Om metode: Bruk av inter-observatør enighet og inter-rater reliabilitet i NTA, og forslag til utvidelse av repertoaret i atferdsanalytiske studier

Anders Dechsling[1], Roald Øien[2,3], og Anders Nordahl-Hansen[1]
[1]Høgskolen i Østfold, [2]Universitetet i Tromsø, [3]Yale University School of Medicine

For at atferdsanalytiske studier skal inkluderes i evidensbasene må forskningsdesign og de tilhørende statistiske analyser være av god kvalitet. En kvalitetsindikator innen psykologiforskning er bruk og rapportering av enighet mellom observatørene under datainnsamling og analyser. Denne artikkelen beskriver omfanget av artikler i Norsk Tidsskrift for Atferdsanalyse som har metodedel, som oppgir å ha målt enighet mellom observatører, og som oppgir hvordan forfatterne har målt denne enigheten. Resultatene viser at de aller fleste artiklene med metodedel oppgir grad av enighet i form av prosentvis enighet. Vi gir en oversikt over og redegjør for ytterligere strategier for å måle observatørenighet og argumenterer for at atferdsanalytikere bør benytte seg av mer robuste statistiske mål for enighet mellom observatører.

Nøkkelord: Observasjon, Inter-observatør enighet, Reliabilitet, Cohen's Kappa, Intra class correlation, metode

---

Reichow et al. (2018) konkluderer med at det er svak evidens for at *Early Intensive Behavioral Interventions* (EIBI) er en effektiv behandlingsform for barn med autismespekterforstyrrelser (ASF). Konklusjonen baseres mye på forfatternes vurdering av designsvakheter. De argumenterer videre med at studiene må benytte seg av mer sofistikerte design for at man skal kunne ha større sikkerhet med tanke på slutninger man trekker i forbindelse med effekten av behandlingen for barn med ASF (Reichow et al., 2018). I en ny meta-analyse publisert i *Psychological Bulletin* (Sandbank et al., 2020) om tidlige intervensjoner for barn med autisme, konkluderes det med at tre intervensjonsprogrammer (Early Start Denver Model; ESDM, Joint Attention Symbolic Play, Engagement, and Regulation; JASPER, og Pivotal Response Treatment; PRT) viser lovende resultater fra deres studier. Disse intervensjonene omtales gjerne på engelsk som *Naturalistic Developmental Behavioral Interventions* (NDBI) og bygger på prinsipper fra både utviklingspsykologi og atferdsanalyse. EIBI evalueres til å ha "noe støtte" i litteraturen. En av årsaken til at EIBI evalueres med svakere evidensbase enn de såkalte NDBI-intervensjonene er blant annet fordi sistnevnte studier i større omfang benytter seg av design og passende analyser som gjør det tryggere å konkludere om effekter sett i intervensjonen faktisk har noe med intervensjonen å gjøre og ikke «støy» i form

Forfattermerknad: Takk til fagfellene for nyttige innspill til manuskriptet. Det er ingen konflikter mellom forfatterne med hensyn til dette manuskriptet. Korrespondanse vedrørende manuskriptet kan sendes til anders.dechsling@hiof.no

av feilkilder, systematiske og usystematiske målefeil (Sandbank et al., 2020).

Det er grunn til å ta resultater fra disse undersøkelsene på alvor dersom atferdsanalytikere skal påberope seg å drive med evidensbasert behandling. Dette er blitt påpekt tidligere blant annet av Arntzen og Løkke (2015) når det gjelder for eksempel beregning av effektstørrelse og statistiske analyser av data som supplement til visuelle analyser. Det er imidlertid viktig å poengtere at dette ikke kun gjelder for norske forhold, men berører også den atferdsanalytiske tradisjonen internasjonalt (Kratochwill et al., 2010; Shadish, 2014).

Inter-observatør enighet

Det er en rekke aspekter fra forskningsprosessens start til mål som påvirker kvaliteten til en studie. Her fokuserer vi på den delen som har å gjøre med påliteligheten til de observasjoner vi gjør i forskningsstudien, altså inter-observatør enighet (IOE, av noen også omtalt som mellom/fler-observatør enighet, inter-observer agreement) og inter-rater reliabilitet (IRR). IOE viser til enighet eller konformitet mellom observatører mens reliabilitet kan defineres som variasjonen i skårer fra ulike observatører for samme person delt på den totale variasjonen. Dermed vil IRR kunne si noe om muligheten for å differensiere mellom observatørene (De Vet, 2005). Observerbar atferd er en av grunnsteinene i psykologiforskning og kan valideres ved å måle enighet (IOE) eller variasjon (IRR) mellom observatører. I atferdsdefinerte diagnoser, som for eksempel autisme eller ADHD er både IOE og IRR essensielt blant annet når: diagnostiske kriterier defineres, ved etablering av beste standardprosedyrer for diagnostisering, ved utvikling av oppgaver for å utforske nye etiologiske teorier, og ved implementering og evaluering av evidensbasert behandling. Både IOE og IRR omhandler hvorvidt dataene og observasjoner som gjøres er til å stole på. Rasjonalen bak bruken av både IOE og IRR handler om å få en høyest mulig grad av objektivitet rundt våre observasjoner. Måling av enighet kan ta ulike former alt ut fra hvordan studien er lagt opp, men en nokså vanlig prosedyre er at en og samme enhet eller observasjon innen en studie skåres minst en gang av to eller flere ulike observatører. Observatørene skal skåre sine observasjoner separat og være «blinde» for den eller de andre observatørers skårer slik at de ikke påvirkes av hverandre i kodingssituasjonen. Imidlertid er det vanlig og anbefalt at observatørene i forkant, sammen med de andre forskningsmedarbeiderne, har samarbeidet og diskutert kodene som skal benyttes slik at en kalibrering seg imellom er sannsynlig. Dette kan blant annet innebære å kode liknende observasjoner til det som skal kodes i det faktiske studiet i fellesskap. Her er det også viktig å nevne at det som regel benyttes observatører som *ikke* arbeider direkte inn i prosjektet. For eksempel i de såkalte randomiserte kontrollerte studier (RCT) og andre typer gruppe-studier er det en nødvendighet at observatørene ikke kjenner til om de skårer et individ som tilhører intervensjons- eller kontrollgruppen i studiet. Kratochwill et al. (2010) påpeker at i så måte at man i single-case design og andre observasjonsstudier bør ha doble observasjoner på minimum 20% av observasjonene.

I bunn og grunn involverer observasjoner av enighet at man sammenligner en observatørs skåre med skårene fra en eller flere observatører. I de fleste tilfeller der forskerne er ute etter nøyaktige observasjoner benyttes vanligvis IOE. Man bør være oppmerksom på skillet mellom observatørenighet og reliabilitet selv om begrepene ofte brukes om hverandre (De Vet, 2005), og det er viktig å huske på at man må skille mellom enighet mellom observatører og hvorvidt observatørene har observert korrekt. Det er altså ikke gitt, selv om det er to ulike observatører, at skårene reflekterer objektets faktiske atferd (Kazdin, 2011). Noen ganger kan atferd ha blitt målt nøyaktig, men med lav enighet mellom observatørene. Andre ganger kan atferden ha blitt målt unøyaktig selv om det er høy enighet mellom observatørene.

Cooper et al. (2014) bruker begrepet *Interobserver Agreement* (IOA, interobservatør enighet [IOE] oversatt til norsk) og skriver at IOE er en av de vanligste indikatorene på kvalitet av observasjoner innen anvendt atferdsanalyse. IOE handler om hvorvidt to eller flere uavhengige observatører er enige om hva de observerer. En vanlig måte for å undersøke og rapportere observatørenighet i atferdsanalytiske studier er *prosentvis enighet,* eller *punkt-til-punkt likhet* (Cooper et al., 2007; 2014). Prosentvis enighet regnes ut ved at antallet observasjoner av den observatøren som har registrert færrest antall forekomster av en atferd, deles på antallet observasjoner fra observatøren som registrerte flest forekomster av atferden. Deretter ganges tallet med 100 for å få prosent. Punkt-til-punkt likhet innebærer at observatørene for eksempel måler samme respons uavhengig av hverandre, og så regnes enigheten ut ved at man deler antall responser det er enighet om på det samme antallet pluss responser det er uenighet om, og deretter ganges det med 100 for å få prosent enighet. Cooper et al. (2014) påpeker at det er flere måter å måle IOE på, men boka deres presenterer kun ulike varianter av prosentvis enighet. Videre argumenterer de for at dette er passende måter å måle på, og legger noen premisser for når, hvor ofte, hva som er kravet til god observatørenighet, og hvordan man bør rapportere IOE. De oppsummerer med at prosentvis enighet er den best egnede teknikken innen anvendt atferdsanalyse (Cooper et al., 2014).

Det finnes imidlertid andre måter å beregne IOE på og om prosentvis enighet er den som er best egnet bør det stilles spørsmålstegn ved. Hovedsakelig er ankepunktet mot å benytte prosentvis enighet at denne måten å beregne enighet på ikke tar hensyn til at det er sannsynlig at man i skåringssekvenser innimellom skårer ulike observasjoner likt ved ren tilfeldighet (Cohen, 1960). I forskningslitteraturen innen medisin, psykologi og ikke minst biostatistikk rundt temaet IOE er det bred enighet om at prosentvis enighet bør velges bort til fordel for mer robuste metoder for beregning av enighet mellom observatører (Cicchetti & Rourke, 2004; Cohen, 1960; Hallgren, 2012; Mitchell, 1979; Spitzer et al., 1967).

Vi har undersøkt hvordan observatørenighet blir beregnet og rapportert i studiene publisert i Norsk Tidsskrift for Atferdsanalyse (NTA). Videre presenterer vi alternativer til prosentvis enighet som ofte benyttes i en del vanlige design innen psykologisk forskning der to eller flere observatører eller observasjoner benyttes.

Metode

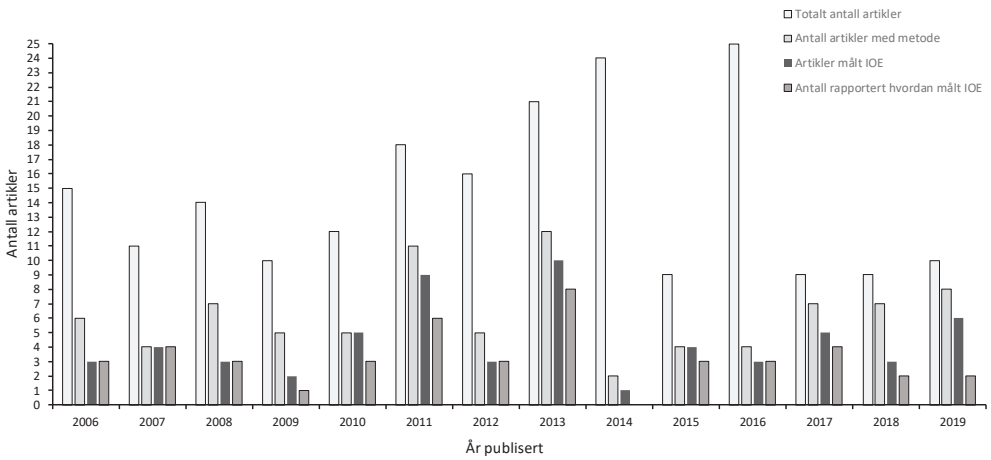
I desember 2019 gjennomgikk vi alle artiklene i alle utgavene på www.nta.atferd.no fra 2006 nr. 1 til og med 2019 nr. 2. En artikkel var registrert på nettsidene, men pdf var ikke vedlagt. Denne artikkelen fikk vi tilgang til ved å henvende oss til en av artikkelforfatterne. Deretter talte vi antall artikler i tidsskriftet som faller innenfor inklusjonskriteriene (se under), og antall artikler totalt. Vi gjennomgikk samtlige inkluderte artikler og undersøkte (i) om de rapporterte hvordan de målte IOE eller IRR, og (ii) og i så fall hvilken beregningsmetode som ble benyttet.

Inklusjons- og eksklusjonskriterier

Artiklene måtte være publisert i Norsk Tidsskrift for Atferdsanalyse på tidsskriftets nettside. Vi inkluderte alle empiriske artikler, litteraturgjennomganger eller andre artikler som hadde en metodedel. Konseptuelle artikler, dikt, og andre typer artikler ble ekskludert.

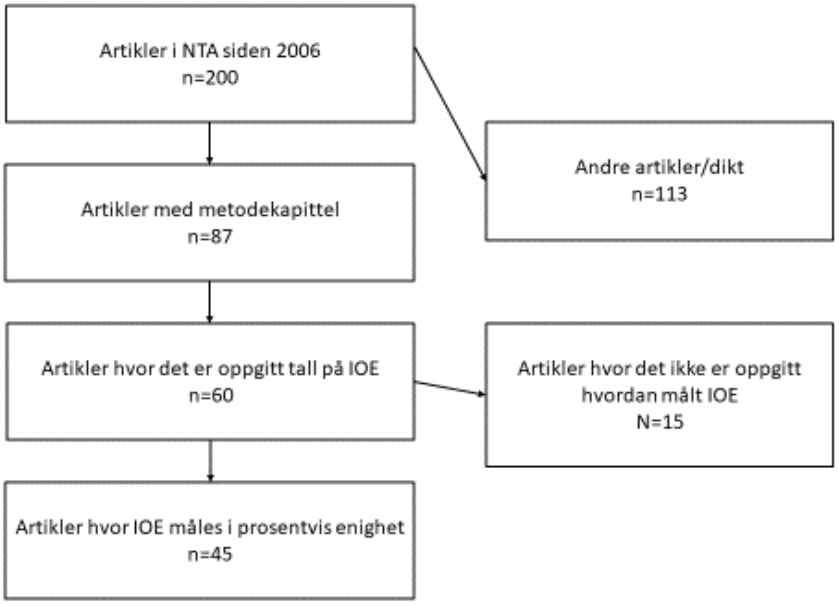
Resultat

I de 38 utgavene av NTA fra 2006–2019 er det totalt 200 publiserte artikler. Av disse var det 87 som ble vurdert til å oppfylle inklusjonskriteriene. Figur 1 gir en oversikt over antall artikler i NTA, antall artikler med metodedel, antall artikler som har målt IOE, og antall som har rapportert hvordan



Figur 1

Note. Figuren gir en visuell framstilling av antallet artikler publisert i NTA, og de øvrige stolpene illustrerer antallet artikler med metode, antall som har målt IOE, og antallet som har rapportert IOE.



Figur 2.

Note. Flytdiagram av gjennomgangen av studiene i NTA.

de har målt IOE. I 60 av de 87 inkluderte artiklene oppgir forfatterne å ha målt enighet eller reliabilitet i en eller annen form, og i 45 (75 %) av disse 60 artiklene oppgis det at de har målt IOE ved prosentvis enighet eller punkt-til-punkt likhet. I 15 av de 60 artiklene er det ikke oppgitt hvordan de har målt IOE (se flytdiagram i Figur 2). Under gjennomgangen indentifiserte vi flere ulike navngivinger av IOE, e.g., Inter-rated score, inter-rated reliabilitet, reliabilitet, mellom-observatør enighet (MOE), mellom-skårer

enighet (MSE), interobserver agreement (IOA), observatørenighet (OE), og intersubjektiv enighetsvurdering.

Diskusjon

Observatørenighet i NTA

Gjennomgangen av artiklene som oppgir hvordan de måler enighet mellom observatører viser at de fleste bruker prosentvis enighet som statistisk analyse. Løkke og Løkke (2006) er den første artikkelen i NTA som viser til hvordan måle IOE der de henviser til Kazdin (1982) når de beskriver fremgangsmåten. De øvrige studiene henviser som oftest til den tilsvarende strategien fra Cooper et al. (2007; 2014).

Cooper et al. (2014) nevner at det finnes flere teknikker for å måle IOE, men går ikke mer inn på disse. Derimot gjennomgår Cooper et al. (2014) forskjellige, men i hovedsak like, formler for å måle prosentvis enighet av ulike hendelser og atferder. Sett i lys av den atferdsanalytiske tradisjonen er det nok riktig å bruke denne strategien i flere av de aktuelle studiene publisert i NTA, men det blir problematisk dersom IOE kun måles på en måte – uavhengig av type studie. Det er hva man observerer som avgjør hvordan man bør måle observatørenighet (Kazdin, 2011).

De fleste artiklene i NTA beskriver heller ikke konteksten der IOE har blitt målt. Dette kan i enkelte tilfeller være en underrapportering av viktige metodiske faktorer. Det kan være helt nødvendig å beskrive konteksten IOE ble målt (Kazdin, 2011; Kratochwill et al., 2010). I for eksempel Dechsling et al. (2019) påpekes det at fordi observatørene var i samme rom og målte med stoppeklokke, så kan observatørenes atferd ha vært påvirket av hverandre og vært en forstyrrende variabel i resultatene fra IOE. Likeledes i Granmo et al. (2017) diskuteres det at selve forsterkerformidlingen påvirker de øvrige observatørene til å skåre likt. Disse variablene kan spille inn og dermed påvirke observasjonsdataene. Forskere bør i første omgang ta dette i betraktning under gjennomføringen av prosedyren, for så å beskrive (og/eller diskutere) målingen av IOE på en slik måte at det er mulig for leseren å vurdere hvordan IOE er målt.

Kazdin (2011) påpeker at det ikke er en avgjort konsensus om hvordan måle enighet i det anvendte feltet. Det finnes ulike strategier for å opparbeide samt å forsikre seg om at enigheten eller reliabiliteten blir god ved studiens start, og hvordan IOE kan opprettholdes og evalueres underveis i kodingsprosessen. Det krever at forsøket designes slik at det lar seg vurdere, og å tilrettelegge studien slik at det er mulig å vurdere statistiske elementer. Et steg i disse forberedelsene kan være at når det velges hvilken målbar atferdsdimensjon (se Cooper et al., 2014) som skal være avhengig variabel i studien bør man være oppmerksom på at operasjonaliseringen av atferden i stor grad vil avgjøre statistisk datatype eller målenivå på variabelen.

Atferd som statistisk målenivå

Innenfor statistikk opererer man gjerne med kategoriske (ikke-parametriske) og kontinuerlige (parametriske) variabler. Under kategoriske variabler finner man: (1) Nominale data, som gjenkjennes ved dataene kan kategoriseres i gjensidig utelukkende kategorier (to eller flere), hvor disse også er uttømmende (i.e., alle de aktuelle dataene faller inn under kategoriene). (2) Ordinale data hvor dataene kan rangeres (mer/mindre), og hvor de er gjensidig utelukkende, men det ikke er lik avstand mellom verdimarkeringene på skalaen. Av de kontinuerlige variablene finner man: (3) Intervalldata, hvor dataene kan rangeres, men det er lik avstand mellom verdimarkeringene på skalaen, noe som gjør at man kan vite mer om relative størrelsesforskjeller mellom verdiene på skalaen. Intervall-data har ikke et naturlig nullpunkt, som for eksempel IQ. (4) Ratio-data, også kalt forholdstallnivå, er data med lik avstand mellom verdimarkeringene på skalaen og hvor skalaen, og da fenomenet man måler, har et naturlig nullpunkt.

Cooper et al. (2014) gjør rede for en rekke målbare dimensjoner av atferd. I Tabell 1 har vi kategorisert noen av de under ulike nivå. Ut fra de ulike målenivåene kan man finne ut hvordan det er best egnet å måle IOE eller IRR. Selv om det er operasjonaliseringene av avhengig variabel som avgjør målenivået, så har vi her illustrert hvordan de ulike dimensjonene av atferd kan kategoriseres i ulike målenivå. Innenfor atferdsanalyse og de målbare dimensjonene ved atferd, så viser det seg at de fleste dimensjonene i Cooper et al. (2014) enkelt kan operasjonaliseres som kontinuerlige variabler. Når man måler enighet av data på forholdstallsnivå så skriver Kazdin (2011) at man kan bruke f.eks. prosentvis enighet eller *pearson product-moment correlation*. Problemet med å bruke disse er nødvendigvis ikke selve beregningsmetodene, men hvordan disse dataene tolkes.

Ved prosentvis enighet (Kazdin [2011] refererer til dette som *frequency ratio*) så måler for eksempel to observatører antall smil hos en person. Deretter deler man det laveste antall observerte forekomster på det høyest antall observerte forekomster, og ganger det med 100 for å få prosent. Denne målemetoden reflekterer nødvendigvis ikke antall forekomster observatørene var enige om, men enighet om et minimum antall forekomster. Vi kan i utgangspunktet ikke vite om de har skåret de samme forekomstene.

Ved *Pearson's product-moment correlation* sjekker man om observasjonene fra de ulike observatørene korrelerer med hverandre og vurderer høy korrelasjon som enighet. Det er en rekke kjente problemer med å bruke korrelasjonsdata for å trekke slutninger, også i måling av enighet ettersom det her kan være høy korrelasjon uten at dataene er knyttet til hverandre (Kazdin, 2011). For eksempel så vil data fra to ulike observatører i prinsippet kunne korrelere perfekt selv om observatørene konsekvent har skåret ulikt, men med en konstant differanse.

Kazdin (2011) hevder at de fleste bruker punkt-til-punkt likhet fordi det er en av

Tabell 1. Atferdsdimensjoner og statistiske variabler

Kategoriske variabler (ikke-parametriske)		**Kontinuerlige variabler (parametriske)**	
Nominal	Ordinal	Intervall	Ratio
Topography, yes/no	Scales	IQ-scale	Count
Event recording (discrete trial)	More than / Less than	Time (of day)	Rate/frequency
			Celeration
Partial-interval recording			Duration
Momentary time sampling			Latency
			Inter-response time
Measuring behavior by permanent products			Strength
			Whole interval recording

Note. Tabellen gir et forslag til inndeling av målbare dimensjoner av atferd under statistiske variabler.

de strengeste formene for å løse noen av utfordringene som for eksempel prosentvis enighet har. Ved punkt til punkt likhet måler man enighet respons for respons i stedet for å se på det totale antallet. Ettersom punkt til punkt likhet brukes mest i single-case design og direkte observasjon hevdes det ulike steder at cirka >80–90 % enighet er bra (Cooper et al., 2014; Kazdin, 2011), men det er viktig å påpeke at prosentandelen vil være påvirket av antallet observasjoner som er gjort. Ved høyfrekvent atferd så vil tilfeldighet spille en rolle ettersom sannsynligheten for «å treffe» blir høyere når frekvensen er høyere. Derfor bør man også ta høyde for slike tilfeldigheter i skåringer fordi man ofte vil forvente høy enighet. Det er en rekke måter å regne ut og korrigere for tilfeldigheter, som for eksempel å regne ut enighet ved tilfeldighet av både forekomster og/eller ikke-forekomster separat (se Kazdin, 2011), eller ved å bruke Cohen's Kappa.

Cohen's Kappa

Cohen's Kappa er en målemetode som brukes til å måle inter- og intraobservatør enighet, og som i tillegg tar inn i beregningene at observatører ved rene tilfeldigheter vil skåre observasjoner likt (Cohen, 1960; 1968). Cohen's Kappa har blant annet blitt brukt til å evaluere kriteriene i diagnosemanualen DSM og som veileder under diagnostisering (Klin et al., 2000; Lord et al., 1989; Taylor et al., 2017). Innflytelsesrike studier i utviklingspsykologi har brukt Cohen's Kappa for å utvikle atferdsmessige og eksperimentelle tester som har resultert i nye teorier om autisme (e.g., Klin et al., 2002; Leekam et al., 2002). Det å bruke Cohen's Kappa til å evaluere og rapportere enighet mellom observatører er nå standard i etablering av effektive behandlinger med innenfor-deltaker design og rapportering av endret atferd (Yoder & Symons, 2010). Fra et utbredelses og policy perspektiv, har Cohen's Kappa blitt brukt til å lede utviklingen av evalueringsmetoder av evidensbasert praksis for autisme (Reichow et al., 2008).

Cohen's Kappa tar utgangspunkt i at observatørene ikke er tilfeldig utvalgt. Dersom observatørene er tilfeldig plukket ut fra et utvalg med observatører, så anbefales Fleiss' Kappa i stedet (Fleiss et al., 1979). Cohen's Kappa (K) regnes ut ved å dele observert enighet mellom observatørene (P_o) minus en tenkt sannsynlighet for enighet ved tilfeldighet (P_e) på 1 minus en tenkt sannsynlighet for enighet ved tilfeldighet (P_e [e.g., tilfeldighet ved gjetning eller annet]). (Figur 3). Denne utregningen ligger klart til bruk og kan gjøres i de fleste statistikkprogram. En kappa på 0 vil indikere at enigheten er på ren tilfeldighet og 1 indikerer fullstendig enighet, 0–0.4 anses som noe enighet, >0.61 er betydelig enighet, og fra 0.81 og oppover er det nær fullstendig enighet. Det er også viktig å merke seg at hva som er akseptabelt nivå på enighet vil variere ut fra hva som studeres. For eksempel vil noe enighet kunne være godt nok i studier som måler effekt av trivselstiltak, men i en rekke medisinske studier så vil det være høyere krav til nivået av enighet og reliabilitet.

Tenkt sannsynlighet for tilfeldighet (P_e) kalkuleres ved å regne ut sannsynligheten for at observatørene har skåret en forekomst av atferd (for eksempel korrekte imitasjoner) ved en tilfeldighet. Dersom Observatør 1 (A) registrerer 36 av 40 korrekte imitasjoner og Observatør 2 (B) registrerer 30 av 40 korrekte imitasjoner, så vil man si at A har 36/40=0.9, og B har 30/40=0.75. Disse indeksene ganger man da slik 0.9 x 0.75 = 0.675, som blir den totale sannsynlighet for at begge har registrert imitasjoner tilfeldig. Deretter regnes det samme ut for ikke registrerte forekomster av imitasjoner. Det vil si at A har 4/40=0.10, og

$$K = \frac{P_o - P_e}{1 - P_e}$$

Figur 3.
Note. Formelen for utregning av Cohen's Kappa (K). P_o er den observerte enigheten mellom observatørene, og P_e er en tenkt sannsynlighet for enighet ved tilfeldighet.

B har 10/40=0.25. Disse ganges 0.10 x 0.25 = 0.025 sannsynlighet for observatørene ikke har registrert forekomst, tilfeldig. Disse legges til slutt sammen, P_e = 0.675 + 0.025 = 0.7. Ettersom begge er enige om 30 riktige og fire ikke korrekte imitasjoner så blir (P_o) = (30 + 4/40=0.85), og Cohen's Kappa regnes da ut ved K = (0.85 – 0.77) / (1 – 0.7) = 0.5. K=0.5 anses å være moderat enighet.

Som et eksempel på når punkt til punkt likhet er nyttig, så nevner Kazdin (2011) og Cooper et al. (2014) for eksempel discrete trials eller forekomst innen intervaller. I slike tilfeller mener vi at man bør se til de statistiske målenivåene og se at det er mulig å operasjonalisere variabelen slike at man får data på nominalnivå. I discrete trial så vil man skåre dikotome data (e.g., ja/nei kategorier), akkurat som i for eksempel forekomst av atferd innen en tidsintervall. Cohen's kappa vil være godt passende i slike studier. Til sammenligning med eksempelet i avsnittet over ville punkt-til-punkt likhet blitt regnet som 30/36 x 100 = 83,3 % enighet. Cohen's Kappa har korrigert for at denne enigheten kan være tilfeldig, og gir i så måte mer reliable data på IOE.

Intra-class correlation

En videreutvikling av Kappa, *Weighted Kappa* (Fleiss & Cohen, 1973) kan også benyttes for data som ikke er dikotome. Her vil vi i stedet presentere *Intra-class correlation* (ICC; Shrout & Fleiss, 1979) som en mulig beregningsmetode for data som er på ordinal, intervall-, og ratio-nivå. ICC benyttes hyppig i forskning for å estimere inter-rater reliabilitet mellom to observatører. Den kan også benyttes som et variasjonsestimat på intra-rater reliabilitet, altså når man undersøker grad av enighet én observatør har over to eller flere observasjoner. I tillegg kan ICC benyttes i spørreskjemaer for å vurdere test-retest reliabilitet. Det er en rekke forutsetninger som må være til stede for at man kan bruke ICC (cf., McGraw & Wong, 1996) og siden ICC er et korrelasjonsmål er overestimering også her en mulighet, men ICC gir anledning til å måle reliabilitet når det er en eller flere subjekter, eller to eller flere (vilkårlige og uvilkårlige) observatører (Nordahl-Hansen et al., 2013). Det som gjør ICC så anvendelig er at den er veldig fleksibel og kan tilpasses, men det er også det som gjør den kompleks og noen ganger vanskelig å forstå. Ulike studier vil bruke ulike varianter av ICC og det er designen på studien som avgjør hvilken variant, men alle har det samme fundamentet som bygger på at det finnes en sann score og en feilmargin (Hallgren, 2012). Utregningen av ICC kan virke noe komplisert, men gjøres enkelt i ulike statistikkprogram som for eksempel R eller SPSS (se Hallgren, 2012, for en oversikt).

Hvilken variant av ICC som passer til studien kan avgjøres i følgende trinn (se Hallgren, 2012; McGraw & Wong, 1996; Shrout & Fleiss, 1979). Først avgjøres om man skal bruke enveis- eller toveismodellen for analysen. Avgjørelsen tas på bakgrunn av hvordan observatørene velges ut. Dersom man har brukt en rekke observatører og plukker et tilfeldig utvalg av disse, så må man bruke en enveis analyse. Der man har bestemt spesifikke observatører så bruker man en toveis analyse for å korrigere for systematiske avvik hos spesifikke observatører.

Det neste trinnet er å definere hva som skal regnes som tilstrekkelig enighet for den aktuelle studien, og om det skal være i form av absolutt enighet eller som konsistente skårer. I atferdsanalytiske studier kan det se ut til at absolutt enighet er det naturlige valget ettersom konsistente avvikende skårer heller vil indikere dårlig operasjonalisering av målatferd. Konsistente skårer kan være ønskelig ved for eksempel bruk av Likert-skalaer. Det tredje trinnet innebærer at forskeren må spesifisere hvorvidt dataene skal baseres på gjennomsnittet av skårene fra flere av observatørene, eller være basert på skårer fra en enkelt observatør.

Det fjerde trinnet vil være å spesifisere hvorvidt observatørene anses som randomiserte eller fast effekt. I atferdsanalytiske studier vil man som regel velge en miks-

effekt modell fordi subjektene er tilfeldige, men observatørene faste. Her er det verdt å merke seg at dette trinnet ikke er viktig for selve utfallet av utregningen, men tolkningen av dette utfallet. Når man vurderer ICC, så vurderes det ut fra en indeks hvor 1 er full enighet, og 0 er at skårene er helt tilfeldig. I de tilfeller det er høy negativ ICC (f.eks. -0.80) så indikerer det en systematisk uenighet. Hallgren (2012) referer til den kjente biostatistikeren Domenic Cicchetti (1994) i hvordan man kan vurdere resultatene fra ICC, hvor ICC <0.4 anses som svak enighet, mellom 0.4 og 0.59 er grei enighet, 0.6 til 0.74 som god enighet, og >0.75 som utmerket enighet. Et av hovedpoengene med ICC er at man måler styrken på enigheten framfor «enig eller ikke-enig».

Veien videre

Det kan være ulike grunner til at prosentvis enighet brukes mest som beregningsmetode på IOE i NTAs studier. For det første er prosentvis enighet praktisk og lettfattelig. For det andre vies prosentvis enighet mye oppmerksomhet i lærebøkene og ser ut til å «ha satt seg» i feltet. En av hensiktene med denne artikkelen er å rette fokus og diskusjon rundt metodiske og statistiske anliggender som kan bidra til å styrke atferdsanalytiske studier.

Mer avanserte beregningsmetoder for enighet mellom observatører vil kreve kunnskap om variasjon og statistikk. I tillegg vil det i mange studier kunne kreve flere observasjoner for at det skal være mulig å ha tilstrekkelig konklusjonsvaliditet. Dermed stilles det høyere krav til fagpersonene og atferdsanalytikernes forståelse av statistiske forutsetninger og beregninger. Et økt fokus på dette kan bidra til utviklingen av mer robuste metoder innen fagfeltet. Arntzen og Løkke (2015) beskriver motstanden mot statistikkbruk i den atferdsanalytiske tradisjonen, men konkluderer med at opplæring og utdanning av atferdsanalytikere bør inkludere grunnleggende kunnskap om statistikk. Vi foreslår at også praksisen i studier inkluderer dette og vi vil påstå at allerede godt trente atferdsanalytikere har kapasitet til å opparbeide seg en bredere forståelse av statistikk.

ICC er et eksempel på en kompleks beregningsmetode som krever grundig forståelse av statistikk selv om statistikkprogrammene tar seg av selve utregningene, og dermed vil muligens Cohen's Kappa være et egnet steg på veien. Cohen's kappa vil og i mange tilfeller være det mest hensiktsmessige å benytte i blant annet N=1 studier. Det er ingen automatikk i at ICC kan brukes i alle tilfeller, men operasjonaliseringen av den avhengige variabelen og en vurdering av datatyper og trinnene som brukes i å beslutte varianten av ICC vil gi en pekepinn. Allikevel kreves det kompetanse for å forstå forutsetningene for analysen og resultatene av ICC. Det kan være problematisk at ICC benyttes av forskere uten tilstrekkelig forståelse av statistikk, og at dette medfører at det publiseres feil verdier for ICC (overestimater) i forskningslitteraturen. Med god nok kunnskap om operasjonalisering av atferd og statistikkbruk, så kan atferd i mange tilfeller operasjonaliseres som nominaldata. I disse tilfellene foreslår vi Cohen's Kappa som et bedre alternativ enn prosentvis enighet, men det vil alltid være typen av studie som avgjør beregningsmetoden.

Konklusjon

Flesteparten av studiene i NTA rapporterer enighet mellom observatører ved å beregne IOE ved hjelp av prosentvis enighet. Vi mener det er grunn til at atferdsanalytikere bør ha bedre kunnskap om hvilke ulike metoder for å forstå og beregne grad av enighet mellom observatører. Vår introduksjon i ulike strategier, som Cohen's kappa og Intra-class correlation, for å måle IOE og IRR bør gjøre atferdsanalytikere nysgjerrige og dermed kritiske når det utvikles nye forskningsdesign i forbindelse med kommende eksperimenter og studier. Vi mener at atferdsanalytikere må være spesielt nøye på bruk av metode og statistikk

dersom det er ønskelig at atferdsanalytiske studier og behandlingsformer skal inkluderes i evidensbasene i framtiden. God og riktig måling og rapportering av IOE og IRR er et steg på veien.

Referanser

Arntzen, E., & Løkke, J. A. (2015). Visuelle analyser av data — er det greit å ikke vite alt? *Norsk Tidsskrift for Atferdsanalyse, 42*(2), 97–105.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. https://psycnet.apa.org/doi/10.1037/1040-3590.6.4.284

Cicchetti, D. V., & Rourke, B. P. (Eds.). (2004). *Methodological and biostatistical foundations of clinical neuropsychology and medical and health disciplines*. CRC Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177%2F001316446002000104

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin, 70*(4), 213–220. https://doi.org/10.1037/h0026256

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied Behavior Analysis* (2nd ed.). Pearson/Merrill Prentice Hall.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2014). *Applied Behavior Analysis* (2nd ed.). Pearson Education Limited.

De Vet, H. (2005). Observer Reliability and Agreement. I P. Armitage & T. Colton (Red.), *Encyclopedia of Biostatistics*. https://doi.org/10.1002/0470011815.b2a04033

Dechsling, A., Larssen, L. M., & Herikstad, Y. (2019). Bruk av tegnøkonomi for å korte ned latenstiden etter friminutt for en elev med Downs syndrom. En systematisk replikasjon. *Norsk Tidsskrift for Atferdsanalyse, 46*(2), 65–69.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement, 33*(3), 613–619. https://doi.org/10.1177%2F001316447303300309

Fleiss, J. L., Nee, J. C., & Landis, J. R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological bulletin, 86*(5), 974–977. https://psycnet.apa.org/doi/10.1037/0033-2909.86.5.974

Granmo, S., Løkke, J. A., Halvorsen, L. R., Dechsling, A., Kvebæk, S., & Navestad, B. E. (2017). Preferansebasert mandopplæring for barnehagebarn med forsinket utvikling i barnehage. *Norsk Tidsskrift for Atferdsanalyse, 44*(2), 63–78.

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Kazdin, A. E. (2011). *Single-Case Research Designs. Methods for Clinical and Applied Settings* (2nd ed.). Oxford University Press.

Kazdin, A. E. (1982). *Single-case research designs. Methods for clinical and applied settings.* Oxford University Press.

Klin, A., Lang, J., Cicchetti, D. V., & Volkmar, F. R. (2000). Brief report: Inter-rater reliability of clinical diagnosis and DSM-IV criteria for autistic disorder: Results of the DSM-IV autism field trial. *Journal of Autism and Developmental disorders, 30*(2), 163–167. https://doi.org/10.1023/A:1005415823867

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry, 59*(9), 809–816. https://doi.org/10.1001/

archpsyc.59.9.809

Leekam, S. R., Libby, S. J., Wing, L., Gould, J., & Taylor, C. (2002). The Diagnostic Interview for Social and Communication Disorders: algorithms for ICD-10 childhood autism and Wing and Gould autistic spectrum disorder. *Journal of Child Psychology and Psychiatry, 43*(3), 327–342. https://doi.org/10.1111/1469-7610.00024

Lord, C., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L., & Schopler, E. (1989). Austism diagnostic observation schedule: A standardized observation of communicative and social behavior. *Journal of Autism and Developmental Disorders, 19*(2), 185–212. https://doi.org/10.1007/bf02211841

Løkke, G. E. H., & Løkke, J. A. (2006). Etablering av ballettdans ved hjelp Presisjonsopplæring (Precision Teaching). *Norsk Tidsskrift for Atferdsanalyse, 33*(3), 111–118.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about som intraclass correlation coefficiens. *Psychological Methods, 1*(1), 30–46. https://psycnet.apa.org/doi/10.1037/1082-989X.1.1.30

Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin, 86*(2), 376–390. https://psycnet.apa.org/doi/10.1037/0033-2909.86.2.376

Nordahl-Hansen, A., Kaale, A., & Ulvund, S. E. (2013). Inter-rater reliability of parent and preschool teacher ratings of language in children with autism. *Research in Autism Spectrum Disorders, 7*(11), 1391–1396. https://doi.org/10.1016/j.rasd.2013.08.006

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). Single-case designs technical documentation. *What Works Clearinghouse.* http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.

Reichow, B., Volkmar, F. R., & Cicchetti, D. V. (2008). Development of the evaluative method for evaluating and determining evidence-based practices in autism. *Journal of Autism and Developmental Disorders, 38*(7), 1311–1319. https://doi.org/10.1007/s10803-007-0517-7

Reichow, B., Hume, K., Barton, E. E., & Boyd, B. A. (2018). Early intensive behavioral intervention (EIBI) for young children with autism spectrum disorders (ASD). *The Cochrane Database of Systematic Reviews.* https://doi.org/10.1002/14651858.CD009260.pub3

Sandbank, M., Bottema-Beutel, K., Crowley, S., Cassidy, M., Dunham, K., Feldman, J. I., Crank, J., Albarran, S. A., Raj, S., Mahbub, P., & Woynaroski, T. G. (2020). Project AIM: Autism intervention meta-analysis for studies of young children. *Psychological Bulletin, 146*(1), 1–29. https://doi.org/10.1037/bul0000215

Shadish, W. R. (2014). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology, 52*(2), 109–122. https://doi.org/10.1016/j.jsp.2013.11.009

Shrout, P. E., & Fleiss, J. L. (1979). Intracalass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037//0033-2909.86.2.420

Spitzer, R. L., Cohen, J., Fleiss, J. L., & Endicott, J. (1967). Quantification of agreement in psychiatric diagnosis. A new approach. *Archives of general psychiatry, 17*(1), 83–87. https://doi.org/10.1001/archpsyc.1967.01730250085012

Taylor, L. J., Eapen, V., Maybery, M., Midford, S., Paynter, J., Quarmby, L., Smith, T., Williams, K., & Whitehouse, A. J. (2017). Brief report: an exploratory study of the diagnostic reliability for autism spectrum disorder. *Journal of Autism and Developmental Disorders, 47*(5), 1551–1558. https://doi.

org/10.1007/s10803-017-3054-z
Yoder, P., & Symons, F. (2010). *Observational measurement of behavior*. Springer Publishing Company.

On Methods: Use of inter-observer agreement and inter-rater reliability in NTA, and suggestions for expanding the repertoire in behavior analytic studies

Anders Dechsling[1], Roald Øien[2,3], and Anders Nordahl-Hansen[1]
[1]Østfold University College, [2]University Of Tromsø, [3]Yale University School of Medicine

To ensure that behavior analytic research will be included as evidence-based research, rigorous designs and statistical analyses is needed. One indicator of quality within psychological research is the measure and analysis of inter-rater reliability (IRR). This article provides an overview of the number of articles in the Norwegian Journal of Behavior Analysis that contain a method section, that reports data on inter-observer agreement (IOA), and how they have measured IOA. Our results show that most of these articles use percentage agreement. We provide an overview and introduction to other strategies of measuring agreement. Further, we argue that behavior analysts should utilize more robust statistical measures of observer agreement.

Keywords: Observation, Inter-rater reliability, reliability, Cohen's Kappa, Intra-class correlation, method