Tidsskrift for samfunnsforskning

Universitetsforlaget

VITENSKAPELIG PUBLIKASJON

Årgang 63, nr. 4-2022, s. 278–289
ISSN online: 1504-291X
DOI: https://doi.org/10.18261/tfs.63.4.3

Behandlingskontrast i eksperimenter innen velferd og utdanning

Treatment contrast in experimental research within welfare and education

Ira Malmberg-Heimonen
Professor, Institutt for sosialfag, OsloMet-Storbyuniversitetet
iram@oslomet.no

Anne Grete Tøge
Førsteamanuensis, Arbeidsforskningsinstituttet, OsloMet-Storbyuniversitetet
anneto@oslomet.no

Sammendrag

Denne artikkelen belyser metodologiske problemstillinger ved eksperimentelle studier på velferds- og utdanningsfeltet. Evalueringer med eksperimentelle design er stadig vanligere på samfunnsvitenskapelige områder, men de kan være utfordrende å gjennomføre. Intervensjonene involverer gjerne flere aktører. De har ulike nivåer av implementering og flere mekanismer som påvirker hverandre. Intervensjonene iverksettes gjerne i komplekse praksiskontekster og for heterogene målgrupper. Ved hjelp av eksempler fra vår eksperimentelle forskning drøfter vi spesielt behandlingskontrast, det vil si forskjellen mellom aktiviteter i eksperiment- og kontrollgruppen, og diskuterer implikasjoner og håndtering. Det er behov for mer eksperimentell forskning innen velferd og utdanning, men politikkutviklere og forskere bør være oppmerksom på at eksperimentelle design stiller betydelige krav til intervensjoner og kontekst. Tilliten til resultatene fra slike eksperimenter avhenger av hvor godt forutsetningene for å gjennomføre eksperimentet er ivaretatt.

Nøkkelord
Eksperimentelle design, randomisert, programteori, intervensjon, behandlingskontrast

Abstract

This article discusses methodological problems regarding the use of experimental designs within the fields of welfare and education. Experimental designs have become more common in evaluations within these fields; similarly, there are challenges involved in applying them. The interventions evaluated often involve several actors, include multiple levels of implementation and have multiple and interacting mechanisms. Interventions are also often implemented within complex practice contexts and for heterogeneous target groups. In the context of our experimental research, we especially discuss the challenges of a treatment contrast, that is the difference between activities in experimental and control groups, how to understand and tackle it. The article concludes that there is a need for experimental research within the fields of welfare and education, but that experimental designs require measurable interventions and contexts. Our trust in results from experiments depends on the extent to which the requirements are fulfilled.

Keywords
Experimental design, randomised, program theory, intervention, treatment contrast

Innledning

Politiske beslutninger og praksis bør være kunnskapsbaserte for å sikre best mulige offentlige tjenester for borgerne (Meld. St. 4 (2018-2019)). I utviklingen av kunnskapsbasert politikk og praksis er derfor spørsmål om effekt sentralt, og evalueringene av samfunnets innsatser for borgerne bør være mest mulig presise og nøytrale (Hjelmar & Pedersen, 2015). I denne sammenhengen blir eksperimentell forskning ofte beskrevet som en «gullstandard», et forskningsdesign som også omtales som en «kredibilitetsrevolusjon» for effektforskningen (Angrist & Pischke, 2010).

Grunnprinsippene i eksperimentene er enkle, der en kombinasjon av randomisering (det vil si tilfeldig trekking til en eksperimentgruppe og en kontrollgruppe) og manipulering (ulike grupper får ulikt tiltak) minimerer risikoen for å overvurdere effekter (Hutchison & Styles, 2010; Imbens, 2018). Innenfor velferd og utdanning er eksperimentelle forskingsdesign uvanlig (Breit et al., 2019; Fretheim, 2013; Pontoppidan et al., 2018), men det har vært en bevegelse mot økt bruk av eksperimenter de siste ti årene. Flere departementer og direktorater har i økende grad finansiert eksperimentelle evalueringer av intervensjoner i Norge fordi de ser dette som et steg på veien mot mer kunnskapsbasert praksis (Breit et al., 2019).

Innen felt som velferd og utdanning kan gjennomføringen av eksperimenter være utfordrende. Ofte gjennomføres eksperimentene i, eller på toppen av, ordinær praksis. Intervensjonene har virkningsmekanismer på flere nivåer, og flere aktører kan delta i implementeringen. Implementeringen preges også av læringsprosesser og tilpasning av intervensjonen til lokale kontekster, der tilpasningen ofte er en forutsetning for en varig praksisendring. Ofte kan også behandlingskontrasten være svak, det vil si at eksperiment- og kontrollgruppen i praksis mottar relativt lik behandling (Leko, 2015; Ling, 2012).

I denne artikkelen diskuterer vi utfordringer med eksperimentelle design som evalueringsmetode innenfor velferd og utdanning. Artikkelen begynner med en presentasjon av de mest sentrale prinsippene i eksperimenter: randomisering, manipulering og kontrafaktisk sammenligning samt en diskusjon om hvilke utfordringer som kan oppstå når eksperimenter brukes innen disse feltene. Behandlingskontrasten, det vil si forskjellen mellom det som tilbys i eksperiment- og kontrollgruppen, er kjernen i de fleste evalueringer av nye tiltak og reformer. Hamilton og Scrivener (2018) understreker at man derfor ikke kan komme utenom beskrivelser av behandlingskontrasten, og tilhørende fremgangsmåter for å måle denne kontrasten, for å få en substansiell forståelse av hva målte effekter representerer. Spørsmålet er hvordan man skal gjøre dette i praksis. I denne artikkelen presenterer vi konkrete eksempler fra vår forskning og diskuterer betydningen av behandlingskontrasten for studiene og tolkning av resultater. Hensikten er å bidra til at politikkutviklere, praktikere og forskere får større forståelse for og mer kunnskap om betydningen av behandlingskontrast i eksperimentelle design innen velferd og utdanning, og på den måten bidra til økt forståelse for hvordan eksperimenter best mulig kan styrke kunnskapsbasert praksis.

Grunnprinsipper i eksperimentelle design

Personer som mottar en gitt intervensjon, og profesjonelle rundt dem, kan oppleve at intervensjonen er effektiv. Når man spør dem, kan de gjerne fortelle at intervensjonen har vært nyttig. Men opplevd nytte betyr ikke nødvendigvis at intervensjonen har ønskede effekter, slik som forbedret livssituasjon. Det er ikke nødvendigvis intervensjonen som har skapt en forbedret livssituasjon; personen kunne ha opplevd tilsvarende forbedringer uten intervensjonen. Det er derfor avgjørende å ha kunnskap om hvor bra en intervensjon er, sammen-

lignet med fraværet av den. Samtidig kan ikke en og samme person både utsettes og ikke utsettes for en intervensjon i ett og samme tidsrom.

Det eksperimentelle forskningsdesignet bidrar til å løse akkurat dette problemet ved å løfte spørsmålet til gruppenivå: Vil like grupper av individer utvikle seg forskjellig avhengig av hva de utsettes for? Ved å ta en stor gruppe individer eller enheter og tilfeldig trekke ut en del gjennom en såkalt randomisering, forventer man å få to (eller flere) grupper med tilsvarende kjennetegn. Jo flere enheter som randomiseres, jo mer sannsynlig er det at sammensetningen av gruppene er lik, både når det gjelder observerte (for eksempel andel kvinner, gjennomsnittlig alder) og uobserverte (for eksempel holdninger, preferanser og motivasjon) kjennetegn. Dermed oppstår muligheten for en kontrafaktisk sammenligning mellom en gruppe som utsettes for intervensjonen (eksperimentgruppe), og en tilsvarende gruppe (kontrollgruppe) som ikke utsettes for det. Ved å måle hvordan det går med de ulike gruppene når det gjelder utfallet man er interessert i, for eksempel inntektsgivende arbeid, er det mulig å beregne effekten av intervensjonen. Differansen mellom eksperiment- og kontrollgruppen etter eksponering for intervensjonen kan tolkes som effekter av intervensjonen (Flannelly et al., 2018). Randomisering og manipulering samt kontrafaktisk sammenligning er grunnprinsipper i eksperimenter og gjør det altså mulig å trekke kausale slutninger om effekter (Barton, 2000; Flannelly et al., 2018).

Eksperimentelle design innen velferd og utdanning

Eksperimenter som gjennomføres i laboratorier, kan ideelt sett holde omgivelsene konstante, for eksempel gjennom å eksponere deltakere for like mye lys eller lyd. Forskerne har da kontroll over konteksten eksperimentet gjennomføres i, og forskjellen mellom eksperiment- og kontrollgruppen kan dermed begrenses til at eksperimentgruppen eksponeres for den antatt virksomme behandlingen. Kausale spørsmål som dukker opp innen velferd og utdanning, for eksempel om identifisering og oppfølging av elever med risiko for frafall fører til at flere fullfører videregående skole, er det ikke mulig å få svar på ved hjelp av laboratorieforsøk. Dersom denne typen spørsmål skal besvares gjennom eksperimentell forskning, må det gjennomføres i praksisfeltet, for eksempel i skoler.

Intervensjoner innen utdanning og velferd har også ofte virkningsmekanismer på flere nivåer, og flere aktører kan delta i implementeringen (Funnell & Rogers, 2011). Tar vi utgangspunkt i en intervensjon for å bedre oppfølgingen av lavinntektsfamilier, kan intervensjonen virke på individnivå der koordinatorer følger familiene tett opp på ulike målområder. Den kan også virke på systemnivå, der koordinatorer, ledere og samarbeidspartnere koordinerer den systemorienterte innsatsen til familiene (Malmberg-Heimonen et al., 2018). Sammenlignet med laboratorieforsøk er det vanskeligere å identifisere nøyaktig hva som bidrar til effekter når man gjennomfører eksperimenter ute i praksisfeltet.

Det er flere som er kritiske til bruken av eksperimentelle design innenfor velferd og utdanning (Connolly et al., 2018). Deaton og Cartwright (2018) er tonegivende i kritikken og bekymret for at forskere innen disse feltene er for naive i møte med eksperimentelle design. Deres bekymring er hovedsakelig knyttet til heterogene målgrupper, det vil si målgrupper som består av enheter med nokså ulike kjennetegn, som øker risikoen for at randomisering ikke gir sammenlignbare grupper. Langt fra alle er enige med Deaton og Cartwright (2018). Både Imbens (2018) og Raudenbush (2018) mener at nettopp randomiseringen er selve styrken ved eksperimentelle design. Selv om man ikke alltid kan trekke konklusjoner om presise effektstørrelser, lykkes med å få til sammenlignbare grupper eller

får prøvd ut intervensjonen i kontrollerte omgivelser, gir eksperimenter et mer nøkternt bilde av tiltakseffekter enn mange andre forskningsmetoder.

For å kunne måle en effekt av en intervensjon gjennom eksperimentelt design innen velferd og utdanning må intervensjonen representere en tydelig kontrast til ordinær praksis. Dette kalles behandlingskontrast. Hamilton og Scrivener (2018) presiserer at analyser av behandlingskontrasten, altså hva man sammenligner intervensjonen med, er like viktig som å studere implementering og implementeringskvalitet. Hamilton og Scrivener (2018) peker også på at det å analysere behandlingskontrasten bidrar til å identifisere de spesifikke spørsmålene eksperimentet kan besvare, øke forståelsen av hvilke intervensjonselementer som bidrar eller ikke bidrar til effekter, identifisere elementer som er like i begge grupper, samt øke forståelsen for ulike effekter mellom enheter og subgrupper.

I mange tilfeller inneholder intervensjonene imidlertid elementer som allerede eksisterer i ordinær praksis, og kontrasten mellom intervensjon og ordinær praksis kan bli svak (Hamilton & Scrivener, 2018). Eksempelvis bruker NAV-veiledere ofte relasjonelle ferdigheter i det ordinære oppfølgingsarbeidet, og en intervensjon som inneholder relasjonelle elementer, vil da delvis ligne på den ordinære praksisen og ha de samme virkningsmekanismene (Tøge et al., 2020). Det kan også være etisk problematisk å *la være* å tilby kontrollgruppen tjenester de har rett til (Hutchison & Styles, 2010). Dermed kan kontrasten mellom eksperiment- og kontrollgruppen i mange tilfeller bli svakere enn forskerne ideelt sett kunne ønske seg. Det er altså helt sentralt å forstå hva det kontrafaktiske er, hva ordinær praksis inneholder, og hva forskjellen mellom ordinær praksis og intervensjonen som evalueres, er. Svak kontrast er imidlertid ikke alltid noe negativt. For eksempel kan vi lure på om det er bedre å benytte mer relasjonelle metoder i oppfølgingen av NAVs brukere. Selv om ansatte i NAV til en viss grad allerede benytter slike metoder i dag, kan hypotesen være at mer systematisk bruk av relasjonelle metoder vil gi bedre resultater for brukerne. Kontrasten er relativt svak, men vi forventer likevel bedre resultater sammenlignet med ordinær oppfølging. Det essensielle er at behandlingskontrasten er identifisert, målbar og i tråd med det man ønsker å måle effekter av.

En annen kilde til svak kontrast kan være smitte. Oftest er det ikke mulig å isolere kontrollgruppen (Campbell et al., 2012; Murray et al., 2004). I skolen vil det for eksempel være problematisk å randomisere noen elever til å få alternative undervisningsopplegg og samtidig unngå at elevene i kontrollgruppen blir påvirket. Elever og lærere snakker med hverandre på tvers av gruppene de er randomisert til. Kontrollgruppen er ikke lenger «ren»; den er smittet av intervensjonen som testes ut. I dette tilfellet oppstår en uønsket svekket behandlingskontrast. Vi sammenligner ikke med en reell ordinær behandling, men med en behandling som er delvis influert av intervensjonen vi ønsker å teste.

Det er i tillegg vanskelig å «*blinde*» deltakere for hvilken gruppe de er randomisert til. Elevene er ofte klar over at de er med i et eksperiment, og både vissheten om deltakelse og hvilken gruppe de tilhører, kan påvirke resultatene. Dette kan gjøre det vanskelig å tolke om eventuelle effekter av intervensjonen skyldes selve intervensjonen eller deltakernes forventninger og motivasjon knyttet til å være randomisert til den ene eller andre gruppen (Goodman et al., 2018). Observerte effekter kan dermed til en viss grad være drevet av hva deltakerne forventer av intervensjonen de får.

Videre kan også varierende eller manglende implementering i eksperimentgruppen lede til svak behandlingskontrast. Også i dette tilfellet oppstår en uønsket svekket behandlingskontrast. I en slik situasjon inneholder intervensjonen vi måler effekter av, ikke alle elementene, eller har ikke samme intensitet eller kontinuitet, som intervensjonsbeskrivelsen tilsier. I slike tilfeller kan det være fristende for forskere å ekskludere enheter eller individer som ikke har

gjennomført intervensjonen, men med en slik analyse av eksperimentelle data introduserer forskeren en potensiell partiskhet. Ved å ta ut de enhetene som implementerer svakt eller individer som i liten grad har deltatt, risikerer man å ta ut de minst motiverte fra eksperimentgruppen. Dermed er ikke lenger enhetene eller individene tilfeldig fordelt; det kan være andre kilder til forskjeller mellom eksperiment- og kontrollgruppen enn selve intervensjonen. For å unngå dette må alle enheter som er randomisert, inkluderes i analysene. Analysen gjennomføres basert på behandlingsintensjon, en så kalt *intention-to-treat*-analyse (Gupta, 2011).

Hva er da implikasjonene av svak kontrast, variabel implementering eller smitte? Finner man ikke effekter, men vet implementeringen generelt sett har vært svak, eller at det har vært smitte fra eksperiment- til kontrollgruppen, kan man ikke være sikker på at intervensjonen faktisk *ikke* har effekter (Funnell & Rogers, 2011). Det er en mulighet at den er effektiv, men ikke er blitt implementert tilstrekkelig, delvis implementert i både eksperiment- og kontrollgruppen, eller at behandlingskontrasten er for svak. Ved å undersøke behandlingskontrast og grad av implementering, og da helst på en måte som gjør det mulig å fange opp praksisutøvelse i både eksperiment- og kontrollgruppen, kan forskere undersøke om effektene har sammenheng med hvilke intervensjonselementer som er implementert, og få kunnskap om hvor nøyaktig og omfattende implementeringen har vært, samt de aktivitetene kontrollgruppen har deltatt i. Her bidrar kvalitative data til innsikt. Observasjoner og intervjuer gir verdifull informasjon som vil kunne belyse implementeringskvalitet, men også den ordinære praksisen, som intervensjonen sammenlignes med (Grissmer et al., 2009).

Noen eksempler fra velferds- og utdanningsfeltet

Som vi har diskutert, kan eksperimenter innen velferd og utdanning være utfordrende å gjennomføre i henhold til grunnprinsipper i eksperimentelle design. I det videre tar vi for oss fem eksperimenter vi har gjennomført. Vi løfter spesielt frem utfordringer knyttet til sammenligningsgrunnlaget, det vil si kontrasten mellom eksperiment- og kontrollgruppen, før vi diskuterer implikasjoner og håndtering.

Tabell 1 redegjør i korthet for fem av eksperimentene, og gjengir intervensjonen vi evaluerte, designet vi brukte, og hovedfunnene. Det er flere likheter mellom eksperimentene, men det er også noen forskjeller. Et av eksperimentene (nettverksråd) har individ-randomisert design der deltakerne ble randomisert, mens de resterende har klynge-randomisert design der skoler eller NAV-kontorer ble randomisert. To eksperimenter er fra utdanningsområdet (IKO- og LOG prosjektene), mens tre (nettverksråd, HPMT- og HOLF-prosjektene) er fra velferdsfeltet. I hvert av eksperimentene ble intervensjonen implementert i en praksiskontekst, enten dette var NAV-kontorer eller skoler. I intervensjonene inngår ulike nivåer og aktører. I samtlige evalueringer har vi kombinert ulike typer kvantitative og kvalitative data for å studere implementeringsprosesser og virksomme elementer. Vi har også brukt kvalitative data for å forklare effektfunnene.

Eksperimentene har alle sitt utspring i et behov innen politikk og/eller praksis. Forskerne har vært sentrale i utviklingen av forskningsdesignene, mens hvor mye ansvar de har hatt for utviklingen av intervensjonen, har variert. I LOG-prosjektet hadde forskerne en betydningsfull rolle i utviklingen av intervensjonen, mens forvaltningen hadde hovedansvar for utviklingen av intervensjonene innen HPMT- og HOLF-prosjektene. I IKO og Nettverksråd ble eksisterende intervensjoner justert i samarbeid mellom forskere og praksis og implementert i nye kontekster. I samtlige prosjekter har praksisfeltet hatt hovedansvaret for implementeringen og forskerne for evalueringen. Vår vurdering er at det er hensiktsmessig med et tydelig skille mellom implementerings- og evalueringsansvar. Uansett om

utviklingen av intervensjonen har vært styrt fra politikk og praksis eller fra forskere, har avklaring av roller og ansvar i forkant av prosjektet bidratt til at de ulike aktørene har hatt en felles forståelse av intervensjonen og målsettingen med den.

Tabell 1. Eksperimenter innen velferd og utdanning.

Navn og område	Intervensjon	Design og målgruppe	Hovedfunn
Nettverksråd (2007–2011)[a]	Familieråd fra barnevernet brukt i en voksen kontekst (nettverksråd) med mål om å mobilisere deltakernes private og profesjonelle nettverk.	Individrandomisert design, der 149 langtidssosialhjelps-mottakere ble randomisert til eksperiment- og kontrollgruppe.	Resultatene viste positive kortidseffekter på mental helse og sosial tilhørighet, mens langtidseffektene var svake. Kvalitative data viste at årsaker til svake langtidseffekter var nettverkets manglende ressurser og sosialtjenestens manglende oppfølging av deltakere i etterkant av nettverksrådet.
Helhetlig oppfølging av deltakere i kvalifiseringsprogrammet – HPMT (2010–2014)[b]	Helhetlig, prinsippstyrt metodisk tilnærming (HPMT-modellen) med mål om økt arbeidsdeltakelse. Ble implementert innen rammen for kvalifiseringsprogrammet.	Klynge-randomisert design med 18 NAV-kontorer, der ni kontorer ble randomisert til å implementere HPMT-modellen.	Positive effekter av HPMT-modellen på deltidsarbeid for deltakere i kvalifiseringsprogrammet. Positive effekter på de profesjonelles opplevelse av kompetanse. Kvalitative data viste at de profesjonelle verdsatte fleksibiliteten i modellen.
Helhetlig oppfølging av lavinntektsfamilier – HOLF (2016–2020)[c]	HOLF-modellen, en modell for helhetlig oppfølging av lavinntektsfamilier med mål om å øke arbeidsdeltakelsen, forbedre familienes økonomi og boligsituasjon samt barnas situasjon. HOLF-modellen ble implementert i NAV-kontorer.	Klynge-randomisert design der 29 NAV-kontorer deltok. Alle kontorene ansatte koordinatorer. 15 kontorer ble randomisert til å implementere HOLF-modellen, mens de resterende kontorene utviklet lokale praksiser for familieoppfølging. Målgruppen var langtidsmottakere av sosialhjelp med hjemmeboende barn. Koordinatorene trakk ut familier fra målgruppen og gav dem tilbud om oppfølging.	Svak tendens til økt overgang til arbeid som følge av koordinatorrollen, men effekten var ikke statistisk signifikant. Analysene viste ingen signifikante effekter av selve HOLF-modellen på de fire målområdene sammenlignet med de lokale praksisene. Både kvantitative og kvalitative data tyder på at foreldrene hadde gode erfaringer med oppfølgingen fra koordinatorene. HOLF-modellen hadde positive effekter for koordinatorenes opplevelse av profesjonell kompetanse.
Systematisk oppfølging av frafall i videregående opplæring. Evaluering av IKO-modellen (2016–2019)[d]	IKO-modellen, en modell for identifisering, kartlegging og oppfølging av elever i fare for frafall fra videregående opplæring. Målet var at modellen skulle føre til at flere elever fullførte videregående skole.	Klynge-randomisert design med 42 videregående skoler, hvorav 20 implementerte IKO-modellen, mens 22 skoler arbeidet som før med frafall.	Resultatene viste få tydelige effekter ett og to år etter at modellen ble innført. Noen funn pekte på at IKO-modellen reduserte andelen elever med høyt fravær og styrket elevenes opplevelse av støtte fra lærerne samt reduserte skoletilpasningsproblemer. Vi viste at det var bedre utvikling for elevene når implementeringskvaliteten er høy, spesielt gjaldt dette to år etter implementering av IKO-modellen. Kvalitative data pekte på at enkelte lærere opplevde modellen som arbeidskrevende.
Et lag rundt eleven. En klynge-randomisert evaluering av LOG-modellen- (2017–2020)[e]	LOG-modellen er en modell for ledelse, organisering og gjennomføring med hensikt å forbedre det tverrprofesjonelle arbeidet i skolen. Å ansvarliggjøre ledere i skolen og kommunen er sentralt, samt bedre bruk av eksisterende arenaer, slik som ressursteam og styringsgrupper.	Klynge-randomisert design, der 35 barneskoler fra fire kommuner ble randomisert. 19 barneskoler implementerte LOG-modellen, 16 arbeidet som før med tverrprofesjonelt samarbeid.	Resultatene viste positive effekter for samarbeids-partnere, men ikke for lærere eller elever. Kvalitative data pekte på at modellens virkningsmekanismer kun ble delvis aktivert, og at lærere ikke ble tilstrekkelig involvert i implementeringen.

a. Malmberg-Heimonen, 2011; Malmberg-Heimonen & Johansen, 2014.
b. Malmberg-Heimonen, 2015; Malmberg-Heimonen et al., 2016.
c. Gyüre et al., 2022; Malmberg-Heimonen & Tøge, 2022; Tøge et al., 2020.
d. Sletten et al., in press; Malmberg-Heimonen et al., 2019.
e. Hynek et al., 2020; Malmberg-Heimonen et al., 2020.

Slik vi diskuterte i innledningen, bør kontrasten mellom eksperiment- og kontrollgruppen være eksplisitt uttrykt og målbar. I flere av eksperimentene har vi sett en svak kontrast mellom eksperiment- og kontrollgruppen. Et eksempel er HOLF-prosjektet. Der ble 29 NAV-kontorer randomisert til å implementere henholdsvis lokale familieprosjekter og HOLF-modellen utviklet av Arbeids- og velferdsdirektoratet. I begge gruppene av kontorer hadde familieprosjektene mange fellesnevnere: Familiekoordinatorene skulle arbeide med alle familiemedlemmer, koordinere tjenester og følge opp flere målområder. Vår evaluering viste at koordinatorer i eksperiment- og kontrollkontorer jobbet tilnærmet likt; det var ikke systematiske forskjeller i hvordan de fulgte opp familier. Familiene i de to gruppene fikk nokså lik utvikling over tid; det var ingen signifikante forskjeller mellom eksperimentgruppe og kontrollgruppe. Disse funnene bidrar likevel til innsikt. De tyder på at det finnes betydelig lokal kompetanse hos NAV-kontorene til å følge opp sårbare grupper, og at selve HOLF-modellen med sine verktøy og skjemaer ikke hadde forventet merverdi for familiene. Kvalitative data fra prosjektet tyder samtidig på at familiekoordinatoren kunne oppdage forhold som ikke var kjent for saksbehandlere i NAV, familieperspektivet ble bedre ivaretatt på NAV-kontoret, og barnas behov ble bedre sett (Malmberg-Heimonen et al., 2021). Vi kartla også relevant arbeidserfaring blant de profesjonelle. Dette forsterket grunnlaget for å sammenligne familieprosjektene med ordinær praksis. Før de gikk inn i rollen som familiekoordinator, var relasjonelle tilnærminger, brukermedvirkning og målfokuserte møter fremtredende hos de profesjonelle, mens de i mindre grad fulgte opp familiene helhetlig og koordinerte deres tjenester (Tøge et al., 2020).

Målgruppene ved NAV-kontorene var imidlertid mye større enn familiekoordinatorene hadde kapasitet til å dekke, så de trakk ut tilfeldige familier som fikk tilbud om oppfølging. Det gav oss også mulighet til å undersøke om koordinatorene, uavhengig av HOLF-modell, var mer effektivt enn ordinær oppfølging ved NAV-kontoret (Gyüre et al., 2022). Analyser av disse dataene viser en tendens til at flere foreldre som fikk oppfølging av en koordinator, gikk over i arbeid og arbeidsmarkedstiltak, men effektene var små og ikke signifikante (Gyüre et al., 2022).

Vi har evaluert intervensjoner som opererer på flere nivåer. I noen tilfeller har mekanismer som opererer på systemnivå, bidratt til svakere kontrast mellom eksperiment- og kontrollgruppen. Et eksempel er evalueringen av LOG-modellen med hensikt å forbedre det tverrprofesjonelle arbeidet i skolen. I modellen var et sentralt element å mobilisere aktuelle tverrprofesjonelle ressurser i skolen og kommunen. Det kan ha bidratt til at også skoler randomisert til kontrollgruppen kan ha fått tilgang til enkelte elementer av intervensjonen. Dette er et eksempel på smitte på systemnivå og dermed en uønsket svekket behandlingskontrast. Samtidig viste studien at det å involvere samarbeidspartnere på kommunenivå var helt sentralt for å nå målsettingene (Malmberg-Heimonen et al., 2020).

I noen av våre eksperimenter har også praksiskonteksten for eksperimentet endret seg underveis, noe som også kan ha bidratt til svakere kontrast. Et eksempel er IKO-prosjektet, der intervensjonen hadde til hensikt å systematisere det frafallsforebyggende arbeidet i eksperimentskoler. En ulempe var at forvaltningen implementerte nasjonalt fraværsreglement samtidig som vi gjennomførte eksperimentet. Dette bidro til at også kontrollskolene la betydelig vekt på systematisert oppfølging av fraværsutsatte elever. Kontrollskolene innførte lignende innsats som eksperimentskolene; kontrasten mellom eksperiment- og kontrollgruppen ble svakere enn forventet, noe som kan ha bidratt til relativt beskjedne effekter av modellen (Sletten et al., 2022).

Kontrasten var tydelig i nettverksrådseksperimentet. Dette eksperimentet var individrandomisert, og selve nettverksrådsprosessen ble gjennomført av en koordinator som *ikke*

var ansatt av sosialtjenesten (Malmberg-Heimonen, 2011). Deltakelsen i nettverksrådet ble dermed et tydelig «tillegg» for eksperimentgruppen som kontrollgruppen ikke fikk tilgang til. Nettverksrådsintervensjonen i seg selv var en komplisert og manualbasert modell som hverken sosialarbeidere eller andre ansatte i sosialtjenesten kunne gjennomføre uten koordinatoren. Vi var dermed rimelig sikre på at deltakere randomisert til kontrollgruppen ikke fikk intervensjonen eller deler av den. Samtidig hadde også dette prosjektet noen andre metodologiske utfordringer. For eksempel ble enkelte sosialarbeidere «skuffet» når brukeren ble randomisert til kontrollgruppe. Sosialarbeiderne kunne også hevde at enkelte personer i målgruppen ikke var egnet til å delta i eksperimentet fordi de ikke hadde noe nettverk å mobilisere. Bekymringen blant sosialarbeiderne var at nettverksrådet snarere ville bekrefte deltakerens manglende sosiale relasjoner enn å styrke dem. Resultatene viste imidlertid at koordinatorene sammen med deltakerne klarte å mobilisere nettverk, også hos dem som i utgangspunktet hadde svake nettverk (Natland, 2011). Intervjuer med deltakere pekte riktig nok på at de svake nettverkene var en av årsakene til at de positive mentale helseeffektene ikke holdt seg over tid (Malmberg-Heimonen & Johansen, 2014).

Samlet sett viser gjennomgangen av studiene at identifisering av behandlingskontrast er avgjørende for tolkningen av resultatene. Når studiene ikke avdekker effekter, kan det være flere årsaker, for eksempel at intervensjonen ikke er tilstrekkelig implementert, at intervensjonen har likhetstrekk med ordinær praksis, eller at intervensjonen delvis smitter over i kontrollgruppen.

Diskusjon og konklusjon

Vi har i denne artikkelen diskutert eksperimentell forskning innen velferd og utdanning samt eksemplifisert noen utfordringer vi har støtt på i forskningen vår. Vi har spesielt pekt på betydningen av behandlingskontrast og vist at i enkelte av studiene mottok eksperiment- og kontrollgruppen lignende behandling og service. I det følgende diskuterer vi strategier for å studere behandlingskontrasten og benytte denne i fortolkningen av resultater.

Ofte ønsker vi ikke bare å besvare spørsmålet om intervensjonens effekter, men også få kunnskap om hvordan og hvorfor intervensjonen virker eller ikke virker. For å måle behandlingskontrast har vi undersøkt aktiviteter i eksperiment- og kontrollgruppen gjennom tilsvarende data. For å få en dypere innsikt har vi i samtlige eksperimenter benyttet både kvalitative og kvantitative data, slik også andre forskere anbefaler (Grissmer et al., 2009). I begge gruppene har vi intervjuet deltakere og profesjonelle parallelt med å samle inn data gjennom spørreundersøkelser. For å få et mer presist bilde av effekter har vi også brukt registerdata. Med tanke på at deltakere kan bli påvirket av randomisering til den ene eller andre gruppen, kan registerdata gi anledning til å måle effekter på utfall som er mer objektive. Ofte samsvarer de ulike metodologiske tilnærmingene godt med hverandre, men i noen tilfeller kan de tegne til dels ulike bilder. For eksempel kan effektdataene vise nøytrale effekter av en intervensjon, mens både deltakere og profesjonelle i intervjuer forteller om positive erfaringer og god måloppnåelse. Videre kan observasjoner tyde på svak deltakelse i implementeringen blant bestemte aktører med påfølgende svak behandlingskontrast, mens intervjuer eller spørreundersøkelser gir inntrykk av aktiv deltakelse. Motstridende funn kan være vanskelige å tolke og formidle. Både aktører fra politikk- og praksisfeltet og forskerne selv kan bli usikre på hvilken virkelighet som er den «riktige», og hvilke data og analyser man bør legge mest vekt på.

Vanligvis utfører man eksperimenter fordi man har en eller annen hypotese om årsaksforhold, det vil si hva en spesifikk intervensjon skal bidra med, og hvordan. For å vite hva

man skal måle, både av intervensjonselementer, aktiviteter kontrollgruppen deltar i, og forventede effekter, er det hensiktsmessig å utvikle en programteori. En programteori er en antakelse om hvordan intervensjonen er tenkt å virke, herunder elementer, aktiviteter og forventede utfall (Funnell & Rogers, 2011). Jo mer eksplisitt denne teorien (eller teoriene dersom man har flere) er, jo bedre mulighet har man til å innhente data, planlegge metoder og finne frem til adekvate måleinstrumenter for å undersøke antatte mekanismer og effekter. Å identifisere behandlingskontrasten bør være en sentral del av programteorien. Uansett om utviklingen av intervensjonen primært er forskerstyrt eller styrt fra behov innen politikk og praksisfelt, bør forskere, oppdragsgivere og aktører fra praksis gå sammen om å utvikle en programteori, slik at alle aktørene har en felles forståelse av hvordan intervensjonen er tenkt å virke, hva som er behandlingskontrasten, og hvordan effekter skal analyseres (Malmberg-Heimonen et al., 2018; Tøge et al., 2020). For å sikre at analyser og rapportering av funn baserer seg på opprinnelige hypoteser, bør en detaljert beskrivelse av studien offentliggjøres i forkant av gjennomføringen. De to vanligste måtene å offentliggjøre planlagte studier på er registreringer i databaser (for eksempel ClinicalTrials.gov) eller gjennom en protokoll. Ved rapportering skal forskere bruke fastsatte standarder. Den mest utbredte i velferds- og utdanningsfeltet er CONSORT (Consolidated Standards of Reporting Trials) (Campbell et al., 2004; Eldridge et al., 2016; Montgomery et al., 2013; Schulz et al., 2010). I disse protokollene og standardene må forskerne beskrive behandlingskontrasten i detalj. Standardene inneholder også lister over hva forskerne må rapportere ved publisering. Disse listene er nyttige for forskere også i planleggingsfasen fordi de fungerer som en huskeliste over de ulike elementene man må vurdere og avgjøre før selve gjennomføringen av eksperimentet.

Å utvikle intervensjoner og evaluere dem i komplekse kontekster handler ikke kun om å gjennomføre *én* studie av *én* intervensjon. Det handler også om en langvarig og syklisk forskningsinnsats med forarbeid, implementering og testing i kontrollerte former samt videreutvikling, implementering og evaluering i praksis. Etter evalueringen vil det være aktuelt å justere programteorien for så å implementere og evaluere igjen. Pilotering av forskningsdesign, intervensjon og behandlingskontrast anbefales på samtlige felt (Eldridge et al., 2016), men det er sjelden at intervensjoner innen samfunnsvitenskapelige områder blir evaluert gjennom tilsvarende faser som innen medisin, der intervensjoner ofte evalueres i opptil fem faser (Vicki, 2014). Det kan være kostnadskrevende å evaluere i flere faser, men nøyaktig pilotering av design og intervensjon vil være med på å optimalisere kunnskapsutbyttet fra eksperimentet. Gjennom pilotering av selve randomiseringen og intervensjonen kan forskere oppdage utfordringer ved måten randomiseringen gjennomføres på, eller at behandlingskontrasten ikke er i tråd med det man ønsker å måle effekter av. Det er da mulig å justere prosedyrene før igangsettelse av et fullskala-eksperiment. I denne utviklingsprosessen er også begrepet etterprøvbarhet sentralt, som betyr at studien skal være mulig å gjennomføre på samme vis senere. Hensikten er å teste ut om en intervensjon viser de samme resultatene i en annen kontekst, for eksempel et annet land, i en annen praksiskontekst eller for en annen målgruppe.

De senere årene har forvaltningen i økende grad utlyst forskningsmidler med ønske om eksperimentelt design for å gi innsikt i effekter av tiltak og politikkutforming. Disse prosjektene har bidratt til innsikt i umiddelbar nytte og på den måten gitt politikere og byråkrater grunnlag for å mene og beslutte. I flere av våre studier var behandlingskontrasten relativt svak, og vi identifiserte lignende elementer i eksperiment- og kontrollgruppen, noe som bidro til å forklare relativt marginale effekter av intervensjonene. Dersom man har gode empiriske målinger av implementering og behandlingskontrast, slik vi har hatt som

målsetting å ha i våre studier, kan slike data være med på å belyse årsakene til nøytrale eller marginale effekter. Studiens kvalitet, ikke resultatene, må ligge til grunn for hva som vektlegges i kunnskapsproduksjonen, den politiske diskusjonen og den videre politikkutformingen. En ting som bidrar til god kvalitet, er tydelige beskrivelser av hva som tilbys til eksperiment- og kontrollgruppen. Brukere av forskning bør være oppmerksom på denne behandlingskontrasten, for det er nøyaktig den som studiene måler effekter av.

Referanser

Angrist, J.D. & Pischke, J.S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *The Journal of Economic Perspectives, 24*(2), 3–30. https://doi.org/10.1257/jep.24.2.3

Barton, S. (2000). Which clinical studies provide the best evidence?: The best RCT still trumps the best observational study. *BMJ,* (321), 255–256. https://doi.org/10.1136/bmj.321.7256.255

Breit, E., Fossestøl, K. & Pedersen, E. (2019). Kunnskapsbasert praksis innenfor en samstyringsmodell. *Tidsskrift for velferdsforskning, 22*(03), 184–197. https://doi.org/10.18261/issn.2464-3076-2019-03-01

Campbell, M.K., Elbourne, D.R. & Altman, D.G. (2004). CONSORT statement: extension to cluster randomised trials. *BMJ, 328*(7441), 702–708. https://doi.org/10.1136/bmj.328.7441.702

Campbell, M.K., Piaggio, G., Elbourne, D.R. & Altman, D.G. (2012). Consort 2010 statement: extension to cluster randomised trials, *BMJ, 345*, e5661. https://doi.org/10.1136/bmj.e5661

Connolly, P., Keenan, C. & Urbanska, K. (2018). The trials of evidence-based practice in education: a systematic review of randomised controlled trials in education research 1980–2016. *Educational Research, 60*(3), 276–291. https://doi.org/10.1080/00131881.2018.1493353

Deaton, A. & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine, 210*, 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005

Eldridge, S.M., Chan, C.L., Campbell, M.J., Bond, C.M., Hopewell, S., Thabane, L. & Lancaster, G.A. (2016). CONSORT 2010 statement: extension to randomised pilot and feasibility trials. *BMJ, 355*, i5239. https://doi.org/10.1136/bmj.i5239

Flannelly, K.J., Flannelly, L.T. & Jankowski, K.R.B. (2018). Threats to the Internal Validity of Experimental and Quasi-Experimental Research in Healthcare. *Journal of Health Care Chaplaincy, 24*(3), 107–130. https://doi.org/10.1080/08854726.2017.1421019

Fretheim, A. (2013). Kunnskapsbasert politikkutforming. *Norsk epidemiologi, 23*(2), 205–210.

Funnell, S.C. & Rogers, P.J. (2011). *Purposeful program theory: Effective use of theories of change and logic models*. San Francisco: Jossey-Bass/Wiley.

Goodman, L.A., Epstein, D. & Sullivan, C.M. (2018). Beyond the RCT: Integrating Rigor and Relevance to Evaluate the Outcomes of Domestic Violence Programs. *The American Journal of Evaluation, 39*(1), 58–70. https://doi.org/10.1177/1098214017721008

Grissmer, D.W., Subotnik, R.F. & Orland, M. (2009). *A guide to incorporating multiple methods in randomized controlled trials to assess intervention effects*. American Psychological Association. Hentet fra https://www.apa.org/ed/schools/teaching-learning/randomized-control-guide.pdf

Gupta, S.K. (2011). Intention-to-treat concept: a review. *Perspectives in clinical research, 2*(3), 109–112. https://doi.org/10.4103/2229-3485.83221

Gyüre, K.T., Tøge, A.G. & Malmberg-Heimonen, I. (2022). The Effects of Service Coordination on Disadvantaged Parents' Participation in Activation Programs and Employment: A Randomized Controlled Trial. *Research on Social Work Practice*, *32*(4),402–414. https://doi.org/10.1177/10497315211046523

Hamilton, G. & Scrivener, S. (2018). Measuring Treatment Contrast in Randomized Controlled Trials. MDRC Working paper. Hentet fra https://www.mdrc.org/sites/default/files/MTC_Paper_MDRC_WEBSITE_VERSION.pdf

Hjelmar, U. & Pedersen, P.V. (2015). Kompleks evaluering – tre metodiske læringspunkter fra en case. *Metode & Forskningsdesign, 2*(2), 32–56.

Hutchison, D. & Styles, B. (2010). *A guide to running randomised controlled trials for educational researchers* NFER Slough.

Hynek, K.A., Malmberg-Heimonen, I. & Tøge, A.G. (2020). Improving interprofessional collaboration in Norwegian primary schools: A cluster-randomized study evaluating effects of the LOG model on teachers´perceptions of interprofessional collaboration, *Journal of Interprofessional Care*, 1–10. https://doi.org/10.1080/13561820.2019.1708281

Imbens, G. (2018). Understanding and misunderstanding randomized controlled trials: A commentary on Deaton and Cartwright. *Social Science & Medicine, 210*, 50–52. https://doi.org/10.1016/j.socscimed.2018.04.028

Leko, M.M. (2015). To Adapt or Not to Adapt: Navigating an Implementation Conundrum. *Teaching Exceptional Children, 48*(2), 80–85. https://doi.org/10.1177/0040059915605641

Ling, T. (2012). Evaluating complex and unfolding interventions in real time, *Evaluation, 18*(1), 79–91. https://doi.org/10.1177/1356389011429629

Lortie-Forgues, H. & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher, 48*(3), 158–166. https://doi.org/10.3102/0013189X19832850

Malmberg-Heimonen, I. (2011). The effects of family group conferences on social support and mental health for longer-term social assistance recipients in Norway. *British Journal of Social Work, 41*(5), 949–967. https://doi.org/10.1093/bjsw/bcr001

Malmberg-Heimonen, I. & Johansen, S. (2014). Understanding the longer-term effects of family group conferences. *European Journal of Social Work, 17*(4), 556–571. https://doi.org/10.1080/13691457.2013.818528

Malmberg-Heimonen, I. (2015). Social workers´ training evaluated by a cluster-randomized study: reemployment for welfare recipients? *Research on Social Work Practice, 25*(6), 643–653. https://doi:10.1177/1049731515569357

Malmberg-Heimonen, I., Natland, S., Tøge, A.G. & Hansen, H.C. (2016). The effects of skill training on social workers´professional competences in Norway: Results of a cluster-randomized study. *British Journal of Social Work, 46*(5), 1354–1371. https://doi.org/10.1093/bjsw/bcv073

Malmberg-Heimonen, I., Tøge, A.G. & Fossestøl, K. (2018). Program Theory within Policy-Initiated Evaluations: The Norwegian Low-Income Family Study. *Journal of Evidence-Informed Social Work, 15*(4), 337–350. https://doi.org/10.1080/23761407.2018.1455161

Malmberg-Heimonen, I., Sletten, M. Tøge, A.G., Alves, D., Borg, E. & Gyüre, K. (2019). *Å forebygge frafall i videregående opplæring.* OsloMet skriftserien 2019:1. Hentet fra https://skriftserien.oslomet.no/index.php/skriftserien/article/view/623/139

Malmberg-Heimonen, I., Tøge, A.G., Lyng, S.T., Borg, E., Pålshaugen, Ø., Bakkeli, V., … Fossestøl, K. (2020). *Et lag rundt eleven. En klynge-randomisert evaluering av LOG-modellen*. AFI rapport 2020:7. Hentet fra https://oda.oslomet.no/oda-xmlui/bitstream/handle/20.500.12199/6455/Et%20lag%20rundt%20eleven%20En%20klyngeevaluering%20av%20LOG-modellen.pdf?sequence=4&isAllowed=y

Malmberg-Heimonen, I., Tøge, A.G., Rugkåsa, M. & Bergheim, B. (2021). The child perspective within family intervention projects: a cluster-randomised study with a mixed methods design. *European Journal of Social Work*, 1–13. https://doi.org/10.1080/13691457.2021.1977247

Malmberg-Heimonen, I. & Tøge, A.G. (2022). Family intervention projects as poverty-alleviating measures: results from a Norwegian cluster randomised study. *Social Policy & Society*, 1-16. https://doi.org/10.1017/S1474746422000124

Meld. St. 4. (2018–2019). Langtidsplan for forskning og høyere utdanning 2019–2028. Kunnskapsdepartementet.

Montgomery, P., Grant, S., Hopewell, S., Macdonald, G., Moher, D., Michie, S. & Mayo-Wilson, E. (2013). Protocol for CONSORT-SPI: an extension for social and psychological interventions. *Implementation Science, 8*(1), 99. https://doi.org/10.1136/bmj.c869

Murray, D., Varnell, S. & Blitstein, J. (2004). Design and Analysis of Group-Randomized Trials: A Review of Recent Methodological Developments. *American Journal of Public Health, 94*(3), 423–432. Hentet fra https://pubmed.ncbi.nlm.nih.gov/14998806/

Natland, S. (2011). Spørsmålet om den «gode timing»: nettverksråd i sosialarbeiderens hverdag. *Fontene Forskning, 2011*(2), 40–52.

Pontoppidan, M., Keilow, M., Dietrichson, J., Solheim, O.J., Opheim, V., Gustafson, S. & Andersen, S.C. (2018). Randomised controlled trials in Scandinavian educational research. *Educational Research, 60*(3), 311–335. https://doi.org/10.1080/00131881.2018.1493351

Raudenbush, S.W. (2018). On randomized experimentation in education: A commentary on Deaton and Cartwright, in honor of Frederick Mosteller. *Social Science & Medicine, 210*, 63–66. https://doi.org/10.1016/j.socscimed.2018.04.030

Schulz, K.F., Altman, D.G. & Moher, D. (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ, 340*, c332. https://doi.org/10.1136/bmj.c332 .

Sletten, M.Å., Tøge, A.G. & Malmberg-Heimonen, I. (2022). Effects of an early warning system on student absence and completion in Norwegian upper secondary schools: A Cluster-Randomised Study. *Scandinavian Journal of Educational Research*, 1–15. https://doi.org/10.1080/00313831.2022.2116481

Tøge, A.G., Malmberg-Heimonen, I., Liodden, T., Rugkåsa, M., Gyüre, K. & Bergheim, B. (2020). Improving follow-up with low-income families in Norway. What is new and what is already regular social work practice? *European Journal of Social Work*, *23*(5), 729–741. https://doi.org/10.1080/13691457.2019.1602513

Vicki, L.M. (2014). Clinical Trial Phases. *International Journal of Clinical Medicine, 5*(21), 1374–1383. https://doi.org/10.4236/ijcm.2014.521175