RAPPORT 44/2009

Muligheter og utfordringer ved bruk av kontrafaktisk analyse i forskningsbaserte evalueringer

Inge Ramberg

NIFU STEP

Rapport 44/2009
ISBN 978-82-7218-652-3
ISSN 1504-1824

For en presentasjon av NIFU STEPs øvrige publikasjoner, se www.nifustep.no

Norsk institutt for studier av innovasjon, forskning og utdanning
Norwegian Institute for Studies in Innovation, Research and Education
Wergelandsveien 7, 0167 Oslo
Tlf. +47 22 59 51 00 • www.nifustep.no

RAPPORT 44/2009

Inge Ramberg

Muligheter og utfordringer ved bruk av kontrafaktisk analyse i forskningsbaserte evalueringer

NIFU STEP

Forord

Denne rapporten er utarbeidet på oppdrag av EVA-forum (Nettverk for evaluering i staten) som ved årsskiftet 2008/2009 initierte en studie av bruken av kontrafaktisk analyse i evalueringer i Norge.

Kontrafaktisk analyse er en relativt ukjent term i norske evalueringer, men samtidig sentral når man ønsker å måle effekter eller langsiktige virkninger av et tiltak.

I rapporten kartlegges og vurderes bruken av kontrafaktisk analyse i evalueringer innenfor bistands-, innovasjons- og næringsutviklingsfeltet i Norge sett i forhold til erfaringer internasjonalt. Formålet er som det heter i utlysningen av prosjektet, ” å lære av erfaringene med bruk av slike analyser for å kunne bidra til metodeutvikling og dermed forbedre kvaliteten på kommende evalueringer”. Rapporten er skrevet med tanke på personer som planlegger, bestiller og gjennomfører evalueringer og behandler derfor metodediskusjonen på feltet i fugleperspektiv snarere enn på detaljnivået der metodeutviklingen foregår.

Vi takker oppdragsgiverne for et krevende og interessant oppdrag samt gode innspill underveis i prosjektet.

Oslo, november 2009

Bjørn Stensaker
Konstituert direktør

Magnus Gulbrandsen
Forskningsleder

Innhold

Sammendrag

Rapporten omhandler muligheter og utfordringer ved bruk av kontrafaktisk analyse i forskningsbaserte evalueringer og er skrevet med tanke på personer som planlegger, bestiller og gjennomfører evalueringer.

Kontrafaktisk beskriver tilstanden som ville inntruffet dersom et virkemiddel ikke hadde vært tatt i bruk. Den kontrafaktiske tilstanden benyttes for å vurdere virkningen av et tiltak ved å sammenligne utfallet for grupper som har deltatt i tiltaket med tilsvarende utfall for grupper som ikke har deltatt. Kontrafaktisk analyse kan benytte eksperimentelle eller kvasi-eksperimentelle metodedesign.

Rapporten tar hovedsakelig for seg kvantitativt orienterte analyser da kvalitativt orienterte evalueringsstudier med en eksplisitt kontrafaktisk design sjelden benyttes. Den kontrafaktiske tilstanden kan imidlertid også benyttes i casestudier i strukturerte sammenligninger av ulike case for å studere sammenfallet av ulike årsaksfaktorer (betingelser) når antallet enheter er lavt.

I den innledende delen av denne studien kartla vi *foreliggende norske evalueringsstudier innenfor bistandsområdet samt næringsutvikling og innovasjon* som benytter kontrafaktisk analyse. I kapittel 2 av rapporten beskrives utvalgte evalueringsstudier på disse feltene med vekt på kvasi-eksperimentelle design. Kontrafaktisk analyse benyttes sjelden og da kun et fåtall av de tilgjengelige designene. Eksperimentelle design med randomisering av kontroll- og eksperimentgrupper er brukt helt unntaksvis innenfor dette feltet.

Drøftingen av erfaringer og muligheter ved evalueringsstudiene peker mot et *forbedringspotensial* for norske evalueringer av effekter og langsiktige virkninger av gjennomførte tiltak. Hele evalueringsprosessen fra planleggingsstadiet til oppdragstakerens analyse og rapportering er viktig for å kunne benytte robuste evalueringsdesign som gir valide konklusjoner om tiltakets effekter. Eksempelvis krever et før-etter design at det etableres en *baseline* (minst ett målepunkt som viser utgangsnivået) før tiltak iverksettes. I tillegg til *god planlegging og ekstra ressursinnsats* kreves også *økt metodekompetanse* i evalueringsmiljøene.

Vurdering av *tiltakets evaluerbarhet* er et nyttig planleggingsverktøy både for oppdragsgiver så vel som for evalueringsmiljøet. I enkeltevalueringer som er initiert uten *tilstrekkelige forberedelser*, vil det sjelden være mulig å gjennomføre en kontrafaktisk analyse av effekter og virkninger fordi tiltaket fortsatt pågår eller nylig er avsluttet mens langsiktige effekter og virkninger kanskje ikke er målbare før 3-5 år har gått. *Tidspunktet for når et tiltak kan være modent for en effektevaluering* med kontrafaktisk tilsnitt er med andre ord ikke helt tilfeldig.

I flere tilfeller der policyaktører etterspør effektene av tiltak, vil det heller ikke være mulig å gjennomføre kontrafaktiske analyser fordi tiltaket er universelt utformet for populasjonen. Da kan man vanskelig etablere en kontrafaktisk tilstand. Før oppdragsgiver

utarbeider mandatet, bør dette avklares gjennom en ”evaluerbarhetsvurdering”. Er det pågående prosesser eller tiltak som skal studeres, bør mandatet unngå å etterlyse effekter. Da kan det kanskje i stedet være behov for en kvalitativt orientert prosessevaluering.

Generelt er kontrafaktisk analyse best egnet for å vurdere *effekter av avsluttede langsiktige satsinger* basert på tidsseriedata og robuste evalueringsdesign med flere målepunkter også etter avsluttet tiltak/program.

Mandatet for en evaluering må uansett være realistisk i forhold til *oppdragets omfang og tilgjengelige ressurser*. Å vektlegge laveste pris i utlysninger av effektevalueringer er risikabelt for kvaliteten på arbeidet siden det er få relevante tilbydere av slike tjenester nasjonalt samtidig som det er klart behov for kompetanseutvikling på feltet.

Mandatet for effektevalueringer kan med fordel etterlyse at det skal gjennomføres en *summativ evaluering basert på kontrafaktisk analyse* med en veltilpasset metodedesign som gir grunnlag for å trekke slutninger om tiltakets effekter.

1 Innledning

1.1 Om oppdraget og rapporten

Oppdraget er definert som ”en studie av kontrafaktisk analyse i norske evalueringer. Analysen skal omfatte evalueringer hvor Den Norske Stat er bestiller, som innebærer at bistand og utenrikspolitikken er inkludert”. I oppstartsfasen av prosjektet ble det avklart at studien skulle fokusere på analyser innenfor bistandsområdet samt næringsutvikling og innovasjon. Oppdragets omfang gjorde det nødvendig å avgrense det empiriske materialet til de nevnte områdene. Vi bygger imidlertid delvis også på kartlegginger fra andre forvaltnings-/ politikkområder uten at dette materialet adresseres eksplisitt[1].

Konkurransegrunnlaget for oppdraget omtaler videre ”personer som planlegger, bestiller og gjennomfører evalueringer – både i og utenfor staten” som primære brukere av studien. Det er vårt håp at rapporten kan bidra til å fremme formålet med denne studien som ”er å lære av erfaringene med bruk av slike analyser for å kunne bidra til metodeutvikling og dermed forbedre kvaliteten på kommende evalueringer”.

Hovedtema

Konkurransegrunnlaget definerer følgende fire hovedtema for oppdraget:

I. Kartlegging av kontrafaktiske analyser gjennomført i Norge de seneste 10-15 årene
II. Vurdering av de ulike metodene opp mot internasjonal erfaring
III. Sammenfatning av de viktigste lærdommene fra disse evalueringene
IV. Praktiske anbefalinger som kan bidra til videre metodeutvikling

Disse spørsmålene blir behandlet i kapitlene 2 – 4 i denne rapporten. Først følger en beskrivelse av tidligere kartlegginger og avklaring av begrepet ”kontrafaktisk analyse” samt en beskrivelse av relevante metodedesign ved effektevalueringsstudier. De identifiserte evalueringene på feltet framgår av referanselisten til slutt i rapporten.

Vårt referansepunkt for kartleggingen og den videre analysen er teoribygging innen fagfeltet policyanalyse basert på en statsvitenskaplig/ samfunnsøkonomisk tilnærming. I denne sammenheng er Mohr (1992) *Impact Analysis for Program Evaluation*[2] sentral. Mohrs versjon av kvantitativt orienterte kontrafaktiske analyser tar utgangspunkt i

---

[1] Arbeidsmarkedsfeltet (ledighetstiltak), rusmiddelfeltet og trafikksikkerhetsområdet er blant forvaltningsområdene hvor det foreligger kartlegginger av effektstudier (med kontrafaktiske tilnærminger) i norsk sammenheng. Dette står i kontrast til utdanningsfeltet hvor effektstudier er sjeldne i Norge, mens det i amerikansk sammenheng er lang tradisjon for å benytte (kvasi-)eksperimentelle effektstudier.

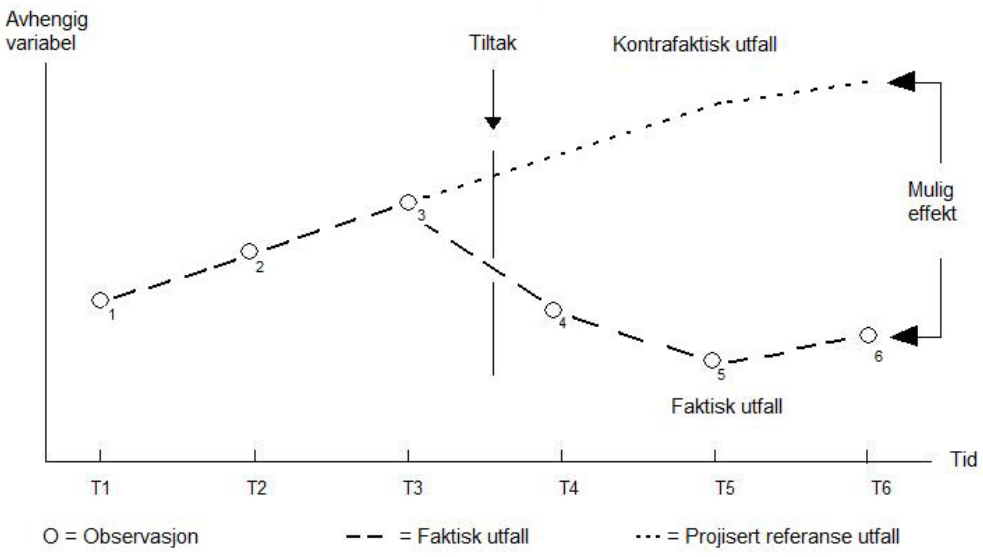
[2] Mohr, L.B. (1992): *Impact analysis for program evaluation*. Sage Publications Inc, CA, USA.

Campbell & Stanleys (1966)[3] ulike pre- og (kvasi)eksperimentelle forskningsdesign, fra det enkleste 'One-shot case study' til det kontrollerte eksperimentelle oppsettet 'Pretest-posttest design med kontrollgruppe'. Mohr omtaler en rekke ulike forskningsdesign og beskriver deres relative styrker og svakheter – i forhold til trusler mot intern og ekstern validitet. Bamberger (2006) er en annen sentral referanse for denne litteraturstudien. Vi benytter Bambergers klassifisering av kvasi-eksperimentelle evalueringsdesign i boka *RealWorld Evaluation* i den videre kartleggingen og drøftingen av evalueringspraksisen, i tillegg til flere andre evalueringsmetodiske bidrag fra det siste tiåret.

1.2 Kontrafaktisk tilstand og kontrafaktisk analyse

"Kontrafaktisk" og "kontrafaktisk analyse" er sjeldne termer i norske evalueringer, men samtidig viktige verktøy for å kunne vurdere langsiktige effekter/ virkningen av et tiltak. *Kontrafaktisk* beskriver tilstanden som ville inntruffet dersom et virkemiddel ikke hadde vært tatt i bruk. Med utgangspunkt i den kontrafaktiske tilstanden, kan en vurdere virkningen av et tiltak ved å sammenligne resultatet for grupper som har deltatt i tiltaket med tilsvarende resultat for grupper som ikke har deltatt. Mohr (1999) definerer den kontrafaktiske tilstanden på følgende måte": *X was the cause of Y if and only if both X and Y occurred and, in the circumstances, if X had not occurred, then neither would Y (the counterfactual)*" Figur 1, hentet fra Vedung (1997), illustrerer dette alternative utfallet, som vil inntreffe dersom et virkemiddel ikke tas i bruk. Her indikeres samtidig forskjellen mellom faktisk og kontrafaktisk utfall som gir mulighet til å beregne effekter av et tiltak.

[3] Campbell, D.T. & Stanley, J.C. (1966): Experimental and Quasi-experimental Designs for Research. Chicago.



Figur 1 Kontrafaktisk utfall i effektevaluering. Kilde: Vedung (1997).

I en evaluering der man ønsker å måle virkningen av et tiltak er den ideelle sitasjonen at en har flere målepunkter eller observasjoner før og etter at tiltaket innføres. I figuren innføres tiltaket mellom tidspunktene T3 og T4 og den langsiktige effekten kan senere beregnes med utgangspunkt i tidspunktet T6, lenge etter at tiltaket er avsluttet.

Kontrafaktisk analyse har sitt opphav i filosofien. Filosofer har siden David Hume (1748) vært opptatt av kontrafaktisk tenkning, mens spesielt David Lewis' (1973) kontrafaktiske teori om kausalsammenhenger[4] har inspirert den senere utviklingen av kontrafaktisk analyse[5]. I evalueringssammenheng kan man gjennomføre en kontrafaktisk analyse ved å benytte en kontrollgruppe (eller sammenligningsgruppe ved ikke-randomiserte opplegg) for å vurdere virkningen av et tiltak. Utfallet for eksperimentgruppen (som har vært eksponert for tiltaket) sammenlignes her med utfallet for sammenligningsgruppen (den kontrafaktiske tilstanden).

Kontrafaktiske analyser er gjerne kvantitativt orienterte, men dette er ingen nødvendighet. Den kontrafaktiske tilstanden kan også utnyttes i casestudier ved hjelp av strukturerte sammenligninger av ulike case (basert på programteorien for det definerte problemet) for å studere sammenfallet av ulike årsaksfaktorer (betingelser) når antallet enheter er lavt[6]. Med for eksempel fem årsaksfaktorer til et utfall vil man teoretisk sett kunne få 2^5=32 mulige

---

4 D. Lewis (1973) 'Causation' in *Journal of Philosophy*, 70:556-67

5 P. Menzies (2001): 'Counterfactual Theories of Causation' in *Stanford Encyclopedia of Philosophy (online edition, updated 2008).*

6 Denne metoden kalles Qualitative Comparative Analysis av Ragin i *The Sage Handbook of Social Science Methodology*.

kombinasjoner av årsaksfaktorene (betingelsene) som kan gi et bestemt utfall/ være opphav til et fenomen. Flere av disse teoretiske kombinasjonene vil ikke kunne observeres empirisk. Men siden kombinasjonene er synliggjort, kan de ifølge Ragin (2007) benyttes som hypotetiske case i tankeeksperimenter ved hjelp av kontrafaktisk analyse[7].

Shadish, Cook og Campbell (2002:499) innvender imidlertid at ikke-eksperimentelle alternativer som intensive kvalitative case studier og teoribaserte evalueringer gir mindre klare kausale sammenhenger. Dels begrunner de dette med at kvalitative metoder gjerne gir upresis kunnskap om den kontrafaktiske tilstanden.

1.3 Ulike metodedesign for effektevaluering basert på kontrafaktisk tenkning

I denne studie forstår vi "kontrafaktisk analyse" som et analytisk verktøy der vurderinger av det faktiske utfallet av et tiltak i forhold til det kontrafaktiske, gir grunnlaget for kausale slutninger om tiltakets effekt. Hensikten med analysen er slik Kvitastein (2000:4) uttrykker det: " … å finne de effekter som kan tilskrives programmet, og samtidig unngå at effekter som ikke skyldes programmet, tilskrives programmet". Kvantitative effektevalueringsstudier (*impact evaluations*) benytter ulike metodedesign avhengig av problemstillingen og i hvilken grad det er mulig å oppnå kontroll over ulike faktorer av betydning for utfallet av et tiltak. Hvis man har eksperimentell kontroll over disse betingelsene, kan en benytte randomiserte eksperimentelle design så lenge forskningsetiske vurderinger og praktiske hensyn tillater det.

Bamberger beskriver i *RealWord Evaluations* det han kaller de sju mest anvendte kvantitative evalueringsdesign (Bamberger et al. 2006: 46, 209). Dette er alle såkalte *kvasi-eksperimentelle evalueringsdesign* hvor det ikke eksisterer noen randomisering av eksperimentgruppen som er gjenstand for tiltaket og kontrollgruppen som ikke er det. Bamberger bygger her blant annet på Campbell & Stanley (1963) *Experimental and Quasi-Experimental Designs for Research* som drøfter metodiske aspekter ved seksten ulike forskningsdesign innenfor gruppene *eksperimentelle design, pre- og kvasi-eksperimentelle design.* Bamberger (2006:232) omtaler dessuten eksperimentelle design hvor både gruppen(e) som er gjenstand for tiltaket og kontrollgruppen(e) som ikke eksponeres, velges ut *tilfeldig*. Han ser få anvendelser av randomiserte eksperimentelle evalueringsdesign i evalueringsstudier utenfor det medisinske, naturvitenskaplige og til dels psykologiske fagfeltet. Selv om han nevner enkelte anvendelser av randomiserte eksperimentelle studier på sitt eget spesialfelt, bistandsevaluering, problematiserer han primært praktiske, etiske og

[7] Empirisk økonometrisk basert evalueringsforskning forutsetter en slik eksperimentell logikk mens nytte-kostnadsanalyser på den andre siden ikke gjør det, men tar for gitt at det eksisterer effekter som kan verdsettes i pengeverdier når effektene er identifisert ifølge Kvitastein (2002:30).

politiske faktorer som sterkt begrensende for eksperimentelle design. I stedet velger han å fokusere på de kvasi-eksperimentelle evalueringsdesignene[8].

De to sterkeste kvasi-eksperimentelle evalueringsdesignene er ifølge Bamberger (2006):

1. **Comprehensive longitudinal design with pre-, midterm, post- and ex-post observations on the project and comparison groups**
   (Benytter komparative tidsserier for både tiltaks- og sammenligningsgruppen)

Før-måling (T_1)	Tiltak (over tid)	Midtveis-måling (T2)	Prosjektslutt måling(T3)	Etter-måling (T4)
P_1	X	$P_{2(n)}$	P_3	P_4
C_1		$C_{2(n)}$	C_3	C_4

Merknad: P og C beskriver henholdsvis målepunktene for tiltaks(prosjektgruppen) og kontrollgruppen mens X symboliserer tiltaket som tiltaksgruppen utsettes for.

2. **Pretest-posttest project and comparison groups**
   (Benytter sammenligningsgruppe med både før- og ettermåling)

Før-måling (T_1)	Tiltak (over tid)	Midtveis-måling (T2)	Prosjektslutt måling(T3)	Etter-måling (T4)
P_1	X			P_2
C_1				C_2

… og de fem mindre robuste evalueringsdesignene:

3. **Truncated longitudinal pretest-posttest project and comparison group design**
   (Benytter sammenligningsgruppe uten før-, men med midtveis- og sluttmåling)

Før-måling (T_1)	Tiltak (over tid)	Midtveis-måling (T2)	Prosjektslutt måling(T3)	Etter-måling (T4)
	X	$P_{1(n)}$	P_2	
		$C_{1(n)}$	C_2	

[8] Bamberger (2006:233) viser imidlertid til Shadish, Cook and Campbell (2002: 269-75) for en oversikt over hvilke betingelser som er gunstige for randomiserte eksperimentelle design.

4. **Pretest-posttest project group combined with posttest analysis of project and comparison groups**
   (Benytter sammenligningsgruppe, men kun med sluttmåling for denne)

Før-måling (T_1)	Tiltak (over tid)	Midtveis-måling (T2)	Prosjektslutt måling(T3)	Etter-måling (T4)
P_1	X		P_2 C_1	

5. **Posttest project and comparison groups**
   (Benytter kun sluttmåling for både tiltaks- og sammenligningsgruppen)

Før-måling (T_1)	Tiltak (over tid)	Midtveis-måling (T2)	Prosjektslutt måling(T3)	Etter-måling (T4)
	X		P_1 C_1	

De to påfølgende evalueringsdesignene benytter ingen sammenligningsgruppe eller kontrafaktisk tilstand og dermed heller ikke kontrafaktisk analyse

6. **Pretest-posttest project group – no comparison group**
   (Benytter kun tiltaksgruppe med før- og sluttmåling)

Før-måling (T_1)	Tiltak (over tid)	Midtveis-måling (T2)	Prosjektslutt måling(T3)	Etter-måling (T4)
P_1	X		P_2	

7. **Posttest project group only**
   (Benytter kun tiltaksgruppe med sluttmåling)

Før-måling (T_1)	Tiltak (over tid)	Midtveis-måling (T2)	Prosjektslutt måling(T3)	Etter-måling (T4)
	X		P_1	

Vedung (1997) omtaler i likhet med Mohr (1992) ytterligere en kvasi-eksperimentell design, *tidsserie uten kontrollgruppe* (Interrupted Time-Series Design Intervention Group), hvor gjentatte før- og ettermålinger kan brukes i de tilfeller det er umulig å finne en kontrollgruppe, eksempelvis når tiltaket er universelt utformet for populasjonen. Denne

designen ligner med andre ord på design 1, med den klare forskjellen at sammenligningsgruppen mangler.

Vi har over valgt ut enkelte metodedesign med utgangspunkt i Bamberger (2006). Omtalen av kvasi-eksperimentelle metodedesign over er på ingen måte uttømmende. I metodelitteraturen finnes flere andre metodedesign som drøftes inngående, deriblant *Regression Discontinuity Design*[9].

Valget av metodedesign for en evaluering avhenger av en rekke faktorer deriblant; tidsperspektiv og formål for evalueringen, hvor lenge evalueringen har vært planlagt, om det foreligger data fra før prosjektet ble startet opp, om det foreligger data eller er ressurser til datainnsamling ved flere målepunkter underveis i programmet, eller om det kan brukes kontrollgruppe eventuelt en relevant sammenligningsgruppe for tiltaksgruppen.

Bamberger (2006:45) vektlegger at det ikke finnes en "beste" design eller metode, men i stedet veltilpassede og relevante evalueringsdesign for den enkelte evalueringen. Hans strategi i "RealWorld Evaluation" er å velge det mest robuste design som kan benyttes innenfor den tilmålte tidsperioden, budsjettbegrensingene, det aktuelle datagrunnlaget samt de aktuelle politiske rammebetingelsene for den enkelte evalueringen. Denne strategien er i tråd med den overordnede tanken om at økt kontroll over alle andre betingelser (variabler) enn årsaksvariablen(e) som forskeren ønsker å trekke slutninger om effekten av, styrker indre (internal) validitet. Samtidig kan imidlertid truslene mot ytre (external) validitet øke[10] noe som påvirker relevansen av evalueringsdesignen for å trekke mer allmenne slutninger om fenomenet som en studerer[11]. Laboratorieeksperimentet har således ofte høy indre validitet noe som kan true eksperimentets realisme. Og tilsvarende, felteksperimenter (kvasi-eksperiment som foregår i fenomenets naturlige miljø) har gjerne høy grad av realisme samtidig som den interne validiteten kan være utsatt for en rekke trusler. Felteksperimentet kan også beskrives som en form for "naturlig eksperiment" siden det tar utgangspunkt i virkelige situasjoner der enhetene faller "naturlig" inn i henholdsvis eksperimentgruppen (prosjektgruppen) og kontrollgruppen (sammenligningsgruppen), eksempelvis fordi et tiltak er selektivt definert som en følge av politisk-administrativt fattede vedtak. Slike naturlige eksperimentelle sitasjoner er interessante i evalueringsøye-med nettopp fordi en rekke av validitetstruslene svekkes.

***

Vi vil i denne kartleggingen rette spesiell oppmerksomhet mot kvasi-eksperimentelle evalueringsdesign og eksemplifisere styrkene og svakheten ved disse med utgangspunkt i evalueringsstudier som er gjennomført i Norge. Her tar vi utgangspunkt i følgende kategorisering av eksperimentelle og kvasi-eksperimentelle studier:

---

9 Se for eksempel Mohr (1992) eller Shadish et al. (2002).

10 Ekstern validitet omhandler i vår sammenheng i hvilken grad man trygt kan generalisere på basis av evalueringen uavhengig av tid og situasjonsbestemte forhold (Mohr,1992).

11 Ref. Ringdal (2001:212).

0. **Randomiserte kontrollerte eksperimenter med eksperiment og-kontrollgruppe**[12]
1. **Komparative tidsserier for både tiltaks- og sammenligningsgruppen** (*Comprehensive longitudinal design with pre-, midterm, post- and ex-post observations on the project and comparison groups*)
2. **Sammenligningsgruppe med både før- og ettermåling** (*Pretest-posttest project and comparison groups*)
3. **Sammenligningsgruppe uten før-, men med midtveis- og sluttmåling** (*Truncated longitudinal pretest-posttest project and comparison group design*)
4. **Sammenligningsgruppe, men kun med sluttmåling for denne** (*Pretest-posttest project group combined with posttest analysis of project and comparison group*s)
5. **Sluttmåling for både tiltaks- og sammenligningsgruppen** (*Posttest project and comparison groups*)
6. **Tiltaksgruppe med før- og sluttmåling** (*Pretest-posttest project group – no comparison group*)
7. **Tiltaksgruppe med sluttmåling** (*Posttest project group only*)

[12] Her kan Campbell & Stanleys "Post-test only control group design" være undervurdert slik Ringdal, (2001:217) antyder. Designen benytter randomiserte eksperiment- og kontrollgrupper etter at tiltaket er gjennomført. Dette kan være nyttig i de tilfellene det ikke foreligger eller det kan rekonstrueres baselinedata fra før tiltaket ble startet opp. Denne evalueringsdesignen vil være en randomisering anvendt på Bambergers design 5 "Posttest project and comparison groups". Hvis randomisering er relevant og forsvarlig å benytte, vil en med "Post-test only control group design" kunne unngå flere validitetstrusler som den ikke-randomiserte versjonen av "etterdesignet" har, f.eks.; seleksjonstrusselen, ulik historie for gruppene samt eventuelle effekter av måleinstrumentet over tid.

Tabell 1: Samlet oversikt over metodedesign med og uten sammenligningsgruppe

Design nummer	Før-måling (T_1)	Tiltak (over tid)	Midtveis-måling (T2)	Prosjektslutt måling(T3)	Etter-måling (T4)
1	P_1 C_1	X	$P_{2(n)}$ $C_{2(n)}$	P_3 C_3	P_4 C_4
2	P_1 C_1	X			P_2 C_2
3		X	$P_{1(n)}$ $C_{1(n)}$	P_2 C_2	
4	P_1	X		P_2 C_1	
5		X		P_1 C_1	
6	P_1	X		P_2	
7		X		P_1	

Kilde: Bamberger (2006)

Matching

Vedung (1997) knytter umiddelbart matchingteknikken til kvasi-eksperimentelle design. En kan oppnå økt kontroll i kvasi-eksperimentelle evalueringsdesign (der randomisering av enhetene i tiltaks- og sammenlingningsgruppen er umulig) ved å benytte ulike *matching*-teknikker som for eksempel *propensity scores.* Her forsøker en å identifisere et tilstrekkelig antall enheter i "par som er like på variabler som er viktige for eksperimentet" (Ringdal 2001:214) for så å sette sammen deltakere til eksperiment- og sammenligningsgruppen fra disse parene. Man konstruerer med andre ord en sammenligningsgruppe som ligner mest mulig på tiltaksgruppa.

Matching kan redusere seleksjonstruslene som de kvasi-eksperimentelle designene er sensitive for. De siste årene har særlig forskningsmiljøer innenfor statistikk og økonometri (deriblant Rosenbaum, Rubin og Heckman) videreutviklet ulike *matching*-teknikker. Ifølge Jakobsen og Kvitastein (2002:67) balanserer *propensity score*-teknikken sammensetningen av tiltaks- og sammenligningsgruppen godt og som regel bedre enn ved randomisert tilordning av enhetene i gruppene noe som kan redusere skjevheter i effektanalyser. Dette er som nevnt nyttig i forhold til seleksjonstrusselen mot den interne validiteten. Vi skal nå kort se nærmere på andre validitetstrusler i kvasi-eksperimentelle design.

Trusler mot intern og ekstern validitet

En styrke ved kvasi-eksperimentelle forskningsdesign er at de gir en god ramme for å vurdere validitetsproblemer i form av faktorer som kan gi opphav til alternative forklaringer på utviklingen og det endelige utfallet. Campbell og Stanley (1966) beskriver åtte generelle kategorier av trusler mot den interne validiteten for kvasi-eksperimentelle design. Disse er endret *historie*, *modning* for de som utsettes for tiltaket, endringer i *måleinstrumentet* som benyttes mellom før- og ettermålingen, effekt av *tidligere målinger*, *regresjon* (naturlig fall i målte verdier mellom før- og ettermålingen fordi før-nivået var ekstremt), *mortalitet* (selve frafallet av enheter mellom før- og ettermålingen påvirker effekten som måles), *seleksjon* (skjevhet i utvalget av enhetene i tiltaks- og sammenligningsgruppene) samt *interaksjon mellom seleksjon og modning*.

Relevante trusler mot intern validitet i for eksempel komparativt tidsseriedesign er; *seleksjon Q*-effekten, avvikende historie og eventuelt smitteeffekter mellom kontroll- og eksperimentgruppen. Disse faktorene diskuteres eksplisitt og om mulig kan en benytte mer robuste design som tar høyde for de aktuelle validitetstruslene.

Når det gjelder trusler mot *ekstern validitet* gir forskningsdesignene også en god mulighet til å vurdere om generalisering av resultatene er mulig og interessant. Tiltak som har vært vellykkede ett sted, gir ikke nødvendigvis samme virkning i en ny kontekst. Det er jo for eksempel slett ikke sikkert at en type arbeidsmarkedstiltak som var effektiv på 80-tallet i Norge vil gi en tilsvarende virkning i dag. Campbell og Stanley (1966) beskriver en tilsvarende sjekkliste over faktorer av betydning for å vurdere den eksterne validiteten i kvasi-eksperimentelle design som Kvitastein (2002:73pp) omtaler.

Etiske og juridiske begrensninger for kontrafaktiske design

Denne studien tar for seg hvordan kontrafaktisk analyse kan gjennomføres i evalueringssammenheng og i mindre grad om etiske og juridiske forhold som må vurderes nøye særlig når eksperimentelle design og randomisering benyttes ved innsamling av individdata.

Forskningsbaserte evalueringer må forholde seg til det forskningsetiske rammeverket og til personvernlovgivningen. I Norge har de nasjonale forskingsetiske komiteer utarbeidet etiske retningslinjer og datainnsamlingen i samfunnsvitenskaplige undersøkelser hvor individdata innhentes krever melding til Norsk samfunnsvitenskapelig datatjeneneste og eventuelt konsesjon fra Datatilsynet. I andre land eksisterer det dessuten profesjonsetiske retningslinjer for evalueringspraksis eksempelvis *The American Evaluation Association* (AEA)s retningslinjer for evalueringspraksis. Disse vektlegger blant annet at informanter skal vises respekt, verdighet og at datainnsamlingen forutsetter informert samtykke og konfidensialitet i rapporteringen. Etiske vurderinger medfører at mange problemstillinger ikke kan undersøkes med eksperimentelt design. Kvasi-eksperimentelle tilnærminger kan i flere tilfeller likevel være aktuelle.

1.4 Tidligere kartlegginger og analyser av evalueringspraksis

Det er tidligere gjennomført få og spredte analyser av norske evalueringsstudier og evalueringspraksis generelt og spesielt innenfor forvaltningsområdene næringsutvikling/ innovasjon/forskning samt bistand og utvikling. Få av disse metastudiene omtaler kontrafaktisk analyse eksplisitt. Flere av studiene viser imidlertid til en økende tendens til at evalueringene vektlegger resultater og virkninger av tiltak samtidig som de i liten grad drøfter metodiske utfordringer.

Departementenes evalueringspraksis

Statskonsult (1997) gjennomførte i samarbeid med Norsk institutt for studier av forskning og utdanning (NIFU) en kartlegging av departementenes evalueringspraksis i perioden 1994-1996 basert på et tilfeldig utvalg av 100 evalueringer. Totalt 206 evalueringer ble identifisert i denne perioden, og disse hadde en anslått totalkostnad på vel 200 millioner kroner. Noen av hovedkonklusjonene var at

> ”Evalueringene er knyttet til ulike virksomheter i de forskjellige departementene, men de fleste er orientert mot effektevalueringer knyttet til det kontrollansvar departementene har i forbindelse med iverksetting av offentlige tiltak. Det tyngste bruksområde for evalueringene er i forbindelse med omorganisering og reformvirksomhet. I hovedsak er det fagavdelingene som er initiativtakere til evalueringene. Det er kun i forbindelse med igangsettelse av evalueringer av reformer at den politiske ledelse markerer seg, men også her er det fagavdelingene som er mest aktive. (…) Hovedtyngden av evalueringene utføres innen instituttsektoren.”

Departementene rapporterte selv om tett oppfølging av evalueringene gjennom styringsgrupper, og at evalueringene ”først og fremst brukes til å forbedre eksisterende tiltak sammen med det å få nye ideer i en sak ” Statskonsult (1997, s.11). Kartleggingen baserte seg på departementansattes egne oppfatninger av sentrale sider ved evalueringsvirksomheten.

Statskonsult (2003) tar for seg departementenes bruk av evalueringer i daglig arbeid med utgangspunkt i ”de undersøkelser, analyser eller utredningsoppdrag som departementene selv refererer til på spørsmål om evalueringer”. Notatet ser nærere på evalueringspraksis innenfor *de tre forvaltningsområdene bistand, høyere utdanning og trygd*. Hovedkonklusjonene i Statskonsult (2003) er blant annet at evalueringer…

> ”… i stor grad brukes av departementene – også instrumentelt – som grunnlag for politikkutforming, læring og utvikling på området. Men variasjonene er store. Og det er gjennomgående slik at evalueringer ikke brukes direkte i beslutningsprosessene på den måten som økonomiregelverket legger opp til Statskonsult (2003, s 10)”.
>
> ”… fungerer som ett av flere instrumenter for å kontrollere utviklingen i sektoren og sjekke oppnådde resultater innenfor eget ansvarsområde”. Statskonsult (2003, s 20)”.

Statskonsult (2003:23) viser til at utfordringene i evalueringsvirksomheten for de tre forvaltningsområdene varierer betydelig. På *bistandsområdet* er det en utfordring knyttet til ”å kunne bruke evalueringene som gjøres – også i styringen av Norad”.

Kommunikasjon, samhandling og samarbeid mellom berørte parter vektlegges imidlertid gjennom hele evalueringsprosessen ifølge Statskonsult (2003:25). På *trygdeområdet* var "de økonomiske aspektene ved ytelsene bedre ivaretatt enn de sosiale aspektene" samtidig som Statskonsult fant at det "innenfor trygdesektoren settes (…) sterkere fokus på forebygging, nye virkemidler og tverrsektorielt samarbeid" i evalueringssammenheng. Innenfor *høyere utdanning* så "departementet en generell utfordring knyttet til mangfoldet av evalueringer og mengden av informasjon" for oppfølging og styring av sektoren. Statskonsult trekker dessuten fram "internasjonale sammenlikninger og kunnskap om kvalitetsnivå og standard ved institusjonene" som voksende områder for evalueringer i sektoren.

Evalueringspraksis innenfor bistands- og utviklingsfeltet

Jerve og Villanger (2008) vurderer på oppdrag av Norad, norske evalueringsstudier som vektlegger *impact* (langsiktige virkninger) av norsk bistand og utviklingssamarbeid. Jerve og Villangers analyse er basert på utvalgte bistandsevalueringsoppdrag etter 1996. Metaanalysen vurderer mandatet/oppdragsbeskrivelsen for evalueringene, relevansen av metodevalget i forhold til formålet, og implementeringen av metodedesignen samt om denne gir det nødvendige grunnlag for konklusjonene som trekkes. Metaanalysen viser bredde i de valgte metodedesignene (fra økonometrisk analyse av surveydata til kvalitative vurderinger basert på foreliggende dokumentasjon). Analysen betoner betydelige utfordringer for fremme kvaliteten i effektevalueringer på bistands- og utvikingsfeltet. I flere tilfeller er effekten av tiltaket som skal vurderes (jf. mandatet/oppdragsbeskrivelsen) umulig å identifisere blant annet fordi man ikke kan isolere fra andre viktigere faktorer. Dessuten påpekes det at valgte metodetilnærminger ofte er svakt utviklet eller brukt på en overfladisk måte på grunn av manglende ressurser. Samtidig beskriver metaanalysen eksempler på at det er mulig å implementere robuste effektevalueringsdesign som kan isolere og måle reelle effekter av et bistandsprogram innenfor en realistisk budsjettramme.

Forss og Bandstein (2008a) beskriver evalueringer på bistands- og utviklingsfeltet internasjonalt i perioden 2004-2007 på basis av en gjennomgang av 80 tilfeldig utvalgte evalueringsrapporter registrert av medlemslandene/ organisasjonene i *Developement Assistent Comittee*, DAC, (OECD). Rapportene er i all hovedsak prosjekt- og programevalueringer på land- eller regionnivå. Nesten 2/3 av prosjektene og programmene var ikke avsluttet da evalueringene pågikk. Mindre enn 1/3 av evalueringene rapporterte resultater og disse evalueringene ble kun unntaksvis gjennomført flere år etter at aktivitetene/tiltakene var avsluttet. Likevel har omkring 60 av de 80 evalueringene "effectiveness" og "impact" som ett av flere evalueringskriterier i sitt mandat/ oppdragsbeskrivelsen – i tillegg til både "efficiency", "sustainability" og "relevance"). Forss og Bandstein ser nettopp den store bredden av DACs evalueringskriterier i mandatene til enkeltevalueringene som hovedforklaringen til at det er gjennomført svært få kontrafaktiske analyser på feltet internasjonalt i denne perioden. Kun én evaluering hadde eksperimentelt design mens et fåtall andre benyttet kvasi-eksperimentelt design. Til sammen benyttet under sju prosent av de utvalgte evalueringene "counterfactual analysis".

Den dominerende metodetilnærmingen i evalueringene var derimot dokumentanalyse og intervjuer. Observasjoner og lengre feltbesøk (studier) var også utbredt.

I Sverige har Forss et al. (2008b) videre vurdert 34 evalueringsrapporter fra Sida (Styrelsen för internationellt utvecklingssamarbete) fra perioden 2003-2005. Evalueringene vurderes i forhold til de fem evalueringskriteriene i OECD/DAC som er viktige i bistands-evalueringer internasjonalt. Selv om effekter og langsiktige virkninger er et av fem kriterier her og i de fleste mandatene for SIDA evalueringene, har ingen av disse benyttet eksperimentelt eller kvasi-eksperimentelt metodedesign. Metaanalysen påpeker et generelt behov for kvalitetssikring og sterkere empirisk basis i de aktuelle evalueringene, særlig gjelder dette evalueringene som skal vurdere effektivitet i forhold til måloppnåelse samt langsiktige virkninger.

Evalueringspraksis innenfor næringsutvikling og innovasjon

På oppdrag av Distriktenes utbygningsfond kartla Rynning (1990) empiriske evalueringer av *bedriftsrettede virkemidler* i Norge som var gjennomført i blant annet i forhold til målsetning, design, resultater og effekter. Omkring 50 studier i tidsrommet 1979-1987 ble tatt med. Rynning (1990, s.64) skriver at vi har

> ”ennå relativt få solide empiriske baserte argumenter for våre synspunkter på effekter av distriktspolitiske virkemidler. Dels skyldes dette manglende kjennskap til de studier som er blitt gjort. Dels vet vi altfor godt hvor vanskelig det er å gjennomføre evalueringsstudier” (…) Hittil har de fleste studiene forsøkt å påvise økt sysselsetting som følge av distriktspolitiske virkemidler, uten til å ta hensyn til hvilken konjunkturutvikling som har rådet, hvilken størrelse og type bedrifter som har dominert søkermassen, eller hvor stabil den eventuelle økte sysselsettingen kan forventes å være.” (…) Designmessig har forskerne oftest valgt standardløsninger og vært mer styrt av datatilgjengelighet enn av selve problemstillingen (…) forskerne har i liten grad diskutert metodologi i studiene. Dette er uheldig for etterfølgere og kolleger – som ikke får anledning til å lære fra andres erfaringer. (…) Forsøk på aktivt å kontrollere effekten av mulige feilkilder må planlegges i forveien, og konklusjonene må modereres etter hvor alvorlige feilkildene er.

Om effektene skriver Rynning (1990, s 66) følgende:

> ”Sysselsettingseffekter av DU-støtte er ofte uklart dokumentert i evalueringsstudiene. På 1980-tallet ser effektene også ut til å ha avtatt. Men veksten i hele økonomien er lavere (…) I hvilken grad støtte har bidratt til å opprettholde arbeidsplasser, er vanskelig å anslå (…) stabil sysselsetting, må være basert på lokale prosjekter og lokal tillit”.

Rynning (1990, s. 67) anbefaler bedre planlegging av framtidige evalueringsstudier:

> ”Når det gjelder å fremme videre forskning om effekter av distriktspolitiske virkemidler, har bedre planlagte, bedre gjennomførte og mer forsvarlige analyserte og rapporterte evalueringsstudier i sin alminnelighet allerede blitt nevnt. Flere grundig planlagte studier vil bidra til på øke almennkunnskapen om effekter av distriktspolitikken, og skape et bedre grunnlag for å sammenligne resultatene studiene i mellom.

Technopolis' (2001) gjennomgang av ulike metoder for å måle den relative effektiviteten ved ulike teknologipolitiske virkemidler i ni ulike land (deriblant Norge) omtaler ikke kontrafaktisk analyse i det hele tatt. Nytte-kostnadsanalyse av resultater midtveis i ulike innovasjonsprogram er derimot en utbredt metode på feltet som beskrives som utilstrekkelig i Technopolis' kartlegging. Kritikken mot denne tilnærmingen rettes særlig mot problemet med å *isolere effekter av ulike virkemidler*. Et annet utbredt problem med nytte-kostnadsanalyser i effektevaluering på innovasjonsfeltet er ifølge Technopolis (2001:71) *tidsforsinkelsen* fra et innovasjonsprogram/ utviklingstiltak iverksettes til eventuelle effekter kan måles for de direkte involverte og i siste instans for samfunnet. Et tredje og generelt utbredt problem med evalueringer av innovasjons- og utviklingsaktiviteter på begynnelsen av 2000-tallet var avhengigheten av *selvrapporterte unøyaktige estimater fra enkeltbedriften* omkring resultater og utfall av ulike satsinger.

Økonometrisk estimering er en annen metode for effektvurdering som omtales og kritiseres av Technopolis (2001:72) primært for å gi en utilfredsstillende approksimasjon av effekten som kan følge av et offentlig tiltak/virkemiddel for de involverte bedriftene. Ifølge rapporten er det svært vanskelig å kunne estimere effekten på grunn av intrikate samspill av innovasjonsfaktorer som påvirker effekten tiltaket kan gi, men som sjelden (kan) inkluderes i estimeringsmodellen.

Problemene med nytte-kostnadsanalyse og økonometrisk estimering i forhold til innovasjon- og næringsutviklingstiltak, kan forklare at ingen av de ni landene bruker disse metodiske verktøyene alene, ifølge Technopolis. Selv argumenterer de for å benytte kvantitative tilnærminger i kombinasjon med kvalitative metoder som casestudier og brukerundersøkelser. Samtidig anbefales systematisk planlegging og gjennomføring av evalueringsinnsatsen på feltet gjerne basert på langsiktige monitoreringsopplegg som gir en rekke indikatorer og målepunkter underveis og etter tiltakene er avsluttet. Norges forskningsråds utvikling av Provis-databasen i samarbeid med Møreforskning samt britiske "UK Business Link surveys" trekkes fram som tidlige eksempler slike verktøy.

Vinnova (2007) sammenfatter nyere svenske satsinger på å utvikle effektevalueringer på innovasjon og næringsutviklingsfeltet siden 2001. Disse blir gjennomført fra 5-20 år etter programmets avslutning. Verket för innovationssystem (Vinnova) samler løpende inn data fra brukerne som kan benyttes senere i effektevalueringer. Metodetilnærmingen i effektstudiene er bredspektret, men videreutvikling av nytte-kostnadsanalysemodeller i samarbeid med Møreforskning framstilles som sentrale for å vurdere langsiktige, overgripende effekter på samfunnsnivå av de offentlige satsingene på "behovsmotivert" (brukerstyrt) forskning.

I dette innledende kapitlet har vi beskrevet oppdraget, begrepet "kontrafaktisk analyse" og ulike kontrafaktiske metodedesign med utgangspunkt i Bamberger (2006). Dessuten har vi kort berørt matchingteknikken *propensity scores* og trusler mot intern og ekstern validitet samt etiske og juridiske begrensinger for kontrafaktiske design. Til slutt har vi omtalt tidligere kartlegginger og analyser av evalueringspraksis innenfor bistands- og utviklingsfeltet samt innenfor næringsutvikling og innovasjon. Disse kartleggingene peker mot at kontrafaktiske analyser er sjeldne i slag og da spesielt analyser som tar i bruk eksperimentelle metodedesign.

2 Kartlegging av kontrafaktiske analyser innenfor bistand, innovasjon og næringsutvikling

I dette kapitlet viser vi hvordan vi fant fram til enkeltevalueringene som analysen bygger på samt gir en innledende beskrivelse av disse før vi i neste kapittel vurderer styrker og svakheter ved de valgte metodedesignene.

2.1 Søk og utvelgelse av evalueringsstudiene

Vi benyttet ulike elektroniske søk i tillegg til informasjon fra nettverket til EVA-forum for å identifisere relevante evalueringer med kontrafaktisk design i perioden 1993-2008:

a) Elektroniske søk etter publikasjoner på hjemmesidene til Forskningsrådet, Innovasjon Norge, Norad, Riksrevisjonen, SSØ, samt Utenriksdepartementet
b) Elektroniske søk i bibliografiske databasene Bibsys, ISI Web of Knowledge og spesifiserte søk på internett (deriblant med Google Scholar)
c) Forespørsler om aktuelle evalueringsrapporter til blant andre referansegruppen.

Publikasjonssøkene var rettet mot å identifisere evalueringsstudier generelt og særlig ”kontrafaktisk analyse” og ”effektevaluering”. De innledende publikasjonssøkene a) identifiserte en lang rekke evalueringsstudier i den aktuelle perioden, flere effektevalueringer, men ga men ingen treff på “kontrafaktisk”. Oversikten over evalueringene finnes i referanselisten.

Bibsys-søkene bekreftet at evalueringer kun unntaksvis indekseres etter metodisk innretning med mindre dette framgår avtittelen. I den aktuelle perioden hadde ingen evalueringer (men en hovedfagsoppgave om stemmerettutvidelse) ”kontrafaktisk analyse i tittelen”. Bibsys-søkene identifiserte ingen evalueringer med ”kontrafaktisk analyse” men flere publikasjoner med ”effektevaluering” i tittelen. To av disse var innenfor innovasjon og næringsfeltet. (I tillegg ble det registrert effektevalueringer blant annet innen arbeidsmarkedsfeltet og på miljøområdet). Supplerende Bibsys-søk med utvalgte oppdragsgivere i kombinasjon med termen ”evaluering” ble dessuten gjennomført. Videre gjennomførte vi målrettede søk på termen kontrafaktisk analyse” og effektevaluering med ISI Web of Knowledge samt søkemotoren Google Scholar. Dette ga flere relevante treff særlig på arbeidsmarkedsfeltet og medisin og helsefeltet, og enkelte innenfor næringsutvikling. (Enkelttreff førte oss videre til andre relevante publikasjoner fra samme fagmiljø. Vi fikk dessuten c) forslag på relevante evalueringsrapporter innenfor bistandsfeltet etter en henvendelse formidlet av Norads evalueringsavdeling.

Det ble lagt ned betydelig ressurser i søkeprosessen uten at dette identifiserte et betydelig antall norske evalueringsstudier som benytter kontrafaktisk analyse. Vi kan dessverre ikke helt utelukke at det likevel er gjennomført evalueringsstudier med kvasi-eksperimentelt design på de utvalgte forvaltningsområdene som vi ikke er kjent med. Tidligere kartlegginger og andre studier som vi er kjent, med omtaler imidlertid heller ikke

evalueringer med kontrafaktisk design før overgangen til 2000-tallet. Dette kan ha sammenheng med den metodeutviklingen som har funnet sted spesielt innenfor økonometriske og statistiske fagmiljøer siden andre halvdel av 1990-tallet. Vi baserer det videre arbeidet på åtte evalueringsstudier som alle er utført fra og med 2000.

2.2 Beskrivelse av de utvalgte evalueringsstudiene

Evalueringer fra bistands- og utviklingsfeltet med kontrafaktisk design

Design 0: Randomisert eksperiment med *post-test only control group design* (kombinert med kvalitative intervjuer og informasjon fra tidligere evalueringer)

Brochgrevink A et al. (2003): *Credible Credit. Impact Study of the Dedebit Credit and Savings Institution (Decsi), Tigray, Ethiopia.* Norwegian Institute of International Affairs.

Denne evalueringen tar for seg effekten av mikrokredittlån for fattigdomsreduksjon i enkelthusholdninger samt utviklingen i berørte lokalsamfunn. Surveydata fra besøksintervjuer i ulike husholdninger og lokalsamfunn som var lånetagere høsten 2002, sammenlignes med tilsvarende data fra husholdninger som ikke var lånetakere og en tredje gruppe som tidligere hadde vært låntagere. Deltakerne i de tre gruppene ble valgt ut tilfeldig med basis i låneinstitusjonens registre og supplerende lokale registre samtidig som det ble tatt høyde for ulikheter i husholdningenes velstandsnivå. I tillegg til surveydataene ble det også gjennomført kvalitativ intervjuing som ga 35 supplerende casebeskrivelser. Evalueringsrapporten betoner nytten av datatriangulering fordi det er vanskelig å isolere effekten av lånene for fattigdomsreduksjon i en kontekst med krig, tørke og andre sosiale prosesser som påvirker utviklingen i husholdningene og lokalsamfunnene som studeres. Metodevalg og begrensninger behandles utfyldig i evalueringsrapporten.

Design 1X: Designen har fellestrekk med design 1, Komparative tidsserier for både tiltaks- og sammenligningsgruppen (Comprehensive longitudinal design with pre-, midterm, post- and ex-post observations on the project and comparison groups)

Holden, S. (2009): ”Impacts of Low-Cost Land Certification on Investment and Productivity” in *Amer.J. Agr. Econ*.

Metodeopplegget i studien er imidlertid ikke detaljert nok beskrevet i artikkelen til at vi med sikkerhet kan karakterisere dette som et eksempel på design 1. Dette er imidlertid en kontrafaktisk analyse som benytter økonometrisk estimering av virkningen av etiopisk jordreform (sertifisering av brukerrettigheter for husholdninger til registrerte landbruks-areal) i forhold til produktivitet og investeringsvillighet for videreutvikling av arealet.

Estimeringen er basert på paneldata fra tre surveyer over en åtte-årsperiode før, under og etter gjennomføringen av reformen. Utvelgelse av enhetene (husholdninger) i regionen er basert på *propensity score matching* av husholdninger avhengig om disse hadde sertifikat for landbruksareal eller ikke. Artikkelen gir en god dokumentasjon på hvordan

estimeringen er gjennomført, men gir i liten grad utdypende informasjon om den bakenforliggende metodedesignen. Artikkelen er akseptert for publisering i et landbruksøkonomisk tidskrift rettet mot fagfeller på feltet.

Evalueringer fra innovasjons- og næringsutviklingsfeltet

Design 5: Sluttmåling for både tiltaks- og sammenligningsgruppen (Posttest project and comparison groups)

Alsos, GA et al. (2006): *Flere og bedre bedriftsetableringer? Evaluering av Innovasjon Norges stipendordninger 1999-2005*. NF-rapport nr. 11/2006. Nordlandsforskning. Bodø

Analysen av hovedproblemstilling 3 – hvilken betydning har Innovasjon Norges bistand hatt for de resultater som er oppnådd? – bygger på en elektronisk survey i 2006 til et utvalg stipendiemottaker etter 1999 samt til en kontrollgruppe av tilsvarende nyetablerte bedrifter (i uke 21-24 i 2002) som ikke har fått stipend fra Innovasjon Norge. Tiltaksgruppen ble etablert ved hjelp av registerdata og telefonscreening. Studien trekker også på kvalitative casestudier med gruppeintervju og personlig intervjuer gjennomført i to fylker (for å besvare kvalitativt orienterte delproblemstillinger). Prosjektrapporten gir en god framstilling av problemstilling også gjennom analysemodell som beskriver progamteorien/ utfallslinjen. Metodedesignen kunne muligens vært styrket ved å benytte *matching* av tiltaks- og sammenligningsgruppen.

Design 5: Sluttmåling for både tiltaks- og sammenligningsgruppen (Posttest project and comparison groups)

Alsos, GA et al. (2000): *SND i Distrikts-Norge. Evaluering av de bedriftsrettede distriktspolitiske virkemidlene.* NF-rapport nr. 21/2000. Nordlandsforskning. Bodø

Deler evalueringen berører effekten av virkemidlene for bedriftene som har mottatt støtte samt effekter for utvikling av næringsstrukturen. Det benyttes registerdata samt survey-data fra to sammenligningsgrupper, bedrifter som ikke søkte om støtte i perioden samt bedrifter som søkte, men fikk avslag. I tillegg benyttes casestudier og andre intervjudata. Rapporten beskriver evalueringsmetodikken, deriblant sammenligningsgruppene og deres representativitet relativt inngående.

Design 6: Tiltaksgruppe med før- og sluttmåling (Pretest-posttest project group – no comparison group)

Alsos, GA et al. (2007*): Evaluering av SkatteFUNNs adferdsaddisjonalitet. I hvilken grad har SkatteFUNN ført til endret FoU-adferd i bedriftene?* NF-rapport nr. 13/2007. Bodø

Evalueringen benytter panel-surveyundersøkelse om bedrifters utvikling fra de har fått godkjent til etter at de har fullført et utviklingsprosjekt med støtte fra SkatteFUNN-

programmet. Ingen sammenligningsgruppe er benyttet fordi SkattteFUNN er rettighetsbasert og det finnes ifølge forfatterne ingen naturlig kontrollgruppe. Imidlertid vurderes de observerte endringene avhengig av om bedriftene ”oppgir å ville ha satt i gang prosjektet også uten SkatteFUNN” i motsetning til de som oppgir tiltaket ”var avgjørende for at prosjektet ble gjennomført”.

Ifølge forfatterne spør man tradisjonen tro ”dem som vet”. Disse vil imidlertid ha egeninteresse av å *svare strategisk*, for å opprettholde tiltaket, slik at verdien av denne designen er diskutabel. Rye (2002) viser på sin side til at betydningen av strategiske svar omkring addisjonaliteten fra utviklingsprosjekter med offentlig støtte er overvurdert.

Tidsseriedesign er følsomme for *bortfall* av respondenter etter at før-målingen er foretatt. Bortfallsanalysen i evalueringsrapporten viser at omkring 1/3 av respondentene som besvarte for-undersøkelsen ikke besvarte etter-undersøkelsen selv etter gjentatte purringer. Systematiske skjevheter i utvalg av respondenter vil være problematisk for effektivitetsvurderingene og drøftes derfor eksplisitt i rapporten.

Design 7x: Tiltaksgruppe med ”sluttmåling” – underveis; en variant av ”Posttest project group only”

Clausen TH og Rasmussen, E. (2008): *Resultatevaluering av SIVAs industri-inkubatorprogram*. Nordlandsforskning. Bodø. NF-rapport nr. 04/2008. Nordlandsforskning. Bodø

Industri-inkubatorprogrammet startet i 2004 og besto i 2008 av 16 bedriftsinkubatorer. Programmet er beregnet til å løpe ut 2011. To av disse var i en tidlig oppstartsfase og inngår derfor ikke i evalueringen. Konkurransegrunnlaget definerte oppdraget som ”resultat- og effektevaluering av SIVAs industri-inkubatorsatsing”.

Valide effektevalueringer kan imidlertid sjelden gjennomføres før det har gått flere år etter tiltaket er gjennomført. Evalueringens tidshorisont var tre måneder. Før-måling foreligger ikke, og ingen kontrollgruppe er benyttet. Valget av design 7 er en naturlig med de begrensingene som er satt. Med mer tid og ressurser til rådighet (enn tre sommermåneder/ 480.000 NOK inkl. mva), kunne evalueringen for eksempel benyttet sammenlignings-gruppe (eventuelt randomisert kontrollgruppe). Men fortsatt vil en mangle målepunkter etter programmet er avsluttet for å kunne besvare spørsmål om inkubatorenes effekter fra utlysningen. Evalueringsoppdragets begrensninger medfører at effektvurderingene omkring programmets addisjonalitet og måloppnåelse, alene baseres på intervjuer og spørreskjemadata med de involverte i inkubatorbedriftene og bedrifter med tilknytning til disse. Dette problemet drøftes, men kan ikke løses i rapporten.

Vi har i dette kapitlet beskrevet søk og utvelgelse av evalueringsstudiene som senere ble beskrevet med utgangpunkt i ulike eksperimentelle og kvasi-eksperimentelle design. I det påfølgende kapitlet vurderes styrker og svakheter ved metodedesignen for evalueringene.

3 Erfaringer, muligheter og utfordringer ved bruk av kontrafaktisk analyse i evalueringer

I forrige kapittel beskrev vi de utvalgte evalueringsstudiene i kartleggingen. Her vil vi oppsummere hvilke metoder som har vært brukt og diskutere styrker og svakheter ved disse. Deretter vurderer vi valget av kontrafaktiske analysedesign i lys av internasjonale erfaringer og standarder for evalueringsvirksomhet. Kapitlet avrundes med en oppsummering av de viktigste lærdommene.

3.1 Styrker og svakheter ved anvendte metoder

Hvilke metoder har vært brukt?

Vi har foran vist eksempler på bruk av følgende metodedesign i de identifiserte effektevalueringsstudiene:

0. **Randomisert eksperiment med kontrollgruppe**, i vårt tilfelle *post-test only control group design*
1. **Komparative tidsserier for både tiltaks- og sammenligningsgruppen** (*Comprehensive longitudinal design with pre-, midterm, post- and ex-post observations on the project and comparison groups*)
5. **Sluttmåling for både tiltaks- og sammenligningsgruppen** (*Posttest project and comparison groups*)
6. **Tiltaksgruppe med før- og sluttmåling** (*Pretest-posttest project group – no comparison group*)
7. **Tiltaksgruppe med sluttmåling** (*Posttest project group only*)

Vi har imidlertid *ikke* funnet eksempler på anvendelse av design 2,3 og 4 som alle benytter sammenligningsgrupper.

2. **Sammenligningsgruppe med både før- og ettermåling** (*Pretest-posttest project and comparison groups*)
3. **Sammenligningsgruppe uten før-, men med midtveis- og sluttmåling** (*Truncated longitudinal pretest-posttest project and comparison group design*)
4. **Sammenligningsgruppe, men kun med sluttmåling for denne** (*Pretest-posttest project group combined with posttest analysis of project and comparison groups*)

Design 2 forutsetter før-måling for både tiltaks- og sammenligningsgruppen og derved en utvidet planleggings- og ressursinnsats som sjelden vil være tilgjengelig. Imidlertid er det som Bamberger (2006) beskriver, mulig å rekonstruere før-målinger fra andre (sekundære)

datakilder som foreligger eventuelt benytte tilgjengelige registerdata når forholdene ligger til rette for det. Design 3 og 4 krever ingen før-måling for sammenligningsgruppen og skulle være mulig å benytte i flere effektevalueringer med en kortere planleggingshorisont. Designene kan for eksempel benyttes i forlengelsen av en prosessevaluering (underveisevaluering) Det vil da kanskje foreligge nok informasjon til å rekonstruere en før- og eventuelt midtveis-måling for prosjektgruppen.

Hva er styrkene og svakhetene ved de ulike metodene?

Drøftingen er her avgrenset til de fem identifiserte designene og tar utgangspunkt i metodelitteratur som er beskrevet innledningsvis og da spesielt Mohr (1992), Vedung (1997) og Bamberger (2006).

0. **Randomisert eksperiment med kontrollgruppe,** i vårt tilfelle, *post-test only control group design.* Når en benytter randomisering ved utvelgelsen av eksperiment- og tiltaksgruppe, øker generelt kontrollmulighetene over andre faktorer som kan gi opphav til målbare effekter ut over selve tiltaket. Samtidig øker gjerne truslene mot den eksterne validiteten fordi designen fort kan bli lite realistisk i forhold til den sosiale situasjonen som tiltaket inngår i. Eksperimentelle design vil dessuten ofte øke kostnadene og tidsrammen for gjennomføringen av evalueringen.
   Post-test only control group designen er imidlertid blant de mer anvendelige siden den ikke benytter før-målinger samtidig som ettermålingen ikke er følsom for validitetstruslene fra historie, modning og endringer i måleinstrumentet. Randomiseringen skal sikre at eksperiment- og tiltaksgruppen er ekvivalente. Følgelig kan det avledes fra designen at verdien på den uavhengige variabelen for de to gruppene også er identisk ved baseline (målepunktet før tiltaket iverksettes) ifølge Vedung (1997:179).

1. **Komparative tidsserier for både tiltaks- og sammenligningsgruppen** (*Comprehensive longitudinal design with pre-, midterm, post- and ex-post observations on the project and comparison groups*) Ifølge Bamberger (2006:213) gir denne kvasi-eksperimentelle evalueringsdesignen potensielt opphav til færre og mindre alvorlige validitetstrusler når den er implementert riktig. Han betoner at det ikke finnes noe perfekt design og at validiteten må vurderes på ulike trinn i evalueringen. Bamberger framstiller design 1– komparative tidsserier som den sterkeste, men samtidig mest (tid)krevende kvasi-eksperimentelle metodedesignen. Den er velegnet for å analysere virkningen av uprøvde tiltak som på sikt vurderes brukt i stor skala. Seleksjonsskjevhet, historie- og smitteeffekter mellom tiltaks- og sammenligningsgruppen er blant de potensielle validitetstruslene for design1 jf. Mohr (1992:146).

5. **Sluttmåling for både tiltaks- og sammenligningsgruppen** (*Posttest project and comparison groups*) er en variant av design 0 (*post-test only control group design)* som kan benyttes der randomisering er umulig og det ikke eksisterer eller kan rekonstrueres

en før-måling. Designen er anvendelig når evalueringen iverksettes underveis eller kort tid etter tiltaket er avsluttet. Uten bruk av *matching* er denne designen truet av betydelige seleksjonseffekter i tillegg til historieeffekter (eksempelvis betydningen av økonomiske, politiske hendelser som inntraff i løpet av tiltakets levetid)

6. **Tiltaksgruppe med før- og sluttmåling** (*Pretest-posttest project group – no comparison group*). Denne elementære før-etter designen er følsomt for de fleste validitetstruslene, men kan brukes dersom det er umulig å etablere en sammenligningsgruppe. Dette kan være tilfelle når tiltakene er universelt tilgjengelige. Design 6 er ifølge Bamberger (2006:222) egnet til å vurdere tiltak der man allerede kjenner effekten av tilsvarende tiltak benyttet i lignende omgivelser. Han understreker likevel at designen bør baseres på en velutviklet programteorimodell samt kombineres med innsikt fra prosessevalueringer. Bamberger beskriver designen som uegnet de man ønsker presise estimater om betydningen av et mindre tiltak.

7. **Tiltaksgruppe med sluttmåling** (*Posttest project group only*). Flere, deriblant Bamberger viser til at denne designen er svært utbredt i effektevalueringssammenheng, men dette betyr slett ikke at den anbefales. Designen gir ingen kontrafaktisk analyse og er beheftet med en lang rekke alvorlige validitetstrusler. Uten en omfattende drøfting av disse truslene i den enkelte evalueringen er denne designen uinteressant og uegnet for å trekke slutninger om effekter av tiltak.

Design 5-7 er generelt de svakeste designene samtidig som disse gir lavere kostnader og stiller lavere krav til tilgjengelige data. Dette kan være viktig for å kunne gjennomføre evalueringer innenfor knappe tids- og kostnadsrammer samtidig som det er begrensende for hvor klare konklusjoner som kan trekkes om eventuelle effekter.

I hvilken utstrekning diskuteres metodevalgene i rapportene?

Evalueringsrapportene vi har tatt for oss er hovedsakelig rettet mot oppdragsgiver og metodespørsmålene er nedtonet. Tidsskriftsartikkelen vi har sett nærmere på, gir derimot utdypende informasjon om estimeringsmodellen som er benyttet.

Metodevalgene diskuteres i varierende, men til dels utilstrekkelig grad i rapportene. Vi finner imidlertid flere interessante eksempler på diskusjon av metodevalg og validitetstrusler i enkelte av rapportene, blant annet i rapportene fra Nordlandsforskning, og i evalueringen fra CMI om mikrokreditter: I dette eksempelet på ”design 0” (*Post-test only control group*) drøftes validitetsproblem med bakgrunn i at rangeringen på velstandsmålet for husholdningene ikke er gjennomført så systematisk og objektivt som ønskelig. Forfatterne vektlegger for øvrig betydningen av at funnene i denne eksperimentelle studien kan trianguleres med kvalitative data som ble innsamlet samtidig. Dette er av betydning når en i dette eksperimentet forsøker å beskrive effektene av et

selektivt økonomisk utviklingstiltak i en kontekst av ekstreme eksistensielle faktorer (tørke og en nylig avsluttet krig i nabolandet Eritrea som gir opphav til folkevandring).

Seleksjonsskjevhet, historie- og smitteeffekter mellom tiltaks- og sammenligningsgruppen er potensielle validitetstrusler for design 1 – komparative tidsserier jf Mohr (1992:146). Disse vurderes i begrenset grad i eksempelet med økonometrisk estimering av virkningen av etiopisk jordreform (sertifisering av brukerrettigheter for husholdninger til registrerte landbruksareal) i forhold til produktivitet og investeringsvillighet. Derimot beskrives matching-prosedyren for de to gruppene relativt detaljert av Holden et al. (2009).

3.2 Vurderingen av metodene i et internasjonalt perspektiv

I hvilken utstrekning samsvarer de utvalgte evalueringene med internasjonal standard?

Det foreligger ingen enhetlig internasjonal standard for evalueringsvirksomhet. Enkelte internasjonale og overnasjonale organer har imidlertid i likhet med nasjonale fagmiljøer, utarbeidet standarder ofte også benevnt som "retningslinjer" og "kriterier for god evalueringspraksis". Enkeltorganisasjoner som European Evaluation Society (EES) vektlegger i stedet for en felles europeisk evalueringsstandard, et mangfold av kulturelt betingede evalueringstradisjoner og tilrettelegger erfaringsutveksling mellom nasjonale evalueringsorganisasjoner i stedet for utvikling av en EES standard (Beywl, 2006:14).[13].

I dette mangfoldet av prinsipper og retningslinjer er det en utfordring å drøfte av hvordan de utvalgte evalueringsstudiene i kartleggingen samsvarer med "internasjonal standard". Vi begrenser oss her til noen sentrale formuleringer i retningslinjer fra OECD/ DAC, FN og EU.

OECD-organet DAC[14] (*Developement Assistance Comittee*) har nedfelt generelle retningslinjer for evalueringsvirksomheten og siden 1991 følgende fem evalueringskriterier for bistandsevalueringer (OECD 2008, 13):

1. Relevans; er bistandsaktiviteten tilpasset giverlandets mål, samarbeidslandets behov og prioriteringer?
2. Måloppnåelse (*effectiveness*); er målene for aktivitetene/ tiltakene innfridd?
3. Bærekraftighet (*sustainability*); vil aktiviteten kunne videreføres etter at tiltaket opphører?
4. Produktivitet (*efficiency*); brukes de tilgjengelige bistandsmidlene mest mulig effektivt for å innfri de ønskede resultatene?
5. Virkning (*impact*); langtidsvirkninger som er tilsiktet eller utilsiktede, direkte eller indirekte

[13] AEA uttaler på sin side eksplisitt i retningslinjene at disse er "utviklet in the context of Western cultures, particularly the United States, and may reflect the experiences of that context. The relevance of these principles may vary across other cultures, and across subcultures within the United States" AEA (2004, preface). AEA-retningslinjene er bygd på fem generelle prinsipper som berører a) systematisk, empiriske undersøkelser, b) evaluatorens kompetanse, c) evaluatorens integritet, d) evaluatorens respekt for interessentene samt e) ansvaret for befolkningen og dens velferd.

[14] OECD-organet DAC er et viktig forum for giverland og multilaterale organisasjoner på bistandsfeltet som vil styrke effektiviteten for midlene som gis og samtidig koordinere utviklingsaktivitetene.

Her er det særlig det siste kriteriet som er sentralt i forhold til kontrafaktiske tilnærminger som benyttes i sammenheng med effektevalueringer i etterkant av at tiltaket er avsluttet. DACs impact-kriterium berører:

> The positive and negative changes produced by a development intervention, directly or indirectly, intended or unintended. This involves the main impacts and effects resulting from the activity on the local social, economic, environmental and other development indicators. The examination should be concerned with both intended and unintended results and must also include the positive and negative impact of external factors, such as changes in terms of trade and financial conditions.

STANDARD versus PRAKSIS I

Vurderinger av både *positive, negative, forventede* så vel som *uforventede* effekter av det aktuelle tiltaket står sentralt i effektevalueringer. DAC -standarden omtaler ikke ulike design for effektevaluering. Betoningen av *eksterne faktorer* for virkningene som et tiltak måtte ha, tilsier imidlertid en kontrafaktisk metodedesign som inkluderer bruk av kontroll/ sammenligningsgruppe. Uten en slik design kan man vanskelig gjennomføre en interessant effektevaluering i tråd med DACs impact-kriterium[15].

I vår kartlegging finner vi flere eksempler på evalueringer som skal adressere effekter av et tiltak uten bruk av kontroll-/sammenligningsgrupper. Slike design er utbredte også internasjonalt. Bamberger (2006: 203) referer til at majoriteten i en metaevaluering av 67 studier i CARE International de siste to årene, kun var basert på tiltaksgruppe med sluttmåling (design 7). Lignende funn er gjort i de svenske SIDA-evalueringene jf. Forss og Bandstein (2008a).

DACs prinsipper for bistandsevalueringer (OECD, 2008: 7) betoner for øvrig; a) en eksplisitt evalueringspolicy for bistandsorganet, b) en upartisk og uavhengig evaluering, c) en åpen prosess med bred formidling av resultatene, d) bruk og oppfølging av evalueringen, e) nær kontakt mellom bistands- og mottakerorganisasjonen, og f) at evalueringen er med i planene allerede fra når tiltaket utformes.

STANDARD versus PRAKSIS II

Dette sistnevnte prinsippet er sentralt for effektevalueringer generelt. Når en evaluering er godt planlagt, vil det for eksempel være enklere å etablere en *baseline* (før-måling) forut for tiltaket iverksettes så vel som målepunkter underveis, ved avslutningen og etter at tiltaket er avsluttet. Med en rekke målepunkter, gjerne i kombinasjon med en prosessorientert underveisevaluering, vil datagrunnlaget for en robust effektevaluering være tilstede.

---

[15] Chianza (2008) argumenterer for behovet av å revidere de fem DAC-kriteriene, deriblant å se *effectiveness* i sammenheng med *impact*-kriteriet. Han ser nettopp *impact*-kriteriet som overordnet de øvrige kriteriene.

I 2006 utformet DAC egne *kvalitetsstandarder* for evalueringsarbeid (OECD, 2008:19) som bygger på DAC-prinsippene og skal gi retningslinjer for gjennomføring av evalueringer. Flere av formuleringene her er konkrete spesielt i forhold til analytisk tilnærming og metodiske krav, vi skal ta for oss enkelte av disse[16]:

STANDARD versus PRAKSIS III

Prinsippet om at "*the intervention logic (…) distinguishes between findings at the different levels: inputs, activities, outcomes and impacts*" er en nyttig retningslinje for evalueringer - som kan klargjøre og strukturere evalueringer for både oppdragstakere, oppdragsgiver og brukere av rapporten. Det er imidlertid ikke hovedregelen at evalueringsrapporter spesifiserer sammenhengen mellom problem, tiltak, aktiviteter og utfall. Kartleggingen ga eksempler på begge deler.

STANDARD versus PRAKSIS IV

DAC-standarden, understreker *kravet til innsiktsfull metodeforklaring med diskusjon av begrensninger og mangler* ved den valgte tilnærmingen og andre vurderinger omkring evalueringsstudiens validitet og reliabilitet. Herunder kommer også kravet til spesifikasjon av prosedyrene knyttet til utvalgstrekking, med vektlegging av eventuelle begrensinger i representativiteten i datagrunnlaget.

Få oppdragsrapporter vil innfri en slik standard fullt ut. Her konfronteres rapportforfatteren med informasjonsasymmetri-dilemmaet. Samtidig som begrensninger og mangler ved evalueringsmetodiske tilnærminger bør beskrives entydig, forutsetter avanserte reliabilitets- og validitetsdiskusjoner evalueringsmetodisk spesialisering og fordypning i samfunnsvitenskaplig metode. Utfordringen blir da å finne en hensiktsmessig balanse i evalueringsrapporten slik at sentral informasjon blir formidlet samtidig som detaljerte metodediskusjoner med fagfeller primært føres i andre kanaler.

FNs normer og standarder for evalueringsarbeid er utformet i 2004 og bygger i stor grad på OECD/DAC-standarden (United Nations Evaluation Group, 2005a, 2005b). Eksempelvis krever standard 4.9 at evalueringsmetoden beskrives utfyldig, på en "transparent" måte og da særlig eventuelle kritiske aspekter slik at brukeren kan "trekke

---

[16] *2.2 Intervention logic and findings*: The evaluation report briefly describes and assesses the intervention logic and distinguishes between findings at the different levels: inputs, activities, outcomes and impacts.
4.1 Explanation of the methodology used: The evaluation report describes and explains the evaluation method and process and discusses validity and reliability. It acknowledges any constraints encountered and their impact on the evaluation, including their impact on the independence of the evaluation. It details the methods and techniques used for data and information collection and processing. The choices are justified and limitations and shortcomings are explained.
4.2 Assessment of results: Methods for assessment of results are specified. Attribution and contributing/confounding factors should be addressed. If indicators are used as a basis for results assessment these should be SMART (specific, measurable, attainable, relevant and time bound).
4.4 Sampling: The evaluation report explains the selection of any sample. Limitations regarding the representativeness of the evaluation sample are identified.

egne konklusjoner om datakvaliteten" blant annet ut fra datainnsamlingsmåten og utvalget av respondenter.

Standard 4.11 adresserer spesifikt evalueringsdesignen og etiske betraktninger knyttet til denne som rasjonale for designen og "mekanismer som kan ivareta deltagernes konfidensialitet, verdighet og velferd". Dette er grunnleggende hensyn nedfelt i forskningsetiske rammeverk som er av betydning i ulike former for intervjuundersøkelser og eksperimentelle eller kvasi-eksperimentelle tilnærminger. Etiske normer diskuteres sjelden utfyllende i evalueringsrapporter, men tas gjerne for gitt.

For det tredje vektlegger også FNs evalueringsstandard 4.12 og 4.15 (i likhet med OECD/DAC – kriteriene) logiske sammenhenger i analysemodellen der funn og konklusjoner følger av utfall som er lenket, men separate fra tiltak/ aktiviteter.

EU-kommisjonens standarder (2004) knyttet til kvalitet i rapporteringen har lignende formuleringer[17] og også her finner vi igjen vektleggingen av logiske sammenhenger i analysemodellen. De øvrige prinsippene er særlig knyttet til administrative rutiner og forankring av evalueringsarbeid i EU-systemet.

Vi finner imidlertid mange likhetstrekk mellom nevnte internasjonale retningslinjer for evalueringsarbeid og da særlig mellom OECD/DAC og *FNs standarder* (United Nations Evaluation Group, 2005a), som eksplisitt refererer til førstnevnte. Samtidig er det ikke overraskende å se at disse standardene ikke innfris i enkeltevalueringer både internasjonalt og i en norsk sammenheng. Innfrielse av de enkelte standardene må imidlertid sees i lys av forutsetningene som ligger til grunn for den aktuelle evalueringen. Standarder eller retningslinjer gir likevel et nyttig utgangspunkt for å fremme kvaliteten for enkeltevalueringer generelt. Når det gjelder evalueringer med kontrafaktisk metodedesign nevner ikke de tre organenes foreliggende retningslinjer dette eksplisitt. I siste del av rapporten vil vi trekke inn enkeltstandarder i drøftingen, der dette er naturlig – blant annet i forhold til delspørsmålet om forbedringspotensial.

Hvilket forbedringspotensial finnes?

Gjennomgangen av evalueringsmetodisk litteratur beskrevet i det innledende kapitlet, samt retningslinjer for god evalueringspraksis indikerer et forbedringspotensial for evalueringer

---

[17] D2. The evaluation reports shall describe the purpose of the evaluation and its context and also the objectives, questions, procedures, results and reasoned conclusions of the evaluation, so as to make available the essential information in an easily understandable form.
D3. The report shall describe the information sources in such detail that the correctness of the information can be assessed. The data collected or selected shall be adapted to the methodologies used and be sufficiently reliable for the expected use.
D4. The prospects and reasoning on which interpretation of the results is based shall be described and explained. The results should follow on logically and be substantiated by data analysis and interpretations based on carefully-presented explanatory hypotheses.
D6. The conclusions and any recommendations shall be rigorous and not distorted by personal or partisan considerations. The recommendations shall be comprehensible, useful, applicable and detailed enough to be brought into effect.

av effekter og langsiktige virkninger av gjennomførte tiltak. Hele prosessen fra planleggingsstadiet til oppdragstakerens analyse og rapportering er da relevant hvis mer robuste forskningsbaserte evalueringsdesign skal kunne tas i bruk. Forbedringspotensialet konkretiseres i listen under:

1. Gjennomgangen viser at evalueringsdesign som brukes hyppig i dag kommer til kort når man skal vurdere langsiktige virkninger først og fremst fordi de ikke benytter *sammenligningsgrupper*. Uten slike grupper kan man ikke avlede effekter ut fra sammenligning av det faktiske utfallet for tiltaksgruppen versus det kontrafaktiske utfallet for sammenligningsgruppen.
2. Bruken av *før-, underveis- og ettermålinger* vil kunne styrke effektevalueringer. Med flere målepunkter vil ulike effekter av tiltak lettere kunne dokumenteres da man sjelden kan forutse på hvilke tidspunkt (etter at tiltaket er iverksatt) som de mer langsiktige effektene og virkningen vil kunne inntreffe.
3. Evalueringsmetodisk litteratur fra det siste tiåret, deriblant Shadish et al. (2002) vektlegger *bruken av randomiserte eksperimentelle* design - når mulighetene ligger til rette for dette – framfor kvasi-eksperimentelle design og sammenligningsgrupper med ulike statistiske teknikker for å kunne sikre at sammenligningsgruppen. Shadish mener å se et underforbruk av eksperimentelle design i effektevalueringer der forholdene ligger til rette for det. Oppfatningene om randomiserte eksperimentelle design er imidlertid delte. Bamberger et al. (2006) ser få anvendelser av slike design innenfor bistandsfeltet og vektlegger problemene ved slike design sterkere enn fordelene. De vil da også ofte være politisk og etisk uforsvarlig eller evalueringsfaglig sett uinteressante å benytte randomiserte eksperimentelle design i mange sammenhenger slik som Bamberger påpeker. Vår vurdering basert på litteraturgjennomgangen er likevel at en bør vurdere om randomiserte eksperimentelle design kan være relevante i den enkelte effektevalueringen. I tilfeller der dette er mulig og etisk forsvarlig vil det ofte være en styrke å benytte eksperimentell design som kan påvise effekter og virkninger av tiltak uten at truslene mot intern validitet kommer i forgrunnen. I mange tilfeller vil imidlertid slike design hverken være mulige eller interessante. Dessuten er truslene mot den eksterne validiteten ofte betydelige for eksperimentelle design slik at generalisering ut fra den aktuelle evalueringen blir problematisk.
4. *Gjennomført bruk av programteori med utfallslinje* kan være fordelaktig både for å strukturere oppdraget, evalueringsdesignen så vel som rapporteringen. Programteorien framkommer gjerne gjennom en kvalitativ analyse av problemet ut fra beskrivelser i policydokumenter og intervjuer med sentrale aktører. Ut fra dette kan en beskrive utfallslinjen i Mohrs terminologi (Mohr, 1992); dvs. hvordan tiltakets målsettinger, aktiviteter og resultater er tenkt å gi de ønskelige utfallene (løse det definerte problemet). Programteori med utfallslinje er et godt verktøy ikke bare i effektevalueringer, men i evalueringsstudier generelt. Programteori med utfallslinje er av betydning for både prosessevalueringer (f.eks. underveis i et program) så vel som for summative evalueringer (effektevalueringer).

5. *Bruk av registerdata og tidsseriedata fra monitoreringsopplegg er en underutnyttet kilde til mer robuste effektevalueringsdesign* (jf. bla. Vedung ,1997 og Bamberger, 2006). Et hovedproblem for effektevalueringer er gjerne målrettede og pålitelige før-målinger/ baselinestudier. Det er videre ofte problematisk å rekonstruere reliable før-målinger når tiltaket er modent for evaluering. Dersom det allerede fra start er tilrettelagt et monitoreringsopplegg for programmet/ tiltaket eller det er mulig å benytte innsamlede registerdata fra relevante målepunkter før/etter, vil evalueringsdesignen styrkes. Dersom registerdataene er sentrale i effektanalysen vil man gjerne samtidig ha tilgang til tilsvarende målinger for sammenligningsgruppen. Dette er en styrke sammenlignet med frafallsproblematikken for surveydata og sammenligningsgrupper (som ikke har vært eksponert av et tiltak)
6. Analyser basert på *registerdata/ monitoreringsdata* kan dessuten gi interessante oppfølgingsstudier over lengre tidsperioder slik at man kan oppnå en strategisk kunnskapsbygging innenfor et større politikkområde. Dette kan være av stor betydning i tillegg til å gjennomføre separate og avgrensede evalueringer av bestemte virkemidler som inngår i et større system som sjelden er gjenstand for evaluering.

I hvilken grad er det realistisk å gjennomføre kontrafaktiske analyser under de tilgjengelige ressursene for evalueringen?

Vi har i arbeidet med dette oppdraget ikke hatt systematisk tilgang til opplysninger om ressurser og tidsrammer for de utvalgte evalueringene. Derfor er det problematisk å gi et entydig svar på problemstillingen over. Vårt generelle inntrykk basert på utlyste evalueringer det siste tiåret, er imidlertid at robuste kontrafaktiske evalueringsdesign som kreves for å kunne evaluere effekter og virkninger av tiltak, ofte ligger utenfor den angitte tidshorisonten og ressursrammen. Vi har i delkapittel 2.2 gitt et eksempel på hvor det er et klart misforhold mellom ambisjonene i utlysningen og hvilke effekter det er mulig å måle innenfor den sterkt begrensede tids- og ressursramme i utlysningen.

Det er minst to måter til å håndtere dette misforholdet på; enten avvente effektevalueringen inntil det er akkumulert tilstrekkelige midler for at en slik evaluering kan gjennomføres når datagrunnlaget gir grunnlag for det, *eller* å gjennomføre en avgrenset resultatvurdering. Den andre opsjonen vil være en relevant måte å håndtere problemet på ifølge Bamberger (2006). Samtidig kan dette være mindre tilfredsstillende både for oppdragstaker så vel som oppdragsgiveren og berørte interessenter, siden evalueringen trolig vil munne ut i en rapport med svært mange forbehold omkring målbare effekter. Vår egen vurdering sammenfaller her til dels med Jerve og Vinnanger (2008:21). De vektlegger at det vil være bedre å utlyse midler til én omfattende og dyptpløyende evaluering som er tilrettelagt for å kunne gi interessant kunnskap om effekter, framfor å utlyse to svært avgrensede evalueringer som ikke kan gi robuste konklusjoner om eventuelle effekter.

Vedung (1997:157) gir på sin side et nyttig inntak for vurderinger omkring effektevalueringer med sin to-stegs prosess i vurderingen av det han benevner som *The Impact problem*. Første steg, "pre-evalueringen" er gjerne en vurdering av "evaluerbarhet" før den egentlige evalueringen kan påbegynnes. I pre-evalueringen vurderes; hvorvidt tiltaket er tilstrekkelig modent og om det er utviklet (eventuelt kan utvikles et teknisk design) for den kommende evalueringen – samtidig som det etableres et klima blant de sentrale interessentene slik at en lovende fullskala-effektevaluering kan gjennomføres som senere kan brukes i den videre politikkutformingen.

3.3 Sammenfatning av lærdommer fra evalueringene

Tredje hovedkomponent i oppdraget berører de viktigste lærdommene fra evalueringene i forhold til erfaring med sammenligningsgrupper samt muligheter og begrensninger ved kontrafaktiske analyser.

Hvilke erfaringer har man med bruk av sammenlikningsgrupper?

Inndelingen av enhetene i eksperiment- og kontrollgruppe versus tiltaksgruppe-/ og sammenligningsgruppe er en av hovedfordringene ved å tilrettelegge gode eksperimentelle og kvasi-eksperimentelle evalueringsdesign for kontrafaktisk analyse. Eksemplene i vår kartlegging viser også dette. Sammenligningsgruppen i kvasi-eksperimentelle design er imidlertid en forutsetning for å kunne etablere en kontrafaktisk tilstand for å gjennomføre effektanalysen – siden kausaleffekten utgjør forskjellen mellom den faktiske og kontrafaktiske tilstanden. Uten en sammenligningsgruppe (eller kontrollgruppe i en eksperimentell design) blir effektmålet lite robust fordi man vesentlig svekker mulighetene til å kontrollere betydningen av andre relevante faktorer som kan virke inn på den målte effekten av det aktuelle tiltaket.

Det kan være krevende å etablere gode sammenligningsgrupper og den kontrafaktisk tilstanden som kan benyttes for å beregne tiltakets effekt når det ikke foreligger "naturlige sammenligningsgrupper". Målsettingen er at sammenligningsgruppene holder høy kvalitet selv om de ikke kan bli perfekte (Shadish et al. 2002:6). Da må en unngå flere fallgruver. King og Zeng (2005) vektlegger at det er viktig å unngå sammenligninger basert på ekstreme kontrafaktiske tilstander. Faren for å velge en ekstrem kontrafaktisk tilstand er størst ved kvantitative tilnærminger ifølge King og Zeng som spesielt refererer til statsvitenskaplige studier av internasjonale relasjoner og demokratiutvikling. Sensitivitetsanalyser kan være et hjelpemiddel for å etablere gode sammenligningsgrupper når man har relevante kvantitative data. King og Zeng har imidlertid utviklet en dataapplikasjon som kan avgjøre avhengigheten av ulike modeller uten å måtte gjennomføre tidkrevende sensitivitetsanalyser.

Flere nyere kontrafaktiske studier internasjonalt tar i bruk ulike *matching*-teknikker for å sikre at sammenligningsgruppen benyttet i kvantitativt orienterte effektstudier er reelle og så like som mulig på alle andre målbare parametre enn tiltaket som evalueres.

Økonometriske og statistiske fagmiljøer har gjort betydelige framskritt de siste ti årene deriblant Heckman et al. (1998) og Pearl (2000). Enkelte norske fagmiljøer innen arbeidsmarkedsøkonomi og næringsutvikling har dessuten begynt å anvende denne statistisk orienterte tilnærmingen.

Andre spesialister innenfor forskningsbasert evaluering som Shadish et al. (2002:503) er imidlertid tilbakeholdne overfor denne utviklingen og vektlegger i stedet (kvasi)eksperimentelle *metodedesign* framfor statistiske kontrollteknikker som *matching* som de eventuelt vil bruke som en siste utvei:

> ”The position we do not like is the assumption that statistical adjustment techniques such as those advocated by statisticians and econometricians [propensity scores/ structural equation modeling] are so well developed that they can be used to obtain confident results in nonexperimental and weak quasi-experimental contexts.

Shadish et al. (2002) viser til at andre problemer må løses først. Man må for eksempel unngå å velge kontroll-/ sammenligningsgrupper ut fra sentrale registre når tiltaksgruppen er valgt ut fra en lokal kontekst. Ifølge Shadish et al. er det dessuten lite støtte i de siste tiårenes empiriske forskning for at statistiske kontroller kan løse problemer i longitudinelle (tidsseriebaserte) surveys hvor individer med ulike erfaringer kontrasteres for å estimere effekter ut fra erfaringsforskjellene. Hovedpoenget til Shadish et al. (2002) kan forenkles slik; jo sterkere evalueringsdesign, jo mer forståelige blir resultatene – i motsetning til analyser med komplekse økonometriske manipulasjoner hvor brukere uten denne spisskompetansen stilles overfor valget om å stole på metoden og resultatet eller ikke.

Bamberger (2006:234) trekker på sin side fram andre praktiske utfordringer for utvelgelsen av sammenligningsgruppe; utvelgelse av tiltaksgruppe på basis av administrative kriterier eller *selvseleksjon*. Begge disse formene for utvelgelse av deltakere i kvasi-eksperimentet skaper problemer for utvelgelsen av sammenligningsgruppen. Dersom deltakerne selv kan velge om de skal inngå i titaksgruppen, vil de trolig ha spesielle årsaker for å delta som kan være vanskelig å gjenspeile i sammenligningsgruppen. Derfor vil en med evalueringsdesignen gjerne forsøke å unngå selvseleksjon av deltakere. Dersom derimot tiltaket er selektivt tildelt etter administrative kriterier men store deler av populasjonen likevel faller inn under utvelgelseskriteriene, vil det også være vanskelig å lokalisere relevante enheter i sammenligningsgruppen. Hvis tiltaket som evalueres har pågått over tid, kan en løsning være å inkludere enheter (personer/grupper)som ble eksponert for tiltaket i en tilbakelagt fase (fase 1), men nå ikke gjør dette lenger i fase 2.

Etter å ha fokusert på erfaringene med sammenligningsgrupper, vil vi nå se nærmere på muligheter og begrensninger generelt for bruken av kontrafaktisk analyse i evalueringer.

Hvilke muligheter og begrensninger finnes når man skal gjennomføre kontrafaktiske analyser?

Erfaringene i den evalueringsmetodiske litteraturen peker mot følgende hovedpunkter i forhold til spørsmålet over:

1. *Mer krevende, robuste og relevante metodedesign vil kunne bidra til å styrke validiteten i effektevalueringer*. Dette igjen kan gi mer interessante funn og konklusjoner omkring langsiktige effekter og virkninger av offentlige tiltak.
2. *Mer avanserte design vil dessuten kreve godt planlagte evalueringer med ekstra ressursinnsats* fra oppdragsgiverens side som igjen kan gi gjennomtenkte datainnsamlingsopplegg, oppbygging av baselinedata og tidsserier som vil være en vesentlig styrke for framtidige effektevalueringer
3. *Mer avanserte metodedesign vil samtidig kreve økt metodekompetanse i evalueringsmiljøene*. Kompetanseheving i eksisterende miljøer og gjennom bidrag fra spesialiserte miljøer med spisskompetanse på feltet kan da være avgjørende.
4. *I enkelte tilfeller vil det ikke være mulig å gjennomføre kontrafaktiske analyser*. Dette er tilfelle hvis det umulig å etablere en sammenligningsgruppe for eksempel fordi tiltaket er universelt utformet for populasjonen. Da kan man heller ikke etablere en kontrafaktisk tilstand og gjennomføre kontrafaktisk analyse.
5. *Metodetriangulering gjennom kombinasjon av kvantitative og kvalitative data kan være til hjelp i evalueringsdesign der en har sammenligningsgrupper, men mangler før- og underveismålinger*. Bamberger (2006) anbefaler å avhjelpe validitetstruslene i metodedesignene ved å supplere med ulike typer datakilder (dersom en ikke kan velge en mer robust design). Design som kun benytter sluttmåling f.eks. kvantitative tverrsnittdata fra en survey, kan for eksempel sammenholdes med kvalitative data for å belyse alternative hypoteser om effektene av et tiltak.
6. *I enkeltevalueringer som er initiert uten tilstrekkelige forberedelser vil det sjelden være mulig å gjennomføre en kontrafaktisk analyse av effekter og virkninger*, fordi tiltaket fortsatt pågår/ er nylig avsluttet, og eventuelle langsiktige effekter og virkninger kanskje ikke er målbare før 3-5 år har gått.
7. *Vurderinger av tiltakets evaluerbarhet* er derfor et nyttig planleggingsverktøy både for oppdragsgiver så vel som evalueringsmiljøet har fått tilliten til å gjennomføre evalueringen. Vedung (1997) beskriver dette nærmere, og EU-kommisjonens standarder for evalueringsvirksomhet så vel som av OECD/DACs og FNs normer[18] vektlegger dessuten denne typen forberedelser av evalueringer.

[18] FN-normen (N7) understreker betydningen av å bygge en ”(…) evaluation approach into the plan. To safeguard independence this should be performed in an advisory capacity only”. Vurderingen av evaluerbarhet består i “…verifying if there is clarity in the intent of the subject to be evaluated, sufficient measurable indicators, assessable reliable information sources and no major factor hindering an impartial evaluation process” (United Nations Evaluation Group, 2005b).

4 Diskusjon av faktorer som kan bidra til videre metodeutvikling i Norge

Fjerde hovedkomponent i oppdraget – praktiske anbefalinger som kan bidra til videre metodeutvikling – tar for seg *oppdragsgivers rolle og mandatutformingen* mens forbedringsmulighetene for evaluator er særlig orientert omkring *kompetanse-oppbyggingsspørsmål*.

Kvitastein (2002) drøfter sju kritiske problemer for at evalueringer skal kunne være tjenelige for politikkutforming. Vi ser her nærmere på tre av disse som er mest relevante i denne sammenhengen;

> *"problemet med at ulike typer evalueringsoppdrag ikke skilles klart nok både i anbudsinnbydelser og senere fortolkning av resultater"*
>
> *"problemet med at den diskursen som foregår i etterkant av evalueringer, når tiltak eller programmet bringes opp på den politiske agendaen, tvinger fram betraktninger om tiltakenes effekt"*
>
> *"problemet med at den metodikk som evalueringer gjerne krever, er for lite kjent i basismiljøene ettersom evalueringsforskning krever tilnærminger fra flere fagområder"*

Kvitastein (2002, viii) ser i første rekke et behov for et klarere skille basert på evalueringenes formål, dersom de skal være tjenelige i policysammenheng. Han etterlyser et skarpere skille mellom evalueringer som har ambisjon om *støtte underveis i prosjekter* i motsetning til *å dokumentere effekter av gjennomførte tiltak:*

> Sammenblandingen av disse to ulike formene for evalueringer har konsekvenser som kan gi legitimitetstap både for de forskere som gjennomfører evalueringer og de institusjoner som står som oppdragsgivere for evalueringer. Skaden oppstår ved at servile forskere etter mildt press rapporterer effekter av tiltak i underveisevalueringer uten at disse effektene er sannsynlig dokumenterbare.

Scrivens[19] to hovedkategorier; formative (prosessorienterte) underveisevalueringer versus summative (effektorienterte) evalueringer blandes dessverre relativt ofte og ukritisk sammen når mandatet for en evaluering utformes. Nettopp formålet for evalueringen bør være tydelig fordi de to hovedtypene vil kreve vesensforskjellige metodetilnærminger, og ikke minst fordi det ikke er mulig å gjennomføre en valid evaluering av tiltakets eventuelle effekter før dette er avsluttet og det har passert tilstrekkelig tid. Selv om det kan finnes grensetilfeller mellom prosess- og effektevalueringer, er det viktig å ha en klar forståelse av skillet slik at verken oppdragstaker, oppdraggiver og andre brukere av evalueringen

[19] Scriven, M (1991): Evaluation thesaurus (4. th ed.). Sage Publications. Newbury Park, CA.

ikke trekker konklusjoner om for eksempel effekter av et tiltak der dette ikke er evaluert, eller for å sitere Kvitastein (2002,11): ”Skillet bør likevel hevdes for å sikre at rekkevidden av konklusjoner står i forhold til det et evalueringsarbeid som faktisk er gjennomført”.

I motsatt fall kan evalueringen fort få en metodedesign som er uegnet for slutninger om effekter og derfor lett bli offer for det andre problemet vi presenterte over, brukere trekker konklusjoner om effekter som evalueringen ikke gir grunnlag for. Hvis dette skjer, vil det gjerne true evalueringens så vel som det aktuelle evalueringsmiljøets legitimitet.

Hvordan kan oppdragsgiver utforme et mandat som tilrettelegger for en kontrafaktisk analyse?

Hva gjør så en oppdragsgiver som vil ivareta evalueringens legitimitet og unngå at evalueringer av effekter og virkninger *ikke* trekkes på et uetterrettelig grunnlag? Vi gir her noen innspill i listen under på basis av erfaringene i litteraturgjennomgangen:

- Kontrafaktisk analyse er best egnet for å vurdere *effekter av avsluttede langsiktige satsinger* basert på tidsseriedata og robuste effektevalueringsdesign med flere målepunkter også etter avsluttet tiltak/program.
- *God planlegging av evalueringsoppdrag* er viktig for en hver evaluering, og ikke minst for evalueringer av et tiltaks effekter og langsiktige virkninger. Det er viktig å sikre at arbeidet med evalueringen er godt forberedt – allerede fra oppstart av tiltaket – og at mandatet presiserer relevante og gjennomførbare oppgaver.
- *Tidspunktet for når et tiltak kan være modent for en effektevaluering* med kontrafaktisk tilsnitt er ikke helt tilfeldig. I flere tilfeller der policyaktører etterspør effektene av tiltak, vil det ikke være mulig å gi dette. Før oppdragsgiver utarbeider mandatet bør dette avklares, fortrinnsvis gjennom en ”*evaluerbarhetsvurdering*”. Er det pågående prosesser eller tiltak som skal studeres, bør mandatet unngå å etterlyse effekter. Da kan det kanskje i stedet være behov for en kvalitativt orientert prosessevaluering.
- Mandatet må være realistisk i forhold til *oppdragets omfang og tilgjengelige ressurser*. Å vektlegge laveste pris i utlysninger av effektevalueringer er risikabelt for oppdragsgiver og kvaliteten på arbeidet siden det er få relevante tilbydere av slike tjenester nasjonalt samtidig som det er klart behov for kompetanseutvikling på feltet.
- Mandatet for effektevalueringer kan med fordel eksplisitt etterlyse at det skal gjennomføres en *summativ evaluering basert på kontrafaktisk analyse* med en veltilpasset metodedesign som gir grunnlag for å trekke slutninger om tiltakets effekter.
- Praktiske erfaringer for tilrettelegging og planlegging forut for utlysningen av oppdrag kan trolig også høstes fra evalueringsfora nasjonalt som internasjonalt. Innenfor bistands- og utviklingsfeltet synes flere organisasjoner å vektlegge erfaringsoverføring; OECD/DAC og Verdensbanken er eksempler på dette, men også hjelpeorganisasjoner som CARE er opptatt av dette på bistandsfeltet. Innenfor utdanningsfeltet i USA er det videre et vesentlig større evalueringsfaglig miljø der man også har en egen faglig standard for evalueringsarbeid.

Planlegging av evalueringsarbeid er og blir sentralt. Innledningsvis i denne rapporten siterte vi Rynning (1990, s. 67) som også anbefaler bedre planlegging av framtidige evalueringsstudier:

> ”Når det gjelder å fremme videre forskning om effekter av distriktspolitiske virkemidler, har bedre planlagte, bedre gjennomførte og mer forsvarlige analyserte og rapporterte evalueringsstudier i sin alminnelighet allerede blitt nevnt. Flere grundig planlagte studier vil bidra til på øke almennkunnskapen om effekter av distriktspolitikken, og skape et bedre grunnlag for å sammenligne resultatene studiene i mellom.

Mandatet for en effektevaluering basert på kontrafaktisk analyse, setter høye krav til metodisk kompetanse. Få norske miljøer besitter denne kompetanse alene i dag noe som forutsetter kompetanseoppbygging internt og i samarbeid med ekstern ekspertise.

Hvilke forbedringsmuligheter finnes for evaluator?

Det tredje problemet ved evalueringsoppdrag som ble nevnt innledningsvis og som Kvitastein (2002) drøfter; ”den metodikk som evalueringer gjerne krever, er for lite kjent i basismiljøene”, er absolutt relevant i forhold til effektevaluering med kontrafaktiske metodedesign. Refleksjon og faglig diskusjon for eksempel omkring etablering av utfallslinje, kontrafaktisk tilstand, tiltaksgruppe, sammenligningsgruppe, relevansen av evalueringsdesignen og validitetsspørsmålene er generelt viktig for å kunne styrke evalueringspraksisen og metodeutviklingen på feltet.

Det vil være en utfordring for eksisterende evalueringsfaglige miljøer i Norge å kunne utnytte mulighetene som metodeutviklingen internasjonalt innenfor kontrafaktisk analyse har vært gjenstand for de siste 10-15 årene. Kompetansen kan best utvikles i kontakt med ekspertmiljøer og i evalueringsfaglige fora der utfordringene drøftes. Kontakt med fagfeller internasjonalt på evalueringsfaglige konferanser så vel som gjennom opprettelsen av evalueringsfaglige forum, kan her være viktig. Det er liten tvil om at det er behov for denne typen fora. Olsen (1995:96) påpeker nettopp at fragmentering (i faglig tyngde og formidling) av evalueringskompetansen i Norge og ulike miljøer som gjennomfører evalueringer her, er et problem som hindrer kvalitetsforbedringer i evalueringspraksisen.

Evaluator har imidlertid ikke bare en utfordring i forhold til å utvikle og opprettholde metodisk kompetanse i effektevaluering. Det er samtidig en utfordring *å formidle denne metodiske kunnskapen* til oppdragsgivere og andre brukere slik at evalueringsresultater og konklusjoner kan brukes på en god måte av forvaltningsmiljøene i politikkutformingen.

Klare konklusjoner om effekter og virkninger av tiltak virker lite troverdige uten at validitetstruslene er eksplisitt behandlet. Denne diskusjonen vil primært være av interesse for bestilleren og miljøene som utfører evalueringene fordi den forutsetter forståelse av metodiske utfordringer som ikke kan gjengis fullstendig i en hver rapport. Mange brukere for eksempel i politiske og redaksjonelle miljøer, vil sjelden ha interesse for metodiske begrensninger og problemer, samtidig som de krever klare svar på evalueringsspørsmålene.

Dette vil det av og til ikke være mulig å gi, noe som da bør understrekes i sammendragskapitlet og utdypes i rapportens metodekapittel for å kunne fremme etterprøvbarheten og troverdigheten ved evalueringsarbeidet. Scriven (2007) anbefaler eksplisitt at sammendraget i evalueringsrapporten bør klargjøre styrken for konklusjonene som trekkes og vekten av det empiriske materialet for hvor robuste slutninger som kan trekkes. Da kan evaluator ivareta legitimiteten for evalueringen når det evalueringsmetodiske og -empiriske grunnlaget er på plass.

Referanser

Evalueringsfaglige referanser

AEA (1994): *AEA Guiding Principles for Evaluators.* The American Evaluation Association. (Revisions ratified by the AEA membership, July 2004) http://www.eval.org/

Bamberger, M, Rugh, J. and L. Mabry (2006): *RealWorld Evaluation. Working Under Budget, Time, Data and Political Constraints.* Sage, Thousand Oaks, CA.

Beywl, WT: "The Role of Evaluation in Democracy: Can it be Strengthened by Evaluation Standards? A European Perspective" in *Journal of MultiDisciplinary Evaluation*, 6, 10-29. Kalamazoo, Michigan [http://evaluation.wmich.edu/jmde/]

Campbell, D.T. & Stanley, J.C. (1966): *Experimental and Quasi-experimental Designs for Research*. Chicago.

Chianza, T. (2008): "The OECD/DAC Criteria for International Development Evaluations: An assessment and Ideas for Improvement". *Journal of Multidisciplinary Evaluation*, 8, 41-51. Kalamazoo, Michigan [http://evaluation.wmich.edu/jmde/]

European Commission (2004): *Evaluating EU Activities — A practical guide for the Commission services.* DG Budget. Office for Official Publications of the European Communities. Luxembourg.

Forss, K. and S. Bandstein (2008a): *Evidence-based Evaluation of Development Cooperation: Possible? Feasible? Desirable?* NOINE Working Papers no. 8. Network of Networks of Impact Evaluation

Forss, K., Vedung, E., Kruse, S.E., Mwaiselage, A. and A. Nilsdotter (2008b): *Are Sida Evaluations Good Enough? An Assessment of 34 Evaluation Reports*. Sida Studies in Evaluation 2008:1. Swedish International Development Cooperation Agency. Stockholm

Heckman, James J., Ichimura, Hidehiko and Petra Todd (1998): "Matching As An Econometric Evaluation Estimator. Evidence from Evaluating a Job Training Program" in *Review of Economic Studies* 64, 605-654.

Jakobsen, Stig-Erik og Olav A. Kvitastein (2002): *Måleindikatorer og målemetoder for omstillingsarbeidet*. SNF-rapport nr. 36/2002. Stiftelsen for samfunns- og næringslivsforskning. Bergen

Jerve , Alf. M and Villanger, Espen: (2008): *The Challenge of assessing Aid Impact. A Review of Norwegian Evaluation Practice*. Chr. Michelsen Institute, Bergen, published by Norad. Study 1/2008. Norad. Oslo

Johnson, A et al. (2008): *Effektanalys av "offentlig såddfinansiering" 1994 til 2004. NUTEKs och VINNOVAs såddfinansieringsstöd*. Arbetsrapport 2008:80 SISTER: Swedish Institute for Studies in Education and Research

King, G and Zeng, L.(2007): ”When Can History Be Our Guide?” in *International Studies Quarterly*, 51, 183-210. Blackwell Publishing, Oxford

King, G and Zeng, L.(2006): ”The Dangers of Extreme Counterfactuals” in *Political Analysis* 14:131-159. Oxford University Press

Kvitastein, Olav A., Lines, Rune, Hammervoll, Trond, Tobiassen, Anita og Torstein Nesheim (2000): *Evaluering av omstillingsprogrammet for verkstedsindustrien i Sør-Troms og nordre Nordland.* SNF-rapport 46/00. Stiftelsen for samfunns- og næringslivsforskning. Bergen.

Kvitastein, Olav A. (2002): *Offentlige evalueringer som styringsinstrumenter: Kravspesifikasjoner og kontrollproblemer*. SNF-Rapport nr. 30/2002. Stiftelsen for samfunns- og næringslivsforskning. Bergen

Mohr, L.B. (1992): *Impact analysis for program evaluation. Sage Publications Inc, CA, USA*

Mohr, L.B. (1999): “The Qualitative Method of Impact analysis” in American Journal of Evaluation , 20, 69-83

OECD (2008): *Evaluating Development Co-operation. Summary of Key Norms and Standards*. OECD DAC Network on Development Evaluation. website: http://www.oecd.org/dataoecd/12/56/41612905.pdf

OECD (2001): *Evaluation Feedback for Effective Learning and Accountability*. Development Assistance Committee. Report No. 5 from DAC Working Party on Aid Evaluation. Paris

OECD (1999): *Improving Evaluation Practices: Best Practice Guidelines for Evaluation* Background Paper. Public Management Service, Public Management Committee

OECD (1998): *Review of the DAC Principles for Evaluation of Development Assistance*. Paris

Olsen, Odd Einar (1995): ”Den norske evalueringskompetansen om næringspolitikk og tiltak: Status og utviklingstrekk” i *Evaluering av offentlige tiltak for næringsutvikling. Rapport fra et arbeidsseminar 3.-4. mai 1995*. Norges forskningsråd, Oslo.

Pearl, J. (2000): *Causuality: models, reasoning and inference.* Cambridge University Press, Cambridge

Ragin, C.C. (2007): ‘Comparative Method’ in *The Sage Handbook of Social Science Methodology*. Sage. Los Angeles, CA.

Rye, M. (2002): ‘Evaluating Impact of Public Support on Commercial Research and Development Projects’ in *Evaluation* Vol. 8 (2): 227-248. Sage Publications. London

Rynning, Marjo (1990): *Norske empiriske evalueringsstudier om effekter av distrikts-regionalpolitiske virkemidler: foretaks-/bedriftsrettede virkemidler*. Rapport 131. Næringsøkonomisk institutt. Bergen

Scriven, M. (2007). *Key Evaluation Checklist*. The Evaluation Center, Western Michigan University, Evaluation Checklists Web site: http://www.wmich.edu/evalctr/checklists/

Shadish, W.R, Cook, T.D: (2002): *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton Mifflin Company, Boston

Statskonsult (2003): *Departementenes bruk av evalueringer*. Notat 2003:1. Oslo

Statskonsult (1997): *Evalueringspraksis i departementa : rapport frå Statskonsult sin konferanse om evaluering i april 1997* Rapport 1997:13. Statskonsult. Oslo

Technopolis et al. (2001): *An international review of methods to measure relative effectiveness of technology policy instruments. Final report.* Technopolis. Brighton

United Nations Evaluation Group (2005a): *Standards for Evaluation in the UN System*. http://www.uneval.org/

United Nations Evaluation Group (2005b): *Norms for Evaluation in the UN System*. http://www.uneval.org/

Vedung, E. (1997, 2000): *Public Policy and Program Evaluation.* Transaction Publishers, New Brunswick, New Jersey

Vinnova (2007): *VINNOVAs fokus på effekter – En samlad ansats för effektlogikprövning, uppfölljning, utvärdering och effektanalys*. VINNOVA Analys VA 2007:14. Verket för Innovationssystem

I Næringsspolitiske evalueringsstudier

a) Evaluering av næringspolitiske virkemidler generelt

Alsos et al (2000): *SND i Distrikts-Norge. Evaluering av de bedriftspolitiske distrikspolitiske virkemidlene*. NF_rapport nr. 21/2000. Nordlandsforskning. Bodø.

Hatling et al. (2000): *SND og distriktsutvikling – rolle, virkemidler og resultater. Delrapport 2 i evalueringen av Statens nærings- og distriktsutviklingssfond gjennomført av Technopolis Group, STEP-gruppen og Albatross Consulting*. STEP-rapport R-05. 2000, STEP-group. Oslo.

Evaluering av SND 1993-1999. STEP
Arnold, Erik m.fl. Samlerapport. Albatross og Technopolis. September 2000.

SND in International Context. Delrapport 5 i evalueringen av SND
Arnold, Erik, Jari Kuusisto og Soraya Fahmy. Technopolis. September 2000.

Organisation and Structure. Delrapport 4 i evalueringen av SND
Arnold, Erik og Philip Sowden. SND. Technopolis. September 2000.

Målhierarki og styringssystem. Delrapport 3 i evalueringen av SND
Thoresen, Tor-Jørgen. Albatross. September 2000.

SND og bedriftsutvikling – rolle, virkemidler og effekter. Delrapport 1 i evalueringen av SND. Hauknes, Johan, Marianne Broch og Keith Smith. STEP. R-04. 2000.

Når politikken blir rådgivende - en evaluering av avtalen mellom Troms fylkeskommune og SND om SNDs distriktskontor i Troms
Buanes, Arild og Magnar Andersen, NORUT Samfunnsforsking, rapport nr. 2/99. På oppdrag for Troms fylkeskommune og SND.

Evaluering av SND Nord-Trøndelag
Schiefloe, Per Morten (red.) (1997) Nord-Trøndelagsforskning NTF-rapport 1997:6. På oppdrag fra Nord-Trøndlag fylkeskommune og SND.

Fire gode år med SND? Evaluering og perspektiver. Stiftelsen for samfunns- og næringslivsforskning
Reve, Torger og Klaus Walderhaug (red.) (1997): SNF-rapport 8/97. På oppdrag fra Næringsdepartementet.

b) Evaluering av spesifikke næringspolitiske virkemidler

Alsos et al. (2006): *Flere og bedre bedriftsetableringer? Evaluering av Innovasjon Norges stipendordninger fra 1999-2005*. NF-rapport nr. 11/2006. Bodø.

Clausen Høyvarde og Rasmussen (2008): *Resultatevaluering av SIVAs industri-inkubatorprogram*. NF-rapport 2008. Bodø.

Næringsutvikling i et regionalt perspektiv
Dale, Kristin, Frode Kristiansen og Kjelrun Espedal (1993): SNF. Rapport nr 10/93.

Virkninger for sysselsetting og bosetting
Grimsrud, Gro Marit, Steinar Johansen og Jan Mønnesland (1993): NIBR. Rapport 1993:6.

DUs betydning for bedriftene
Hervik, Arild, Reidar Johansen og Dag Magne Berge (1993): Møreforsking-Molde. Rapport nr 9303.

Lokalt ansvar i distriktspolitikken. Samlerapport fra en evaluering av lokalt totalansvar for bedriftsrettet distriktsstøtte
Bjørnsen, Hild-Marte, Åge Mariussen, Jan Mønnesland, Asbjørn Røiseland og Merethe Sollund. NF-rapport nr. 11/97. På oppdrag for Kommunal- og arbeidsdepartementet.

Lykkelig som liten? Evaluering av Jæren Produktutviklings modell for nyskapingsarbeid i lokalmiljø Agderforsking. FoU-rapport nr. 4/96.

Evaluering av SNDs program for regionalsatsing. Et sammendrag
Arnesen, Tor og Paul Pedersen (1995): Østlandsforskning og NORUT Samfunnsforskning AS. ØF- rapport nr 01/1995.

"....og alle var enige om at det hadde vært en fin prosess"
Erfaringer med strategisk næringsplanlegging. Ringholm, Toril (1994): NORUT Samfunnsforskning rapport 19/94.

Regionalt samarbeid. Rapport fra et forsøksprogram
Arnesen, Tor, Paul Pedersen og Terje Skjeggedal (1994): Østlandsforskning og NORUT. ØF-rapport 07/94.

Regionforsøket og resultatene
Arnesen, Tor, Paul Pedersen og Terje Skjeggedal (1994): Østlandsforskning og NORUT. ØF-rapport 27/94.

Evaluering av DUs tiltak med bedriftslederbefaringer. Møteplass eller rasteplass?
Rogalandsforskning. RF 30/93. Kvadsheim, Lie og Stuland Larsen (1993)

En vurdering av DUs regionsatsingsprogram i Valdres
Skjeggedal, Terje (1993): Østlandsforskning. Notat 17/93.

En vurdering av DUs regionsatsingsprogram i Midt-Østerdal
Skjeggedal, Terje (1993): Østlandsforskning. Notat 17/93.

DUs regionsatsing i Vesterålen
Pedersen, Paul og Magnar Andersen (1993): NORUT Samfunnsforskning.

DUs regionsatsing i Ryfylke
Pedersen, Paul og Magnar Andersen (1993): NORUT Samfunnsforskning.

DUs regionsatsing i Lofoten
Pedersen, Paul og Magnar Andersen (1993): NORUT Samfunnsforskning.

DUs regionsatsing i Indre Sør-Troms
Pedersen, Paul og Magnar Andersen (1993): NORUT Samfunnsforskning.

En vurdering av DUs regionsatsingsprogram i Nordfjord
Arnesen, Tor (1993): Østlandsforskning. Notat 21/93.

En vurdering av DUs regionsatsingsprogram i Fosen
Arnesen, Tor (1993): Østlandsforskning. Notat/93.

En vurdering av DUs regionprogram med vekt på bakgrunn og prosesser i DU sentralt
Arnesen, Tor (1993): Østlandsforskning. ØF-notat 1/93.

c) Entrepenørskap

Evaluering av KRDs etablererstipend. Del 2: Utviklingstrekk og resultater 1989-98
Bolkesjø, Torjus og Jens Aarsand Sæter. Østlandsforskning. Rapport 179.2000.

Evaluering av KRDs etablererstipend. Del 1: Vurdering av forvaltningsmodeller
Møller, Geir og Otto Kaltenborn. Østlandsforskning. Rapport 178.2000.

Etablererstipendet i Oslo - en analyse av årgangene 1995-97
Bolkesjø, Torjus, Telemarksforsking – Bø. Arbeidsrapport nr. 3/99. På oppdrag for Oslo kommune.

Evaluering av Etablererstipend 1989-90; Sysselsetting og lønnsomhet
Bolkesjø, Torjus og Eivind Jørgensen, Telemarksforsking rapport 92/95.

Etablererstipend 1989-90; Etablering av levedyktige virksomheter?
Reiersen, Jon og Oddbjørn Raaum, SNF-rapport 14/95.

d) Forsknings- og utviklingskontrakter

Til beste for de beste. En evaluering av offentlige og industrielle forsknings- og utviklingskontrakter
Bugge, Markus, Trine Monsen og Morten Staude. STEP. R-03. 2000.

Evaluering av prosjektet Leverandørutvikling i kommunesektoren – LUK
Åsland, Dag Yngvar, Knut Senneseth og Maj-Britt Haver, Agderforskning rapport nr. 52/99. På oppdrag for Nærings- og handelsdepartementet

Evaluering av Industrielle Forsknings- og Utviklingskontrakter. Sammenfattende rapport
Gjertsen, Eirik B, Steinar Fossen og Sigmund J. Waagø (1996): NTNU-ORAL.

Evaluering av Industrielle Forsknings- og Utviklingskontrakter. Trinn 3.
Gjertsen, Eirik B, Steinar Fossen og Sigmund J. Waagø (1996): NTH-ORAL. Intern rapport.

Evaluering av Industrielle Forsknings- og Utviklingskontrakter. Trinn 2
Gjertsen, Eirik B, Steinar Fossen og Sigmund J. Waagø (1995): NTH-ORAL. Intern rapport.

Evaluering av Industrielle Forsknings- og Utviklingskontrakter - Trinn 1
Fossen, Steinar, Per Gaute Pettersen og Sigmund J. Waagø (1994): NTH-ORAL. Intern rapport.

Evaluering av offentlige Forsknings- og Utviklingskontrakter
Waagø, Sigmund m.fl. (1993): NTH-ORAL. R-42.

Evaluering av statlige Forsknings- og Utviklingskontrakter for perioden 1980-1990
Waagø, Sigmund, Petter Gjørvad og Per J. Nesse (1991): SINTEF.

e) Nasjonale innovasjons- og kompetanseprogram

Industrifondets nettverksprogram. Rapport fra midtveisevalueringens første del
Havn, Vidar og Trond Buland (1993): SINTEF-IFIM.

Evaluering av SMB-U-programmet.
Hervik, Arild (1990): Møreforsking.

Evaluering av Eksportsjef-til-leie ordningen.
Moen, Øystein og Sigmund J. Waagø (1991): NTH-ORAL.

Vilje til vekst. Evaluering av bedriftsutviklingsprogrammet "SmåbedriftsBUNT".
Elvemo, Johan og Monica Rolfsen (1993): NTH/SINTEF-IFIM.

Teknologibasert nyskaping: Evaluering av SNDs program "Etablering med ny teknologi".
Remøe, Svend Otto (1993): Østlandsforskning. Rapport nr 10/93.

Evaluering av VEI-prosjektet.
Moen, Øystein (1994): NTH-ORAL.

Evaluering av Eksportsjef til leie-ordningen.
Moen, Øystein (1994): NTH-ORAL.

Evaluering av Nettverksprogrammet. Stiftelsen for samfunns- og næringslivsforskning.
Nesheim, Torstein (1994): SNF-rapport 59/94 og 67/94 (kortversjon)

Evaluering av FRAM-programmet. Delrapport A: Analyse av programmets målsetninger
Rolfsen, Monica (1994): SINTEF-IFIM. Rapport STF82 A94008.

Evaluering av FRAM-programmet - Delrapport B: Analyse av programmets gjennomføring. Rolfsen, Monica (1994): SINTEF-IFIM. Rapport STF82 A94017.

Evaluering av FRAM-programmet. Delrapport C: Analyse av programmets resultater.
Rolfsen, Monica (1995): SINTEF IFIM. STF82 A95011.
Evaluering av FORNY-programmet
Møreforsking, Pharos DA, Segal, Quince&Wicksteed (1997): Møreforsking Rapport nr. 9703. På oppdrag fra Norges forskningsråd og SND.

Evaluering av FRAM-programmet i SND. Nesheim, Torstein, Olav Kvitastein, Rune Lines, Kjell Grønhaug og Bjarne Espedal (1997). Stiftelsen for samfunns- og næringslivsforskning, SNF-rapport 84/97. På oppdrag for SND.

Evaluering av SNDs Nettverksprogram. Nafstad, Ola (1998) ECON - Senter for økonomiske analyser AS, Econ-rapport 27/98. På oppdrag for SND.

Evaluering av NT-programmet. Sluttrapport
Pettersen, Per-Gaute. NORUT og Ernst&Young. Februar 2000

Evalueringsrapport for SND TAKE-OFF 2000
Gjelland, Are. GREI. NTNU. Desember 2000

Evalueringsrapport for SND TAKE-OFF 2001
Gjelland, Are. GREI. NTNU. Juni 2001.

Evaluering av myndighetenes deltaking i regionale såkorn- og venturefond, Oslo 1. oktober 1998. Ernst & Young, På oppdrag for Nærings- og handelsdepartementet.

Følgeevaluering av ARENA
Ekbacka Consult. Furre, Harald og Eriksson, Bjørn. Ulike arbeidsrapporter, Jan., aug., og desember 2004.

Evaluering av næringsrettet design
AsplanViak. 2004.

f) Forskningsbasert nyskapning

Borlaug Brorstad et al. (2008): *Evaluering av bruken av uinfrastrukturmidlene i FORNY-programmet*. Rapport 34/2008. NIFU STEP. Oslo.

Bolkesjø et al. (2004): *Evaluering av kommersialiseringsenhetene i Forny-progammet. Hovedrapport*. Rapport 213/ 2004. Telemarksforskning-Bø.

Hervik et al. (2006): *Resultatstyring av brukerstyrt forskning 2005*. Rapport 0616. Møreforskning. Molde.

g) Næringsspesifikke evalueringer

Kvitastein, Olav A. (2000): *Evaluering av omstillingsprogrammet for verkstedindustrien i Sør-Troms og nordre Nordland.* SNF-rapport 46/00. September 2000.

Evaluering av kvalitetssikringsprosjekter i bygge- og anleggsbransjen i Hordaland, Sogn og Fjordane og Møre og Romsdal
Ernst & Young (1994). Intern rapport.

Evaluering av SNDs Program for Miljøprodukter
Kvadsheim, Henrik og Odd Einar Olsen (1995): Rogalandsforskning. Rapport RF-95/076.

Følgeevaluering NUMARIO
2002. I regi av FKD.

SNDs fiskerisatsing. Andre delrapport i evalueringen av SNDs strategi og virkemidler overfor fiskeindustrien
Arbo, Peter m.fl. (1996): NORUT Samfunnsforskning og Fiskeriforskning. SF09/96.

SNDs fiskerisatsing. Første delrapport i evalueringen av SNDs strategi og virkemidler overfor fiskeindustrien
Arbo, Peter m.fl (1996): NORUT Samfunnsforskning og Fiskeriforskning. SF04/96.

Følgeevaluering kompetansenavene i VSP-mat
Torjus Bolkesjø. Telemarksforsking – Bø. 2005.

*Følgeevaluering VSP-rein.*2004.

Følgeevaluering VSP-mat. SNF. Fase 1. 2002.

Evaluering av konkurransestrategien for norsk mat (KOSTRAT). ECON.8/02.

Strakstiltak for næringsmiddelindustrien - En evaluering
PricewaterhouseCoopers og ECON - Senter for økonomiske analyser AS, ECON-rapport nr. 34/98 (Prosjektleder Ivar Pettersen, ECON).På oppdrag for SND.

Evaluering av "konkurransestrategier for norsk mat" – et handlingsprogram for landbruket. Borch, Odd Jarl og Øystein Moen. Nordlandsforskning. NF-rapport nr. 13/97.

Evaluering av omstillingstiltakene for den landbruksbaserte næringsmiddelindustrien. 2. Delrapport. Coopers&Lybrand og ECON (1997): Coopers&Lybrand og ECON.

Evaluering av omstillingstiltakene for den landbruksbaserte næringsmiddelindustrien. 1. Delrapport. Coopers&Lybrand og ECON (1996): Coopers&Lybrand og ECON.

Reiseliv, kultur og design

Evaluering av SNDs og Norsk kulturråds forsøksprogram rettet mot kultur og næring. Sæter, Jens Aa. og Svein Erik Hagen, Østlandsforskning (1999), ØF-Rapport nr. 04/1999. På oppdrag fra SND og Norsk kulturråd.

Foredlingsstrategier i norsk reiseliv. Evaluering av SNDs satsing på videreforedlings-apparatet i reiselivsnæringa. Steen Jacobsen, Jens Kristian; Petter Dybedal og Ole Skalpe (1996): Transportøkonomisk Institutt. TØI-rapport 329/1996.

Evaluering av tiltak for næringsrettet design.
Cæcilie Riis. Sluttrapport Asplan Viak H2004-039, november 2004.

Kundeeffektundersøkelser for Innovasjon Norge/ SND:

Fremdeles mer å hente, etterundersøkelse av bedrifter som mottok støtte fra Innovasjon Norge i 2003. Oxford Research AS. Juli 2007.

Kundeeffektundersøkelse, førundersøkelse 2005-årgangen. Innovasjon a la carte. Oxford. Desember 2006.

Kundeeffektundersøkelse av bedriftsrettede virkemidler fra Innovasjon Norge. Etterundersøkelse i 2006 av bedrifter som fikk finansiert bedriftsutviklingsprosjekt i 2002. Nordlandsforskning. Mai 2006. NF-rapport nr. 9/2006.

Kundeeffektundersøkelse av bedriftsrettede virkemidler fra Innovasjon Norge. Etterundersøkelse i 2005 av bedrifter som fikk finansiert bedriftsutviklingsprosjekter i 2001. Brastad og Madsen. Nordlandsforskning. NF-rapport nr. 6/2005.

Kundeeffektundersøkelse blant bedrifter som mottok tjenester i 2004. Penger er ikke alt! Bedrifters vurdering av Innovasjon Norges virkemidler. Oxford. Oktober 2005.

Kundeeffektundersøkelse. Bedrifters vurdering av Innovasjon Norges virkemidler. Undersøkelse blant bedrifter som mottok tjenester i 2003
Harald Furre. Oxford Research AS i samarbeid med SNF AS. April 2005.

Kundeundersøkelse av SNDs virkemidler. Etterundersøkelse blant næringsdrivende som fikk tilsagn i 2000. Borch, Braastad og Madsen. NF-rapport nr. 13. 2004.

Kundeundersøkelse av SNDs virkemidler. Førundersøkelse blant næringsdrivende som fikk tilsagn i 2002. Borch, Braastad og Madsen. NF-rapport nr. 21. 2003.

Kundeundersøkelse av SNDs virkemidler. Etterundersøkelse blant næringsdrivende som fikk tilsagn i 1999. Borch, Braastad og Madsen. NF-rapport nr. 16. 2003.

Kundeundersøkelse av SNDs virkemidler. Førundersøkelse blant næringsdrivende som fikk tilsagn i 2001. Borch, Braastad og Madsen. NF-rapport nr. 15. 2002.

Kundeundersøkelse – etterundersøkelse 1998. Bræin, Lasse og Arild Hervik. Møreforsking. 2002.

Kundeundersøkelse av SNDs virkemidler. Førundersøkelse blant næringsdrivende som fikk tilsagn i 2000. Borch, Braastad og Madsen. NF-rapport nr. 12. 2001.

Kundeundersøkelsene for SND 1994-2000. Indikator for måling av markedssvikt
Bræin, Lasse og Arild Hervik. Møreforsking. Rapport 0003.

Kundeundersøkelse av SNDs virkemidler. Etterundersøkelse i 2000. Resultater for bedrifter med tilsagn om finansieringsbistand i 1996 og oppsummering for alle etterundersøkelsene 1994-96. Bræin, Lasse, Arild Hervik og Bjørn G. Bergem. Møreforsking. Arbeidsrapport M 008.

Kundeundersøkelse av SNDs virkemidler. Førundersøkelse blant næringsdrivende som fikk tilsagn i 1999. Borch, Odd Jarl. Nordlandsforskning nr. 27.2000.

Kundeundersøkelse av SNDs virkemidler, Hovedundersøkelse – bedrifter med tilsagn 1998B. Bræin, Lasse, Arild Hervik og Bjørn G. Bergem, Møreforsking. Arbeidsrapport M 9918

Kundeundersøkelse av SNDs virkemidler. Etterundersøkelse i 1999 av bedrifter med tilsagn om finansieringsbistand i 1995. Bræin, Lasse, Arild Hervik og Bjørn G. Bergem, Møreforsking Arbeidsrapport M 9908.

Kundeundersøkelsene i SND 1994-9. Oppsummerende rapport fra før- og etterundersøkelser. Hervik, Arild og Lasse Bræin (1998), Møreforskning, Rapport nr. 9803. På oppdrag for SND.

II Bistands- og utviklingspolitiske evalueringsstudier

Borchgrevink, A, Helle-Valle, J og Tassew Woldehana (2003): *Credible Credit: Impact study of the Dedebit Credit and Savings Institution (DECSI), Tigray, Ethiopia.* NUPI Report

Borchgrevink, A, Tassew Woldehana, Gebrehiwot Ageba, Woldeab Teshome (2005): *Marginalized Groups, Credit and Empowerment: A study of Dedebit Savings and Credit Institution (DECSI) of Tigray.* AEMFI

Helge Brunborg, Ian Bowler, Abu Yusuf Choudhury and Mahbuba Nasreen: *Appraisal of the Birth and Death Registration Project in Bangladesh.* Documents 2001/13, Statistics Norway. (External appraisal of the birth and death registration project, Bangladesh 12-23 November 2000 for Norad).

Helge Brunborg and Erik Aurbakken: *Evaluation of Systems for Registration and Identification of Persons in Mozambique*. Documents 97/8, Statistics Norway (Evaluation of the person registration systems in Mozambique, Dec. 1994 (for the Norwegian Refugee Council))

Holden, ST, Deininger, K. and H. Ghebru (2009): 'Impacts of low-cost land certification on investment and productivity'. *Amer. J. Agr. Econ*. 91(2) (May 2009): 359–373

Norad Evaluation Reports
(prior to 2004 reports were issued by The Ministry of Foreign Affairs)

3.92 *De Private Organisasjonene som Kanal for Norsk Bistand, Fase l*

1.93 *Internal Learning from Evaluations and Reviews*

2.93 *Macroeconomic Impacts of Import Support to Tanzania*

3.93 *Garantiordning for Investeringer i og Eksport til Utviklingsland*

4.93 *Capacity-Building in Development Cooperation Towards Integration and Recipient Responsibility*

1.94 *Evaluation of World Food Programme*

2.94 *Evaluation of the Norwegian Junior Expert Programme with UN Organisations*

1.95 *Technical Cooperation in Transition*

2.95 *Evaluering av FN-sambandet i Norge*

3.95 *NGOs as a Channel in Development aid*

3A.95 *Rapport fra Presentasjonsmøte av «Evalueringen av de Frivillige Organisasjoner»*

4.95 *Rural Development and Local Govemment in Tanzania*

5.95 *Integration of Environmental Concerns into Norwegian Bilateral Development Assistance: Policies and Performance*

1.96 *Norad's Support of the Remote Area Development Programme (RADP) in Botswana*

2.96 *Norwegian Development Aid Experiences. A Review of Evaluation Studies 1986–92*

3.96 *The Norwegian People's Aid Mine Clearance Project in Cambodia*

4.96 *Democratic Global Civil Governance Report of the 1995 Benchmark Survey of NGOs*

5.96 *Evaluation of the Yearbook "Human Rights in Developing Countries"*

1.97 *Evaluation of Norwegian Assistance to Prevent and Control HIV/AIDS*

2.97 *«Kultursjokk og Korrektiv» – Evaluering av UD/Norads Studiereiser for Lærere*

3.97 *Evaluation of Decentralisation and Development*

4.97 *Evaluation of Norwegian Assistance to Peace, Reconciliation and Rehabilitation in Mozambique*

5.97 *Aid to Basic Education in Africa – Opportunities and Constraints*

6.97 *Norwegian Church Aid's Humanitarian and Peace-Making Work in Mali*

7.97 *Aid as a Tool for Promotion of Human Rights and Democracy: What can Norway do?*

8.97 *Evaluation of the Nordic Africa Institute, Uppsala*

9.97 *Evaluation of Norwegian Assistance to Worldview International Foundation*

10.97 *Review of Norwegian Assistance to IPS*

11.97 *Evaluation of Norwegian Humanitarian Assistance to the Sudan*

12.97 *Cooperation for Health Development WHO's Support to Programmes at Country Level*

1.98 *"Twinning for Development". Institutional Cooperation between Public Institutions in Norway and the South*

2.98 *Institutional Cooperation between Sokoine and Norwegian Agricultural Universities*

3.98 *Development through Institutions? Institutional Development Promoted by Norwegian Private Companies and Consulting Firms*

4.98 *Development through Institutions? Institutional Development Promoted by Norwegian Non-Governmental Organisations*

5.98 *Development through Institutions? Institutional Development in Norwegian Bilateral Assistance. Synthesis Report*

6.98 *Managing Good Fortune – Macroeconomic Management and the Role of Aid in Botswana*

7.98 *The World Bank and Poverty in Africa*

8.98 *Evaluation of the Norwegian Program for Indigenous Peoples*

9.98 *Evaluering av Informasjons støtten til RORGene*

10.98 *Strategy for Assistance to Children in Norwegian Development Cooperation*

11.98 *Norwegian Assistance to Countries in Conflict*

12.98 *Evaluation of the Development Cooperation between Norway and Nicaragua*

13.98 *UNICEF-komiteen i Norge*

14.98 *Relief Work in Complex Emergencies*

1.99 *WlD/Gender Units and the Experience of Gender Mainstreaming in Multilateral Organisations*

2.99 *International Planned Parenthood Federation – Policy and Effectiveness at Country and Regional Levels*

3.99 *Evaluation of Norwegian Support to Psycho-Social Projects in Bosnia- Herzegovina and the Caucasus*

4.99 *Evaluation of the Tanzania-Norway Development Cooperation1994–1997*

5.99 *Building African Consulting Capacity*

6.99 *Aid and Conditionality*

7.99 *Policies and Strategies for Poverty Reduction in Norwegian Development Aid*

8.99 *Aid Coordination and Aid Effectiveness*

9.99 *Evaluation of the United Nations Capital Development Fund (UNCDF)*

10.99 *Evaluation of AWEPA, The Association of European Parliamentarians for Africa, and AEI, The African European Institute*

1.00 *Review of Norwegian Health-related Development Cooperation1988–1997*

2.00 *Norwegian Support to the Education Sector. Overview of Policies and Trends 1988–1998*

3.00 *The Project "Training for Peace in Southern Africa"*

4.00 *En kartlegging av erfaringer med norsk bistand gjennomfrivillige organisasjoner 1987–1999*

5.00 *Evaluation of the NUFU programme*

6.00 *Making Government Smaller and More Efficient. The Botswana Case*

7.00 *Evaluation of the Norwegian Plan of Action for Nuclear Safety Priorities, Organisation, Implementation*

8.00 *Evaluation of the Norwegian Mixed Credits Programme*

9.00 *"Norwegians? Who needs Norwegians?" Explaining the Oslo Back Channel: Norway's Political Past in the Middle East*

10.00 *Taken for Granted? An Evaluation of Norway's Special Grant for the Environment*

1.01 *Evaluation of the Norwegian Human Rights Fund*

2.01 *Economic Impacts on the Least Developed Countries of the Elimination of Import Tariffs on their Products*

3.01 *Evaluation of the Public Support to the Norwegian NGOs Working in Nicaragua 1994–1999 3A.01 Evaluación del Apoyo Público a las ONGs Noruegas que Trabajan en Nicaragua 1994–1999 4.01 The International Monetary Fund and the World Bank Cooperation on Poverty Reduction*

5.01 *Evaluation of Development Co-operation between Bangladesh and Norway, 1995–2000*

6.01 *Can democratisation prevent conflicts? Lessons from sub-Saharan Africa*

7.01 *Reconciliation Among Young People in the Balkans An Evaluation of the Post Pessimist Network 1.02 Evaluation of the Norwegian Resource Bank for Democracy and Human Rights (NORDEM)*

2.02 *Evaluation of the International Humanitarian Assistance of the Norwegian Red Cross*

3.02 *Evaluation of ACOPAMAn ILO program for "Cooperative and Organizational Support to Grassroots Initiatives" in Western Africa 1978 – 1999*

3A.02 *Évaluation du programme ACOPAMUn programme du BIT sur l'« Appui associatif et coopératif auxInitiatives de Développement à la Base » en Afrique del'Ouest de 1978 à 1999*

4.02 *Legal Aid Against the Odds Evaluation of the Civil Rights Project (CRP) of the Norwegian Refugee Council in former Yugoslavia*

1.03 *Evaluation of the Norwegian Investment Fund for Developing Countries (Norfund)*

2.03 *Evaluation of the Norwegian Education Trust Fund for African the World Bank*

3.03 *Evaluering av Bistandstorgets Evalueringsnettverk*

1.04 *Towards Strategic Framework for Peacebuilding: Getting Their Act Togheter.Overview Report of the Joint Utstein Study of the Peacebuilding.*

2.04 *Norwegian peacebuilding policies: Lessons Learnt and Challenges Ahead*

3.04 *Evaluation of CESAR´s activities in the Middle East Funded by Norway*

4.04 *Evaluering av ordningen med støtte gjennom paraplyorganiasajoner. Eksemplifisert ved støtte til Norsk Misjons Bistandsnemda og Atlasalliansen*

5.04 *Study of the impact of the work of FORUT in Sri Lanka: Building Civil Society*

6.04 *Study of the impact of the work of Save the Children Norway in Ethiopia: Building Civil Society*

1.05 *Study of the impact of the work of FORUT in Sri Lanka and Save the Children Norway in Ethiopia: Building Civil Society*

1.05 *Evaluation of the Norad Fellowship Programme*

2.05 *Women Can Do It – an evaluation of the WCDI programme in the Western Balkans*

3.05 *Gender and Development – a review of evaluation report 1997–2004*

4.05 *Evaluation of the Framework Agreement between the Government of Norway and the United Nations Environment Programme (UNEP)*

5.05 *Evaluation of the "Strategy for Women and Gender Equality in Development Cooperation (1997–2005)"*

1.06 *Inter-Ministerial Cooperation. An Effective Model for Capacity Development?*

2.06 *Evaluation of Fredskorpset*

1.06 *Synthesis Report: Lessons from Evaluations of Women and Gender Equality in Development Cooperation*

1.07 *Evaluation of the Norwegian Petroleum-Related Assistance*

1.07 *Synteserapport: Humanitær innsats ved naturkatastrofer: En syntese av evalueringsfunn*

1.07 *Study: The Norwegian International Effort against Female Genital Mutilation*

2.07 *Evaluation of Norwegian Power-related Assistance*

2.07 *Study Development Cooperation through Norwegian NGOs in South America*

3.07 *Evaluation of the Effects of the using M-621 Cargo Trucks in Humanitarian Transport Operations*

4.07 *Evaluation of Norwegian Development Support to Zambia (1991-2005)*

5.07 *Evaluation of Development Cooperation through Norwegian NGOs in Guatemala*

(Evaluation report 06/2008) Norad, January 2009

1.08 *Evaluation of the Norwegian Emergency Preparedness System (NOREPS)* (Evaluation report 01/2008) Norad, February 2008

2.08 *Joint Evaluation of the Trust Fund for Environmentally and Socially Sustainable Development (TFESSD)* (Evaluation report 02/2008) Norad, April 2008

3.08 *Mid-term Evaluation of the EEA Grants* (Evaluation report 03/2008) Norad, September 2008

4.08 *Evaluation of Norwegian HIV/AIDS Responses* (Evaluation report 04/2008) Norad, October 2008

5.08 *Evaluation of the Norwegian Research and Development Activities in Conflict Prevention and Peace-building* (Evaluation report 05/2008) Norad, November 2008