



## BBC 59: HVOR VITENSKAPELIG KAN VI EVALUERE VÅR BEHANDLING?

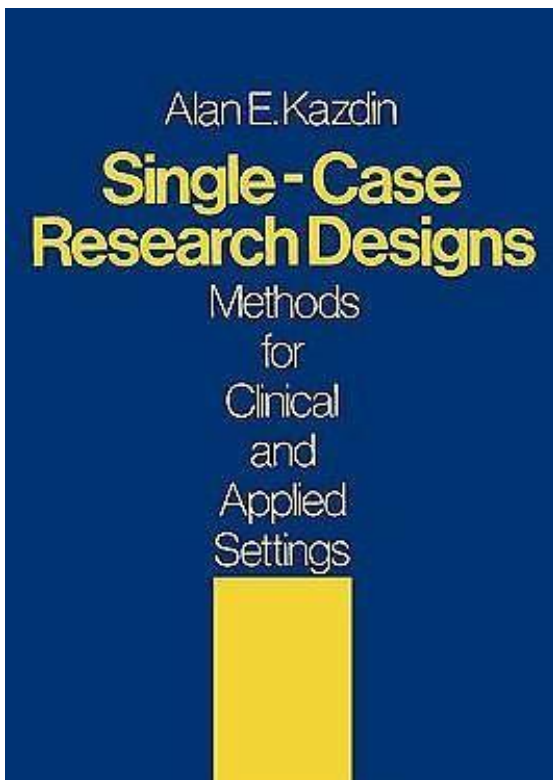
Vi atferdsanalytikere slår oss noen ganger på brystet over hvor vitenskapelig vår praksis er. Med «vitenskapelig» mener jeg da to forhold. Det første er at våre metoder generelt er godt dokumenterte, og at vi driver såkalt empirisk støttet behandling. Det drøftet jeg i BBC 42. Vi kommer bra ut av det, men ingenting er automatisk effektivt fordi det er atferdsanalyse. Det skal også anvendes riktig, og knapt noe som er noenlunde komplekst, lykkes alltid. Det andre er at behandlingstiltak evalueres slik at vi kan vite at det som tilsynelatende virker, faktisk virker. Det siste er et kvalitetsstempel ved atferdsanalyse. Atferdsanalyse har eksperimenter med enkeltindivider som sin viktigste forskningsmetode. Veien kan derfor være kort fra hvordan forskning utføres og evalueres til hvordan behandling utføres og evalueres.



Det er litt billig å si at det er lettere å påvirke atferd, og å påvise hvordan atferden påvirkes, hos dyr i et laboratorium enn hos mennesker i naturlige omgivelser. Vi trenger begge deler, om enn mest det siste. Men det er ikke til å komme forbi at en del eksperimentelle idealer hverken kan eller bør etterleves i forbindelse med behandling.

Det har flere følger for klinisk praksis. Ifølge «boka» skal definisjoner av atferd være så presise at det knapt er tvil om når atferd forekommer og ikke. For å oppnå det skal vi også beregne enigheten mellom uavhengige observatører. Dess større enighet, dess mer presis går man ut fra at definisjonen er, med noen forbehold som jeg kommer til. I klinisk praksis skal det også være en «design» som viser at det virkelig er tiltaket som skaper eventuell endring. En design, eller inndeling av faser som behandling og registrering av atferd skjer i, begynner med baseline. Baseline er som kjent registrering av atferd før tiltaket settes i gang, og regnes som selve grunnlaget for å se om et tiltak virker eller ikke. Vi bør registrere atferden lenge nok til at baseline er stabil. Hvis forekomsten av atferden varierer lite, kan det være nok med en uke eller to. Hvis forekomsten går opp og ned, må vi holde på lenge nok til å få med det

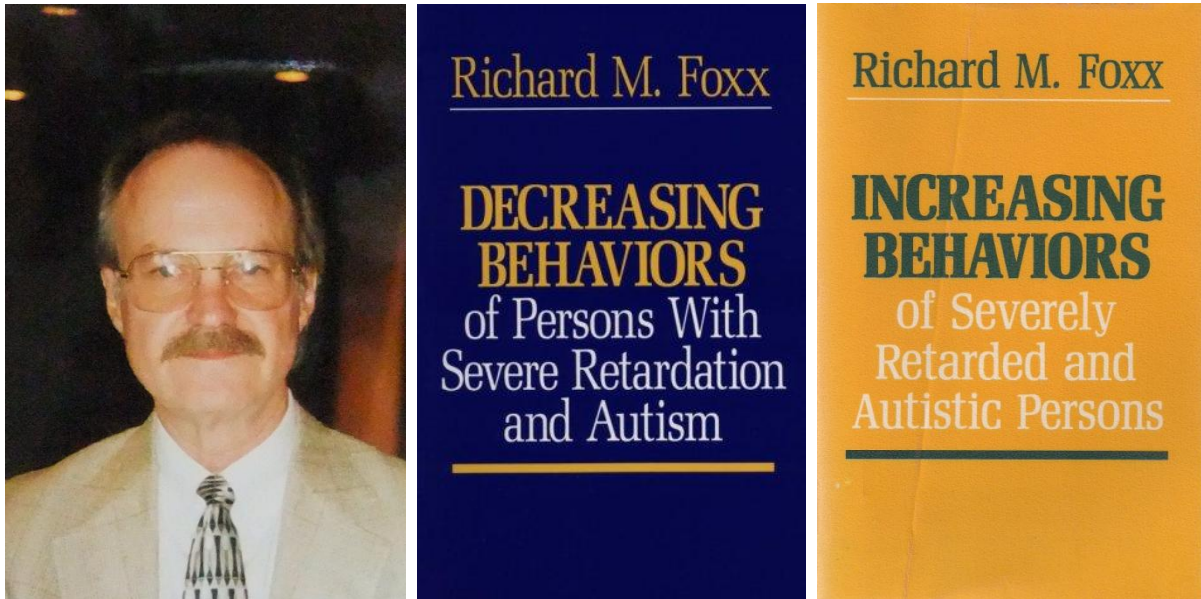
svingende mønsteret, som også er en form for stabilitet. Går baseline i ønsket retning, det vil si at problematferd synker, taler det for å vente med behandling eller at behandling i beste fall er unødvendig. En baseline i uønsket retning taler selvfølgelig for det motsatte. Designer som er spesielt egnet til å se om et tiltak virker eller ikke, er såkalt ABAB- og multipel baseline-design. I ABAB er A baseline- og B behandlingsfaser. Den andre A-en betyr at man midlertidig avbryter behandlingen for å se om endring som er oppnådd i behandlingsfasen, opphører. Bedring både i den første og andre B-fasen, og forverring i den andre A-fasen, er sterke holdepunkter for at tiltaket virker. Multipel baseline betyr at det er flere baseliner, ofte for flere problematferder, og at lengden på baselinene er forskjellige. Hvis det er tre atferder, iverksetter man først behandlingen overfor en atferd, mens baseline for de to andre fortsetter. Så iverksettes behandlingen også for den andre atferden, mens baseline for den tredje fortsetter. Endelig iverksettes behandlingen også for den tredje atferden. Behandlingen får altså tre «sjanser» til å vise at den virker. Hvis alle innfris, er det meget sannsynlig at behandlingen har noe for seg. Nok en design går ut på å endre innholdet i behandlingen. Hvis baseline etterfølges av for eksempel fire faser med forskjellige behandlingsbetingelser, har vi en ABCDE-design, der B, C, D og E står for hver sin behandling. Det kalles en flerbehandlingsdesign. Den enkleste designen er AB-design, der A er baseline og B er behandlingsfase. Men den har lav status, og en del atferdsanalytikere mener at den er for svak.



Alan E. Kazdins bok om enkeltkasus eksperimentelle designer har stått seg utrolig bra siden den først kom ut i 1987. Forfatteren selv (født 1945) ser ut som en ordentlig muntergjøk. Han er også svært allsidig, og har også skrevet om blant annet oppdragelse av opposisjonelle barn.

Høye idealer kan langt på vei skyldes at det aller meste av atferdsanalytisk forskning skjer under temmelig kontrollerte betingelser. Det gjelder ikke bare eksperimentell laboratorieforskning på mennesker og dyr som publiseres i *Journal of the Experimental Analysis of Behavior*, men også mer anvendte studier som publiseres i tidsskrifter som *Journal of Applied Behavior Analysis (JABA)* og *Behavioral Interventions*. I anvendte studier er det vanlig at behandlingen skjer over kort tid i en enkel avgrenset situasjon og ikke i personens naturlige omgivelser, at den dreier seg om en avgrenset problematferd, at den

utføres kort tid om gangen og over en kort periode, og at den utføres av eksperter, som jeg var inne på i BBC 38 og 42. Jeg har lenge ment at det bør lages et nytt nivå av anvendelse: I tillegg til det som hittil er kalt anvendt, og som ikke alltid er så anvendt, bør vi få et *virkelig* anvendt nivå, det vil si behandlingsopplegg av kjøtt og blod fra daglig praksis. Veteranen Richard M. Foxx har vært talsmann for å få fram det virkelige anvendte. Det er flere grunner til at det er lettere å evaluere systematisk i eksperimentelle og anvendte studier enn i daglig praksis: Man kan ta seg tid til en skikkelig baseline, og det er få eller ingen praktiske begrensninger eller etiske betenkeligheter med designer der behandling holdes tilbake. I vår daglige praksis er det imidlertid noen skjær i sjøen som gjør det verre å etterleve idealer.



Richard M. Foxx, fotografert av undertegnede i Chicago i 1997. "I'm truly honored", sa han! Ved siden av to av hans klassiske bøker. Foxx er en av de mest "hands on" av de mest kjente amerikanske atferdsanalytikerne. Han er opptatt av at behandling kan måtte foregå 24 timer i døgnet i mange år, og at det kan være et langt lerret å bleke. Han har vært president i ABA International, og har skrevet filosofisk stoff om blant annet forholdet mellom selvillit og selvrespekt. Han var NAFOs æresgjest til Nafos 25-årsjubileum i 1998, og holdt en meget god og underholdende tale.

For det første kan det faktisk være vrient å definere problematferden. Det kan være glidende overganger mellom at den forekommer og ikke. Hva er et slag, en dytt, et skrik, en trussel, et depressivt utsagn, en rømming og hva det skulle være? Personen kan også ha mange problematferder. Hvis en eller flere av dem er vanskelige å definere klart, er det ikke lett å få overblikk over situasjonen. Behandlingen kan også foregå lenge, kanskje mange år. Da endres trolig bildet. Ikke minst kan det oppstå ny problematferd, kanskje fordi det har blitt mindre av annen problematferd. «Gammel» problematferd kan bli borte uten at det trenger å skyldes behandling. Biologiske endringer, inkludert aldring, kan bli en faktor.

Jeg nevnte at det går an å beregne definisjoners presisjon et stykke på vei. Enkelt sagt går det ut på å se om man er enig i om problematferden forekommer og ikke. Vi kan måle «enighet mellom observatører», eller «interrater reliabilitet». En enkel metode er å se på hvor enige to observatører som er uavhengige av hverandre, er om det totale antallet ganger problematferden har forekommet i en periode. Hvis den ene mener 40 ganger, og den andre mener 50, deler vi 40 på 50 og får 0,8. Ganger vi det med 100, får vi 80 prosent enighet. Vi kan gjøre det samme på grunnlag av for eksempel atferdens varighet. Men metoden har særlig én stor svakhet: De to observatørene trenger ikke å være enige om en eneste observasjon. Begge kan ta grundig feil, men «enigheten» er stor hvis de tar like mye feil. Altså er metoden ganske grov. En annen mulighet er punktvis enighet. Den er egnet for atferd som kan

forekomme ved spesielle anledninger og tidspunkter, slik det ofte er i opplæring, eller i definerte intervaller. Da deles antall observasjoner som observatørene er enige om, på antall observasjoner de er enige om *pluss* uenige om, og finner prosentvis enighet. Metoden er mer presis, men uegnet for atferd som kan skje når som helst. Også her kan enighet bety at begge har observert feil.



Her scorer Leeds' Stephen Warnock på straffe mot Millwall. Som Millwall-fan er jeg skeptisk til at straffen var riktig dømt. Fotball inneholder utallige skjønsmessige og kompliserte dommeravgjørelser der mye spiller inn. Det ligner mer på registrering av atferd enn vi kanskje liker å tro.

I Norge er det uvanlig å måle enighet. Det burde ha vært vanligere, når de nevnte metodene passer. Det er nemlig liten tvil om at en del registreringer som gjøres, er det rene søppel: Registreringer kan være subjektive, og den som registrerer, har knapt satt seg inn i hva som skal registreres. Han eller hun *ønsker* kanskje mindre eller mer av atferden, og under- eller overregistrerer bevisst eller ubevisst. Oppfatningen av definisjonen av atferden kan endre seg, og terskelen for å registrere atferden kan stige eller synke. Noen går lei av å registrere, og noen venner seg til atferden og opplever den som et mindre problem. Begge deler kan svekke motivasjon for å registrere. Jeg har sett figurer med flotte kurver som opplagt er tøv, og som bør erstattes med generelle inntrykk og anekdotiske opplysninger.

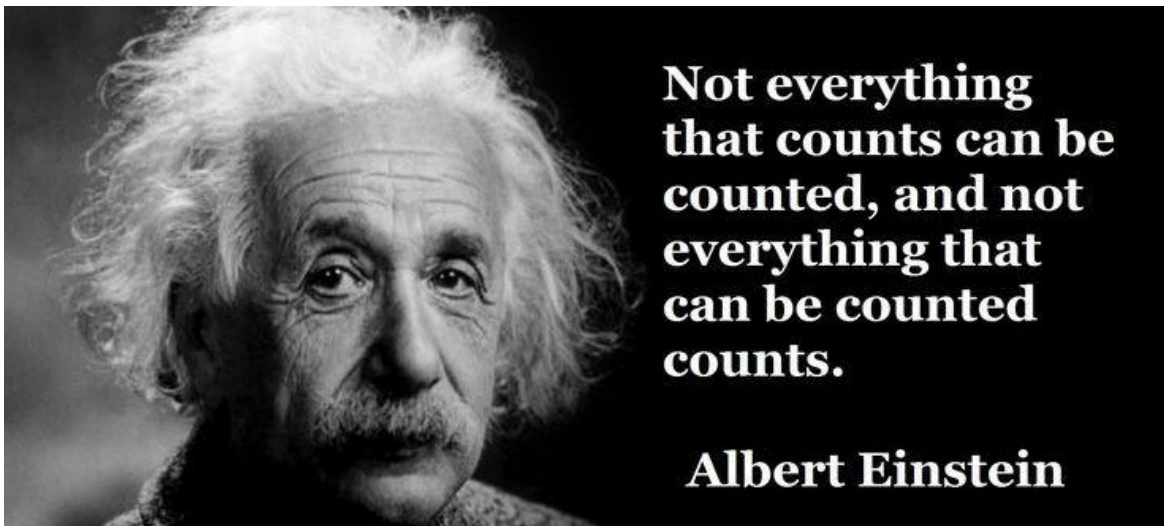
Registrering, særlig av to personer samtidig, er et ressursproblem. Uansett er det uforsvarlig å behandle problematferd uten en best mulig registrering, også fordi riktig observasjon av atferden er nødvendig for å utføre tiltak når de skal utføres, vel og merke hvis tiltak går ut på å forholde seg direkte til atferden. Noe av det lureste og mest økonomiske vi kan gjøre, er derfor å lage en maksimalt tydelig definisjon av atferden, som jeg skriver en del om i eMAAs Atferdsskole leksjon 21 (emaa.no). I beste fall er definisjonen så klar at selve definisjonen neppe kan være årsak til feil registrering. Andre ganger må vi innse at definisjonen må være relativt generell. For eksempel stemning eller angst er ikke så lett å lage entydige indikatorer på, som jeg også er inne på i leksjonen.

For det andre kan det være vanskelig å ta en skikkelig baseline. En grunn kan være at behandling av alvorlig problematferd må settes i gang raskt. Da må vi gjerne *anslå* baseline,

for eksempel omtrent hvor ganger per dag atferden har forekommet. Er vi heldige, gir baseline seg selv. For eksempel sjelden og alvorlig atferd blir ofte nedtegnet detaljert, slik at vi kan lage definisjoner i ettertid. Skolefravær skal framgå av skolens fraværsregistrering.



Eksempler på situasjoner som krever rask behandling, og der det vil være lite tilrådelig med en tradisjonell baseline i form av å la problematferden skje uhindret. Heldigvis går det også an å ta baseline av for eksempel forsøk, selv om det ikke blir helt det samme.



Kloke ord, men de misbrukes for ofte til å "underbygge" at knapt noe skal måles.

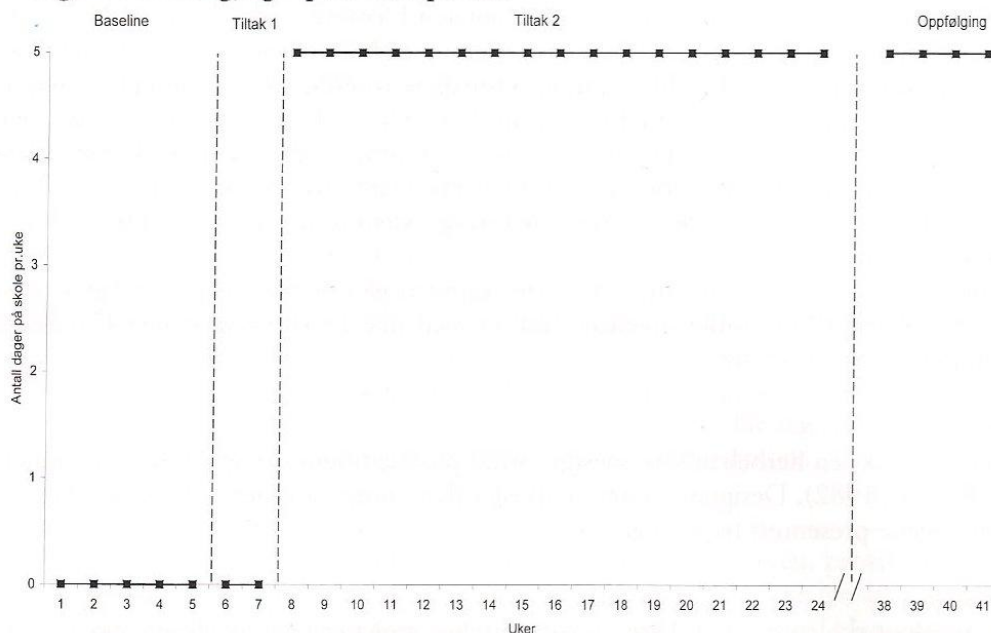
For det tredje kan det være gode grunner til ikke å bruke designer som sterkest kan påvise om tiltak virker eller ikke. Generelt er det uetisk å la være utføre behandling som hjelper, enten det skjer ved å vente med behandling, som i en multippel baseline-design, eller ved å avslutte behandling midlertidig, som i en ABAB-design. I verste fall kan fravær av behandling ha alvorlige følger. Vår oppgave er ikke forskning, men å sørge for at klienten har det bra. En enkel AB-design er derfor ofte det eneste mulige. Men hvis baseline er stabil eller går i uønsket retning, og det er en klar bedring helt fra starten av B-fasen, bør en AB-design være god nok dokumentasjon på at behandling virker. I tillegg er det altså som regel riktig å

fortsette med behandling som tilsynelatende virker, og uetisk å avbryte. Det blir som slagordet «Never change a winning team». Vi skal likevel ikke se helt bort fra ABAB-designen. Det hender blant annet at behandlingen må avbrytes på grunn av endrede rammer, ferier eller fordi man tviler på effekten av den, for så å starte den igjen. Heller ikke en flerbehandlingsdesign er uvanlig. Ofte må vi endre behandling for å lykkes. Da har vi automatisk flere behandlingsfaser, og dermed en flerbehandlingsdesign.

En annen ting er at behandling, i motsetning til de fleste anvendte studiene i Journal of Applied Behavior Analysis og lignende steder, sjelden består av enkeltkomponenter, men av behandlingspakker, det vil si flere tiltak samtidig. Da er det selvfølgelig vanskelig å se hvilken komponent som virker. Samtidig er det viktig å ikke utsette klienten, og for den saks skyld de som utfører behandlingen, for unødvendige tiltak. For å gjøre ting enklest mulig, er det derfor viktig å se hva som kan fjernes. Også da er vi fort inne på en flerbehandlingsdesign.

Evaluering er en balansegang mellom idealer og realiteter, som mye annet. Selv innser jeg at jeg ofte må gjøre en mellomting, selv om jeg også har hatt klienter med så alvorlig og sammensatt problematferd at knapt noen av idealene har kunnet oppfylles. Noen bryr seg for lite om evaluering. Jeg har møtt psykiatere, heldigvis bare noen få, som har avfeid spørsmål

Figur 1: Antall ganger på skolen pr. uke



Dette er en enkel design, men man skal være ganske «vrang» for å være uenig i at fase 2 av tiltaket virket. Det er for øvrig en flerbehandlingsdesign. Hvis vi ser bort fra den korte fase 1, er det en AB-design. Figuren er fra en artikkel i Norsk Tidsskrift for Atferdsanalyse om behandling av skolenekting av Jan-Ivar Sállman og undertegnede.

om evaluering med at ”Hvis medisinen virker, er det så opplagt at ingen er i tvil”. Ingen god holdning, særlig når effekten er svært usikker, og klienten blir stående på medisinen år etter år. Lignende utsagn har kommet fra atferdsanalytikere som hever seg over sunn evaluering og profesjonalitet. Men det er neppe et stort problem i dag. Til slutt nevner jeg at jeg i denne BBC-en har skrevet om behandling. Betrakningene er like aktuelle for opplæring.

Til slutt minner jeg om dagsseminar på Milepælen i Nord-Odal den 11. februar, om hvordan psykiatriske diagnoser kan bli mest mulig nyttige og minst mulig skadelige. Kom og treff atferdskolleger!

16. januar 2015

Børge Holden