariabel (en såkaldt co-variate[8]) i en ny sammenligning mellem grupperne, hvor der er kontrolleret for denne ekstra variabel (Allerup 2012). Gennemføres ovenstående beregninger af signifikans og effect size, i dette tilfælde som *kontrolleret effect size*, finder man, at der ikke er signifikant forskel mellem linjefags-elever og ikke-linjefagselever og den *kontrollerede effect size* beregnes til 0,08, altså halvt så stor som den *ukontrollerede effect size* på 0,15. Inddragelse af variablen "socioøkonomisk niveau" i en multivariat analyse fik altså effect size til at falde.

Denne nedslående fortolkning af linjefagskompetencens betydning for elevpræstationer kan skyldes (Allerup 2012), at de internationalt set meget standardiserede TIMSS-opgaver ikke udgør en stimulus over

8 Elevernes socioøkonomiske baggrund indgår i Hatties rangordning med effect size = 0,57.



Figur 6: *Effect size-beregninger i to studier af Ahn og Choi samt Hattie.*

for eleverne, som kræver "ekspertviden" fra eleverne, men alene trækker på færdigheder, som kan indøves af en hvilken som helst lærer. Over for dette forslag mener Hattie (2014), at det snarere skyldes lærernes manglende evne til at udnytte deres linjefagskompetencer på en måde, som kommer eleverne til gavn. Under alle omstændigheder udstiller disse betragtninger problemer med at acceptere en given effect size-udregning som noget endegyldigt.

Konklusion

Cohen udviklede i 1977 en størrelse, kaldet Cohens d eller effect size, som sammen med adskillige transformationer af den oprindelige version er skabt til at illustrere forskellen mellem to fordelinger. Set gennem statistiske briller kalder Hatties udstrakte anvendelse af effect size-beregninger ved rangordninger af pædagogiske indsatser på kommentarer fra to vinkler: Hvordan skal effect size fortolkes som begreb? Og hvordan skal selve størrelsen af effect size anvendes og fortolkes i praksis?

Der er i artiklen foretaget analyser, som viser, at en simpel fortolkning er tæt forbundet med antagelser om normalfordelte variable, og at der ikke skal store afvigelser til, enten i form af skæve fordelinger eller fordelinger (Cauchy), som har bredere haler end normalfordelingen, før det bliver vanskeligt at fastholde velegnede fortolkninger af effect size og den udregnede størrelse. Det andet spørgsmål belyses via begreber og teststørrelser, som i *matematisk formulering* ligger tæt op ad effect size, og som i konkrete analyser *med kendte antal observationer* gør det muligt at oversætte effect size-værdier til traditionelle p-værdier, som anvendes ved statistisk testteori.

Litteratur

- Ahn, S. & Choi, J. (2004): *Teachers' subject matter knowledge as a teacher qualification. A Synthesis of the quantitative literature on students' mathematics achievement.* AERS.
- Allerup, P. (1994): "Rasch Measurement, Theory of". I Husen, T. & Postlewaite, N. (red.): *The International Encyclopedia of Education*. 2. udgave. Pergamon Press.
- Allerup, P. (2012): *Danske 4. klasseelever i TIMSS 2011.* Aarhus Universitets Forlag.
- Cohen, J. (1992): "A power primer". I *Psychological Bulletin*, 112(1), s. 155-159.
- Hattie, J. (2009): *Visible Learning. A Synthesis of over 800 Meta-analyses Relating to Achievement.* Routledge.
- Poisson, J. (1824): "Sur la probalité des résultats moyens des observations". I *Connaissance des Temps pour l'an 1827.*