SNF-RAPPORT NR. 30/2002

OFFENTLIGE EVALUERINGER SOM STYRINGSINSTRUMENTER:
KRAVSPESIFIKASJONER OG KONTROLLPROBLEMER

AV

OLAV A. KVITASTEIN

SNF- PROSJEKT NR.: 6195 "METODIKK FOR MÅLING AV EFFEKTER AV NÆRINGSTILTAK"
PROSJEKTET ER FINANSIERT AV EVA-FORUM

SAMFUNNS- OG NÆRINGSLIVSFORSKNING AS
BERGEN, JULI 2002

ISBN 82-491-0219-3

ISSN 0803-4036

FORORD

Denne rapporten er ment som enkel orientering for personell i forvaltningen som er eller blir ansvarlige for å gjennomføre evalueringer. Rapporten er en del av et prosjekt i regi av EVA-forum[1]. Dette forumet er opprettet etter initiativ fra ildsjeler i Statens nærings- og distriktsutbyggingsfond (SND), departementene og Norges forskningsråd (NFR), som føler ansvar for at evalueringer blir tatt alvorlig. EVA forumet har siden 1995 hatt årlige konferanser med deltakere fra skandinaviske evalueringsmiljøer.

Rapporten bygger på forarbeider gjennomført i regi av et prosjektet rettet mot metodikk for gjennomføring av evalueringer, med vekt på evalueringer av langtidseffekter. En del av disse arbeidene er av mer teknisk karakter, mens andre gjelder administrative konsekvenser av institusjonalisering av evalueringsfunksjonen. Andre tema som drøftes gjenspeiler diskusjoner som har funnet sted i regi av EVA-forum.

Rapporten har ikke ambisjoner som manual for gjennomføring av evalueringer, slike finnes det flere av. Hensikten er å være veiviser i det villniss av faglige og administrative oppfatninger en gjerne møter i evalueringsarbeidet. Rapporten forsøker å gi en "midt på treet" oppfatning av hvordan evalueringsfeltet ser ut fra *forskernes* side, men varsler også store forskjeller i oppfatninger blant forskere. En må være forberedt på at de fleste forskere en møter har en klar oppfatning av hva som menes med evalueringer og hvordan de skal gjennomføres. For oppdragsgiver er det likevel enkelt å observere at forskerne snakker ut fra vidt forskjellige perspektiver og sverger til forskjellige metoder for gjennomføring av evalueringer.

---

[1] Forum for evaluering av strategier og virkemidler for næringsutvikling (EVA-forum)

Det er en ambisjon for denne rapporten at den skal gi en orientering om hvilke posisjoner og perspektiver en kan forvente å finne blant forskere innen evalueringsfeltet. Presentasjon av ulike skoler og retninger kan være et nyttig redskap for betraktninger om hvilke typer evalueringer en ønsker og hvilke tiltak som kan gjøres for å sikre at en får utført evalueringer av tilfredsstillende kvalitet. Med unntak for EVA-forum er det tatt få initiativ til kvalitetssikring av evalueringsfunksjonen.

Det er også et mål å bidra til refleksjon over bruk av evalueringer, slik at en blir bedre i stand til å gjøre seg opp en mening om hvilket bidrag en evaluering tilfører de oppgaver en har ansvar for.

Rapporten er i stor grad et resultat av diskusjoner i EVA- forum. Professor Arild Hervik og forsker Lasse Bræin ved Høgskolen i Molde samt direktør Erik Arnold ved Technopolis Group, UK har vært gode inspirasjonskilder. Spesialrådgiver Jon Hekland og rådgiver Arne Berge i Norges forskningsråd har vært aktive støttespillere for EVA- forum i mange år og har bidratt generøst til fruktbare diskusjoner. Jeg er særlig takknemlig for mange diskusjoner med avdelingsdirektør Pål Aslak Hungnes i SND's strategiavdeling. Som aktiv pådriver i arbeidet med å sikre kvalitet i evalueringsforskningen gjennom flere år, har han hatt betydelig innflytelse. De synspunkter som fremmes i denne rapporten sammen med de feil og mangler som måtte finnes, er likevel forfatterens ansvar alene.

INNHOLD

SAMMENDRAG

Rapporten drøfter betingelser for at offentlige evalueringer skal fungere som styringsinstrumenter. Drøftingen er avgrenset til å gjelde evalueringer av tidsavgrensede programmer eller enkeltstående tiltak, selv om mange av de forhold som diskuteres også er gyldige for andre typer evalueringer.

Hvorvidt dagens evalueringer er egnede instrumenter for policy- utforming og kontroll av offentlige programmer og tiltak, er betinget av flere forhold enn selve evalueringen. Særlig syv problemer oppfattes som kritiske for at evalueringer skal kunne være tjenlige for politikkutforming. De syv punktene under følger stort sett kapittelinndelingen i rapporten, uten at dette signaliserer noen prioritering.

1. Problemet med at ulike typer evalueringsoppdrag ikke skilles klart nok både i anbudsinnbydelser og senere fortolkning av resultater (kapittel 2).

2. Problemet med at evalueringer blir gitt status som resultat av innsats fra forskere, mens evaluering ikke oppfattes som forskning. (kapittel 3)

3. Problemet med at mange og konkurrerende skoler og retninger innenfor evalueringsforskningen har trekk som begrenser hvilke spørsmål som kan besvares i evalueringsoppdraget. (kapittel 4)

4. Problemer med organiseringen av evalueringsforskningen. (kapittel 5)

5. Problemet med evalueringens forvaltningshistoriske bakgrunn (kapittel 6)

6. Problemet med at den diskursen som foregår i etterkant av evalueringer, når tiltak eller programmet bringes opp på den politiske agendaen, tvinger frem betraktninger om tiltakenes effekt. (kapittel 7)

7. Problemet med at den metodikk som evalueringer gjerne krever, er for lite kjent i basismiljøene ettersom evalueringsforskning krever tilnærminger fra flere fagområder. (kapittel 8)

Til sammen tilsier de syv punktene at det finnes et stort potensiale for forbedringer. For oppdragsgiversiden er det viktig å få klarlagt og systematisert disse problemene, slik at en slipper å komme i forlegenhet om åpenbare svakheter skulle bli avdekket innen eget ansvarsområde. For forskersiden er det viktig å vite hvilke spesielle problemer en møter ved evalueringer.

- AD. 1. PROBLEMET MED AT ULIKE TYPER EVALUERINGSOPPDRAG IKKE SKILLES KLART NOK BÅDE I ANBUDSINNBYDELSER OG SENERE FORTOLKNING AV RESULTATER.

En vesentlig forutsetning for at evalueringer skal kunne som policy-instrumenter er et skarpere skille mellom evalueringer som sikter mot støtte underveis i prosjekter og evalueringer som har ambisjoner om 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.

- AD. 2. PROBLEMET MED AT EVALUERINGER BLIR GITT STATUS SOM RESULTAT AV INNSATS FRA FORSKERE, MENS EVALUERING IKKE OPPFATTES SOM FORSKNING.

Rapporten argumenterer for en sterkere kopling mellom basisdisiplinene og evalueringsforskningen og aksepterer ikke uten videre at det tradisjonelle skillet mellom disiplinforskning og policy-forskning gir aksept for lavere kvalitet innen evalueringsforskningen. Skillet anerkjennes, men kvalitetstapet innen

evalueringsforskningen må tilskrives de institusjonelle prosesser som driver evalueringspraksis, snarere enn at de problemer en møter i evalueringer er så spesielle at forskningen må relegeres til et lavere plan.

- AD. 3. PROBLEMET MED MANGE OG KONKURRERENDE SKOLER OG RETNINGER INNENFOR EVALUERINGSFORSKNINGEN OG UNDERKOMMUNIKASJON AV HVORDAN TREKK VED DE ULIKE SKOLER OG RETNINGER BEGRENSER HVILKE SPØRSMÅL SOM KAN BESVARES I EVALUERINGSOPPDRAGET

De ulike skoler og retninger innen evalueringsforskningen blir beskrevet ut fra hvilke typer evalueringer en kan forvente å få, gitt at den utførende forsker var en dedikert tilhenger av den ene eller den andre skole. Fremstillingen må betraktes som idealtypisk og rent orienterende. I praksis vil de fleste forskere være preget av elementer fra flere skoler og retninger. Internasjonalt finner en forskere som kan betraktes som rene tilhengere av spesielle skoler eller retninger. I det norske evalueringsmiljøet må en forvente at forskere vil tendere mot å gruppere seg rundt de skoler og posisjoner som ligger nærmest egen faglige orientering. I hvilken grad de forventninger som beskrives er treffende, blir derfor et empirisk spørsmål. For oppdragsgivere kan det være nyttig å se hvordan skoler og retninger systematisk avgjør hvilken type rapporter som produseres.

- AD. 4. PROBLEMER MED ORGANISERINGEN AV EVALUERINGSFORSKNINGEN

Organiseringen av evalueringsaktiviteten oppfattes som uavklart og problematisk. Nye oppfatninger om at administrasjon i det offentlige stort sett er lik administrasjon i det private gir økt fokus på kontrollproblemer. Institusjonell teori tilsier at de rutiner og prosedyrer som etableres kan være vanskelig å endre, selv om det er nokså tilfeldig etablert og ikke fungerer helt etter intensjonene. Større beslutningsautonomi på alle nivå i den offentlige administrasjon og styring gjennom kontrollsystemer i stedet for hierarkisk styring, endrer

standard operasjonsprosedyrer og rutiner. For evalueringspraksis innebærer dette betydelige farer for tilstivning i lite tjenlige former. Det er vanskelig å gi anvisninger på den rette organisering, men det åpenbart nødvendig å ha en beredskap mot at praksis festner seg i uhensiktsmessige prosedyrer.

- AD. 5. PROBLEMET MED EVALUERINGERS FORVALTNINGSHISTORISKE BAKGRUNN

Evalueringer oppfattes som en integrert del av *New Public Management* (NPM). Den endring i offentlig administrativ praksis som evalueringer er en integrert del av, har sine historiske røtter. Både idéhistorisk opphav og praksis i andre land viser at vektleggingen av "accountability" i den offentlige praksis har utilsiktede sidevirkninger.

- AD. 6. PROBLEMET MED AT DEN DISKURSEN SOM FOREGÅR I ETTERKANT AV EVALUERINGER, NÅR TILTAK ELLER PROGRAMMET BRINGES OPP PÅ DEN POLITISKE AGENDAEN, TVINGER FREM BETRAKTNINGER OM TILTAKENES EFFEKT

Når evalueringer blir industri får evalueringens retorikk politisk betydning. Det er særlig verdt å merke seg at begrepet *effekt* har gjennomslag som gjør det vanskelig å unngå. Når etterrettelighet er målet, kan det bli viktig å si at det tiltak som er gjennomført har hatt den tilsiktede effekt. På denne måten tvinger evalueringens retorikk evalueringes metodikk mot analysemåter som dokumenterer effekter.

- AD. 7. PROBLEMET MED AT DEN METODIKK SOM EVALUERINGER GJERNE KREVER ER FOR LITE KJENT I BASISMILJØENE ETTERSOM EVALUERINGSFORSKNING KREVER TILNÆRMINGER FRA FLERE FAGOMRÅDER.

Retorikkens tvang innebærer en større vektlegging av metodikk som kan sannsynliggjøre effekter, og det blir viktigere å trekke grensene mellom metoder som kan forsvare slike ambisjoner og metoder som ikke egner seg. Kravene til kompetanse for dette er på langt nær oppfylt i det norske evalueringsmiljøer.

Rapporten indikerer et betydelig behov for oppjustering av komptanse, både på forskersiden og på oppdragsgiversiden dersom en ønsker at offentlige evalueringer skal kunne fungere som styringsinstrumenter.

Evalueringer utføres som regel med intensjoner om nøytralitet og faglig uavhengighet. Det er viktig å nevne at de problemer som omtales ikke innebærer et forsøk på å plassere ansvar. De fleste av de problemer som drøftes har sine egne mekanismer og sin egen logikk og eksisterer uavhengig av de beste intensjoner.

1 INNLEDNING

1.1 Hvorfor evalueringer?

I følge rapporter fra OECD (OECD, 1995) har det fra 1980 tallet funnet sted et globalt paradigmeskift når det gjelder kontroll og organisering av offentlig sektor. Det Weberianske byråkrati har gradvis blitt erstattet med modeller og tenkemåter fra alminnelig næringsvirksomhet. Dette innebærer blant annet at beslutningsmyndighet har blitt desentralisert slik at den enkelte tjenestemann/kvinne har fått større ansvar for konsekvenser på eget virksomhetsområde. Basert på den enkle idéen at administrasjon i offentlig sektor ikke er vesensforskjellig fra annen administrasjon, har denne dreiningen skjedd gradvis. Denne endringen, som kan sees som drevet frem av organisasjoner som Organisasjonen for økonomisk utvikling og samarbeid, OECD[2], Det Internasjonale pengefondet, IMF[3] og Verdensbanken samt av politiske trender i Anglo-Amerikanske land, (Christensen, Lægreid, & Wise, 2002) kalles gjerne for the New Public Management (NPM). Det finnes ingen klar definisjon på hva NPM egentlig er, men det kan betraktes både som et komplekst begrep og en reformpakke som vektlegger økonomiske verdier og effektivitet i forvaltningen. Evalueringer må betraktes som en integrert del av the New Public Management.

Det er stor enighet om at bruk av evalueringer innebærer et ansvar ut over det å utføre ordrer som leveres nedover i en hierarkisk kommandokjede (Wallis & Dollery, 1999). Ledere på avdelingsnivå bærer nå også ansvar for gjennomføring og resultatvurdering for programmer og prosjekter. I USA og EU har

[2] Organisation for Economic Co-operation and Development

[3] International Monetary Fund

evalueringer gått fra å være en sporadisk aktivitet til å bli en permanent offentlig institusjon.

1.2 Hva vil vi med evalueringer?

Denne rapporten handler først og fremst om analyser av offentlige intervensjoner som har spesifikke økonomiske og samfunnsmessige mål som hensikt. Den primære hensikt er å etablere noen retningslinjer for analyser av næringspolitiske tiltak som har en identifiserbar begynnelse og slutt. I den grad analyser av mer permanente institusjoner drøftes, er malen for diskusjonen at også disse blir diskutert ut fra sin uttrykte målsetting. Slike analyser av gjennomføring, måloppnåelse og resultater, kalles gjerne *evalueringer*, et begrep Vedung (2000) har karakterisert som en *semantisk magnet*. Med dette tenker han på den språklige, positive kraft som ligger i begrepet evaluering. Begrepet varsler en grundig, kunnskapsbasert vurdering. Når denne er gjennomført, er det etablert en gjennomtenkt og troverdig konklusjon om det som er vurdert. Som Vedung påpeker, kan nettopp den positive betydning begrepet *evaluering* tillegges, være en kilde til problemer. Begrepets egenskaper som semantisk magnet gjør at mange ulike aktiviteter ordnes inn under begrepet. Dokumenter som tidligere ble kalt utredninger, har de siste årene gjerne fått status som evalueringer, uten at særlig mye av metodikk og tilnærmingsmåter er forandret. For offentlige institusjoner som søker ekstern faglig støtte for sine informasjonsbehov, er det av betydning at det finnes a) en typologi som gjør det klart hvilken type oppgave som søkes løst og b) kravspesifikasjoner som sier klart hvilke krav som stilles til valg av løsningsmåte og c) hvilken status oppdragsgiver kan forsvare at det utførte oppdrag blir gitt.

Vi vil i denne rapporten trekke et skille mellom evaluering *som dokumentasjon av resultater* og *evaluering som implementeringsstøtte.* Dette skillet ligger nært Michael Scriven's (1991) skille mellom *summativ* og *formativ*[4] evaluering. Grovt sett innebærer en summativ evaluering at en rapporterer *om* et tiltak eller program[5] etter at programmet er avsluttet eller har nådd en stabil tilstand eller et stabilt aktivitetsnivå. En formativ evaluering rapporterer *til* et program under utvikling eller gjennomføring.

Selv om dette skillet i praksis ikke alltid er like klart, er det av stor betydning ved tildeling av oppdrag. Sammenblanding av formative evalueringer, eller evalueringer underveis og evalueringer som skal tjene som dokumentasjon av resultater, kan gi skadevirkninger i form av legitimitetstap, noe som i sin tur kan gi budsjettmessige konsekvenser. Det kan oppfattes som lite tillitsvekkende om en positiv underveisrapport benyttes som markedsføring av et tiltak eller program, uten at det gjøres forsøk på å dokumentere at et tiltak/program faktisk har gitt de intenderte resultater. God gjennomføring er ikke synonymt med godt resultat[6]. Fremveksten av evalueringsstudier som fagområde har over lang tid vært preget av spenningen mellom det å benytte evalueringsstudier som støtte for implementering og underveis justeringer av programmer og tiltak, og det å benytte evalueringsstudier som dokumentasjon for resultater av tiltak. Lange diskusjoner av hva evalueringer egentlig er, er lite tjenlige ettersom det

[4] Formative evalueringer kan omfatte flere typer evalueringer basert på forskjellige typer metodikk. Summative evalueringer krever bestemte typer metodikk for å begrunne årsakssammenhenger. Skillet mellom *effektanalyser* og *prosessanalyser* (Mohr, 1992) blir ofte brukt omtrent tilsvarende skillet mellom summative og formative evalueringer.

[5] For enkelhets skyld bruker vi begrepet program både for enkelttiltak og mer varige program når dette ikke endre meningsinnholdet.

[6] Jf. økonomenes skille mellom effisiens og effektivitet.

er nokså klart at dokumentasjon av resultater er en aktivitet som er distinkt forskjellig fra, for eksempel, beslutningsstøtte for gjennomføring av en oppgave.

Det må kreves av en evalueringsrapport at den er *handlingsrelevant* og *etterrettelig* uavhengig om rapporten gjelder dokumentasjon eller beslutningsstøtte. Det er imidlertid klart at kravene til handlingsrelevans må være høyere for en formativ evaluering enn for en summativ evaluering. Omvendt må kravene til etterrettelighet være høyere for summative evalueringer enn for formative evalueringer. Formative evalueringer gjennomføres gjerne under sterkt tidspress, begrunnet i at en evaluering underveis nødvendigvis må følge tiltakets/programmets egen gjennomføringstakt. Summative evalueringer skal dokumentere resultater i ettertid, og er ikke i samme grad avhengig av å følge tiltakenes/programmenes tidsfaser. Handlingsrelevans for formative evalueringer gjelder evne til å rapportere *til* programmet på måter som bidrar positivt. Handlingsrelevans for summative evalueringer gjelder evne til å rapportere *om* programmer/tiltak på måter som dokumenterer resultater på en *etterrettelig* måte.

Rapportens begrep om summative evalueringer er strengere en det vanligvis benyttes ettersom vi reserverer begrepet for analyser som tar sikte på å dokumentere effekter. Når evalueringens retorikk tvinger frem[7] fortolkninger i retning av effekter, kan det være fornuftig å benytte et begrep som gir minst mulig avstand mellom innhold og faktisk bruk.

[7] Jf kapittel 7

1.3 Oppbyggingen av rapporten

Kapittel 2 begrunner hvorfor det innføres et skarpere skille mellom formative og summative evalueringer enn det en gjerne ellers finner i evalueringslitteraturen. Kapitlet gir en oversikt over det mangfold av oppfatninger og definisjoner av evalueringer som finnes. Det tredje kapitlet diskuterer de påståtte skillet mellom evaluering og forskning og prøver å komme ut av den tvetydighet som etableres ved å innføre en kategori som kalles evaluering, men som ikke har status som forskning. Kapittel 4 gir en kortfattet oversikt over de skoler og retninger som har utkrystallisert seg internasjonalt, om enn ikke i samme grad i Norge. Kapitlet skisserer hvilke konsekvenser en kan forvente, gitt en hypotetisk institusjonaliseringsprosess der en av de skoler som diskuteres får en fremtredene rolle i evalueringsarbeidet. Kapittel 5 drøfter organiseringen av evalueringsfunksjonen og viser hvordan rollefordeling og posisjoner i organiseringen av evalueringsarbeidet kan virke inn på utfallet av evalueringer, både når det gjelder evalueringenes kvalitet og evalueringsinstituttets troverdighet. Kapittel 6 gir en kortfattet skisse av evalueringsforskningens idéhistoriske røtter. Kapittel 7 drøfter den diskurssive praksis rundt det forhold som er gjenstand for evaluering og hvilken betydning denne praksisen har for offentlige evalueringers verdi som styringsinstrumenter. Kapittel 8 drøfter evaluerings metodikk og gir en kortfattet oversikt, særlig over hvilke krav som må stilles til analyser som sikter mot å dokumentere resultater. Kapittel 9 viser sammenhenger mellom design av tiltak/programmer og viser hvilke konsekvenser utforming av tiltak har for mulighetene for å dokumentere hva tiltakene/programmene har oppnådd. Kapittel 10 lanserer en del forslag om hvilke krav offentlige myndigheter bør stille til de forskningsmiljøer. Kapittel 11 gir en oppsummering av de implikasjoner rapporten har for veien videre.

2 SUMMATIVE OG FORMATIVE EVALUERINGER

2.1 Evalueringer som dokumentasjon

Et bærende prinsipp i denne rapporten er at evalueringer skal være anvendelige som policy-instrumenter. Med dette mener vi at evalueringer skal tjene som *dokumentasjon* for hvordan programmer eller tiltak iverksatt av offentlige myndigheter har fungert. Slik dokumentasjon er nødvendigvis retrospektiv. Det er først i ettertid en kan observere utfall og ha oversikt over hvordan og i hvilken grad tiltakenes målsettinger har blitt realisert. For overordnete beslutninger om hvorvidt programmer skal fortsette eller avsluttes, om tiltak skal intensiveres eller utfases, er slik dokumentasjon vesentlig. Etterrettelig dokumentasjon etterspørres gjerne av politiske myndigheter når alternativer med ulik politisk valør voteres. Bevilgningsutfall kan derfor være avhengig av offentlige etaters evne til å fremskaffe dokumentasjon av resultater. Erfaringer fra land med sterkere legalistiske tradisjoner enn Norge, viser at kvaliteten på dokumentasjonene gjerne angripes når interessemotsetninger oppstår. Norges integrasjon i EU tilsier at det kan være nødvendig med beredskap mot situasjoner der det stilles spørsmål om dokumentasjonens kvalitet og metodikk.

En oppfatning av evaluering som dokumentasjon innebærer i seg selv en vesentlig avgrensing i forhold til det mangfold av meninger om hva evalueringer *er* eller *bør være*. Det er enkelt å observere at svært mange ulike offentlige aktiviteter har blitt gjenstand for evalueringer. Tidsavgrensede tiltak og permanente institusjoner er evaluert, enkeltstående tiltak og hele programmer har blitt evaluert, både før de er gjennomført og etter avslutning. I mange tilfeller er det vanskelig å se at den metodikk som er benyttet, gjenspeiler karakteren av *hva* som er evaluert og det særegne med valg av tidspunktet for informasjonsinnsamling.

Det kan også virke som om bruk av evalueringer har tiltatt de siste årene og at evalueringer nå fremstår som en integrert del av saksbehandlingen på en rekke områder. Inntrykket er at mange offentlige etater og institusjoner opplever krav om evalueringer som en ny kontrollinstans. Det er likevel vanskelig å se at det finnes retningslinjer eller andre systematiske forsøk på å begrunne *om* en evaluering er nødvendig, eller *når* en evaluering bør gjennomføres.

Dersom en slik situasjonsbeskrivelse er dekkende, har offentlige myndigheter et problem: På den ene side er det et udekket behov for dokumentasjon av resultater som utløser krav om evalueringer. På den annen side er det gjort lite for å sikre at de evalueringer som faktisk utføres kan dekke dette behovet.

2.2 Hva er evaluering? Definisjoner og ulike oppfatninger

Det er varierende oppfatninger av hva som menes med evalueringer. I tillegg til de mange, mer eller mindre presise inndelinger i subgrupper eller typer av evalueringer som summativ og formativ, prosessevaluering og fjerdegenerasjonsevaluering, finnes det en rekke, til dels motstridende definisjoner av selve begrepet evaluering. Micheal Scriven, vitenskapsfilosof og tidligere president i The American Evaluation Association, sier at ”Evaluation is the process of determining the merit, worth, and value of things” (Scriven, 1991:1). Charles Manski sier at ”Program evaluation are efforts to learn from experience in order to improve social decisions” (Manski, 1996). Et mer statsvitenskapelig perspektiv definerer evalueringer som ” the careful retrospective assessment of the merit, worth, and value of administration, output, and outcome of government interventions, which is intended to play a role in the future, practical action situations” (Vedung, 2000:3). Begrepet *programevaluering* som ligger nærmest den avgrensing vi har pålagt oss i denne rapporten er hos økonomen Robert L. Darcy (1981) definert som ”the systematic collection and analysis of information to determine the worth of a purposive organized activity”. Darcy inklu-

derer også en megetsigende fotnote om at "*There are many different views concerning the nature and purpose of evaluation*".

Både Scriven og Darcy er svært generelle i sine definsjoner. Ved å unngå mer spesifikke definisjoner, åpnes det for at mange ulike tilnærminger kan kalles evalueringer. Denne ubestemmeligheten utrykkes nokså klart når det sies at "Evalueringen kan være mer eller mindre forskningslignende og forskningsbaserte. Dette innebærer at de er metodisk og teoretisk forankret, og at forskningens systematikk og grundighet danner et sentralt fundament i analysene. Samtidig blir det også påpekt at evalueringer kan gjennomføres uten å være forskningsbaserte, men likevel basert på systematikk. Begge formene kan være utført på oppdrag fra det offentlige i den hensikt å anvende resultatene i en politisk sammenheng" (Sverdrup, 2002:12). Det er stor enighet, for eksempel innenfor the *American Evaluation Association*, om at det er legitimt å diskutere det mangfold av tilnærminger som preger feltet. Det er også enighet om at de ulike tilnærminger til evaluering gjerne *besvarer ulike spørsmål*. Problemene oppstår gjerne når noen ønsker at et evalueringsperspektiv skal gis rang foran andre synsmåter. Denne type ikke-pragmatiske, totaliserende synspunkter finner en særlig i persondebatter mellom fremstående forskere innen ulike skoler og retninger. For eksempel utrykker Guba og Lincoln (Guba, 1990; Guba & Lincoln, 1989; Lincoln & Guba, 1985) mistro til store deler av den etablerte samfunnsforskning når de sier at "It is our intention to define an emergent but mature approach to evaluation that moves beyond mer science- just getting the facts- to include the myriad human, political, social, cultural and contextual elements that are involved" (Guba & Lincoln, 1989:8). Ambisjonene er prisverdige, men den forskningsstrategi de foreslår innebærer den merkverdighet at datainnsamling og systematisk metodikk blir nesten overflødig dersom forskeren har en "dypere" ambisjon for sine analyser.

Vegringen mot definisjoner som setter strenge grenser for hva skal oppfattes som evaluering, er tydelig. En entydig, avgrensende definisjon for hva som *er* evaluering ville samtidig definere hva som *ikke* er evaluering. Konsekvenser av å avskrive en rekke aktiviteter som ikke-evaluering kan være uoversiktlige. Mangfoldet av de fenomener som blir gjort til gjenstand for evalueringer gjør slike grenseganger problematiske. Kjøreegenskaper, sjødyktighet, reiselivsplaner og forskningsinstitusjoner evalueres, uten at dette tilsier at det er nødvendig å finne det minste felles multiplum som kan gi grunnlag for en entydig definisjon. Det er likevel klart at forskere som driver en virksomhet som de selv aksepterer å beskrive som ubestemmelig, står laglig til for kritikk.

2.3 Evalueringsbegrepet i denne rapporten

Ved formative evalueringer eller prosessevalueringer kan som regel det overordnede spørsmål som danner grunnlaget for hypotesedannelsen formuleres som følger: *Ser det ut som om vi er på rett vei?*

For evaluerer blir oppgaven å arbeide ut fra hypotesen om at *prosjektet er på rett vei*, gitt prosjektets/programmets målsettinger. Teorier og tidligere erfaringer må være grunnlaget for utformingen av prosjekter og programmer og evaluerer har en betydelig oppgave i å etterse at tiltak og programmer faktisk blir utformet på måter som gir de beste forventninger om resultat. En slik oppgave er nesten parallell til den oppgave rederiets kontrollør har i byggeperioden for et skip: Han skal kontrollere at skipet blir bygget i samsvar med de spesifikasjoner kontrakten tilsier. Når skipet er ferdig, er det likevel ikke rederiets kontrollør som får ansvar for å kjenne skipet sjødyktig. Slike oppgaver er tillagt Skipskontrollen og klassifiseringsselskaper.

Parallellen er ikke perfekt ettersom den som jobber med formative evalueringer ikke har skipstegninger å forholde seg til, men må ta til takke med teori og erfaring han/hun gjerne selv er ansvarlig for å skaffe til veie. Parallellen er likevel så nærliggende at tenkemåten kan anvendes på norsk evalueringspraksis. På sammen måte som rederiets kontrollør må ha kunnskaper i samsvar med de oppgaver han/hun har tatt på seg, må evaluerer ha kunnskaper om de sekvenser av tiltak som mest sannsynlig fører til de intenderte mål. Mens rederiets kontrollør står ansvarlig overfor sitt rederi, står evaluerer ved offentlige evalueringer ansvarlig overfor det sivile samfunn, dvs. overfor den åpne meningsdannelse og de demokratiske verdier som gir det konstitusjonelle grunnlaget for statens disposisjoner på fellesskapets vegne. Begge står ansvarlig overfor egen selvrespekt. Dette siste betyr at rederiets kontrollør kan ha problemer med å akseptere minimumsstandarder, dvs. et nybygg som med minst mulig margin passerer skipskontrollen. På samme måte kan evaluerer ha problemer med å akseptere økonomisk verdier som overordnet kriterium. Evaluerer kan i enkelte tilfeller foretrekke at sannsynlighet for målrealisering overordnes kostnadsbetraktinger[8].

Det kan være frustrerende, men også spennende for både evaluerer og rederikontrollør at resultatene av deres innsats ikke lar seg bedømme umiddelbart etter avslutning av arbeidet[9]. For rederiets representant vil det utvilsomt oppfattes som et nederlag om skipet ikke fungerer, særlig hvis det som svikter direkte kan tilbakeføres oppgaver han ikke har utført på en tilfredsstillende måte. For

[8] Dette er et nokså vanlig dilemma. Demsetz (Demsetz, 1989) bruker den amerikanske antitrustlovgivningen som et eksempel på en situasjon er en politisk gitt målsetting støter mot effisiensbetraktninger.

[9] Professor Ole Hallesby mente at det religiøse forfallet i folket skyldes den lange ventetiden mot dommedag. Om destinasjon himmel eller helvete var å betrakte som avgjort i dødsøyeblikket, ville folk bli mer gudfryktige (Hallesby, radiotale 25. januar 1953).

den som evaluerer et offentlig program/tiltak bør ansvarsfølelsen fungere på samme måte, selv om det kan ta lengre tid før det gjennomføres summative undersøkelser om hvorvidt tiltaket har fungert etter intensjonene. Ved formative evalueringer der evaluerers forslag har medført betydelig justeringer eller endringer i et program, vil vurderinger av kvaliteten på disse forslagene bli en integrert del av den summative evalueringen.

I resten av rapporten vil vi benytte evalueringsbegrepet på denne måten: Både formative og summative evalueringer er teoridrevne og forskningsbaserte vurderinger. Formative evalueringer gjelder gjennomføring av prosjekter/tiltak, summative evalueringer gjelder effekter av prosjekter/tiltak. Vi vil i flere anledninger gjerne benytte begreper som underveisevaluering og prosessevaluering. Disse betraktes som typer av formative evalueringer. Begreper som effektanalyse og dokumentasjon av effekter vil også bli benyttet. Disse betraktes som former for summative analyser.

Skillet mellom formative og summative evalueringer er mer hensiktsmessig enn absolutt. Det kan finnes grensetilfeller der det er vanskelig å trekke et entydig skille. Skillet bør likevel hevdes for å sikre at rekkevidden av konklusjoner står i forhold til det evalueringsarbeid som faktisk er gjennomført. Som vi kommer tilbake til senere, vil de to evalueringsformene måtte være undergitt ulike metodiske krav for legitimt å kunne hevde sine ambisjoner.

3 EVALUERINGER OG EVALUERINGSFORSKNING

3.1 Disiplinforskning og policy-forskning

Begrepene ”policy-forskning” og ”disiplin-forskning” er hentet fra sosiologen James S. Coleman, og viser til et skille mellom forskning som eksplisitt skal gi kunnskapsgrunnlag for politikken og tradisjonell akademisk forskning (Knudsen & Wærness, 2001). For oppdragsgiver er dette skillet av interesse, ettersom det ofte etableres et skille mellom *evalueringer* og *evalueringsforskning*. Dette skillet gjelder ikke skillet mellom evaluering og forskning *om* evaluering, for eksempel forskning som gjelder evalueringsmetodikk. Skillet går mellom evalueringer som undergitt *andre krav enn annen samfunnsforskning* og evalueringer som undergitt *samme krav som annen samfunnsforskning*. Dette skillet fremstilles gjerne uklart, men refereres ofte til ved hjelp av idealtypene ”policy-forskning” og ”disiplin-forskning”. I følge Coleman er det fire kjennetegn ved policy-forskningen som gjør den forskjellig fra disiplinforskingen (Knudsen & Wærness, 2001:252):

1. *Tid*. Policy-forskningen må følge tidsplanen til beslutninger i handlingsverdenen i samsvar med politikkens rytme og arbeide ut fra den informasjon som er tilgjengelig.

2. *Språk*. Forskere som er engasjert i policy-forskning må kommunisere med folk som ikke behersker spesialistenes termer og språk.

3. *Konflikt*. Policy-forskningen er preget av motstridende interesser, resultatene kan gripe inn i eksisterende maktforhold og ressursfordeling. Det er vanskelig å unngå at forskerne dras inn i eller påvirkes av konflikter.

4. *Informasjon.* I handlingsverdenen er omfattende forklaringer og tilleggsinformasjon ofte av stor betydning. Displinforskningens krav til teoretisk

> eleganse og kompakte formuleringer gjelder ikke her. I policy-forskningen må en bruke modeller som er enkle i forhold til betingelser en vet kan variere. Policy-forskningen må på en helt annen måte enn disiplinforskningen ta utgangspunkt i verden som den er[10].

Mange evalueringsforskere vil trolig gjenkjenne situasjonsbeskrivelsen i disse fire punktene. Det er en nærliggende konklusjon at evalueringer må betraktes som policy-forskning og følgelig at *andre og lavere krav enn annen samfunnsforskning,* må aksepteres. Dette er problematisk ettersom det gir inntrykk av at forskere legitimt kan senke kravene til kvalitet med henvisning til oppgavens karakter. En bedre fortolkning av Colemans distinksjon er at evalueringer stiller andre krav til *formidling* av forskning enn det som er standarder innenfor disiplinforskningen. En slik fortolkning innebærer blant annet at en aksepterer at evalueringsrapporter må ha en form som er forskjellig fra disiplinforskningen formularer, mens den enkelte forsker er ansvarlig for at metodikk og systematikk er av en kvalitet som gjør rapporten kan oversettes til disiplinforskningens form.

Skillet mellom policy-forskning og disiplinforskning er hensiktsmessig for mange typer evalueringsarbeid, men innenfor vår sammenheng, evalueringer av avgrensede program/tiltak *er det bare tjenlig for formative evalueringer*. Som påpekt av Knudsen og Wærnes (Knudsen & Wærness, 2001:253) kan ikke policy-forskningen unndra seg de faglige krav som gjelder for disiplinforskningen, men de begrensninger som ligger i evalueringers kontekst, slik den er beskrevet i de fire punktene over, begrenser mulighetene for å oppfylle en del krav. Disse begrensningene har likevel ikke samme gyldighet *for de typer evalueringer som pretenderer å dokumentere effekter av program/tiltak*. For

[10] Fritt etter Knudsen og Wærness (Knudsen & Wærness, 2001).

summative evalueringer, og særlig for analyser som foretas med sikte på dokumentasjon av effekter i etterkant av gjennomføring av et tiltak/program er konteksten gjerne en annen. En er i mindre grad avhengig av å følge prosjekters tidsplaner, mindre involvert i umiddelbare problemer og kan ofte holde større avstand til motstridende interesser. For slike analyser må det sikres at konklusjoner kan etterprøves av andre fagfolk, at de strengere krav fra disiplinforskningen opprettholdes.

3.2 Disiplinene som kvalitetsgarantister

Det er flere grunner til å anta at tiltak som bidrar til å fjerne "policy-forskning" fra "disiplinforskning" vil svekke kvaliteten på gjennomførte evalueringer:

1. "Displinforskningen" har en fag- og disiplinmessig inndeling, med tilordnete tidsskrifter som kvalitetsgarantister.

2. Offentlige tjenestemenn/kvinner har som regel displin/fagspesifikk bakgrunn. De har sin utdannelse fra en høyskole eller universitet og er opplært til å se verden fra fagets ståsted.

3. De etablerte disiplinene har en etablert legitim autoritet som evalueringsforskningen mangler.

4. Om disiplinforskningens prosedyrer erstattes med egne oppskrifter for gjennomføring av evalueringer økes faren for prosessrasjonalitet og sermoniell adopsjon av rutiner.[11]

[11] Se pkt. 3.2.4

3.2.1 Disiplinenes spesifikke kompetanse

Inndelingen i fag og disipliner etablerer ulike perspektiver på verden. Økonomer er opptatt av allokering av knappe ressurser, geografer er opptatt av den romlige dimensjonen, sosiologer av den sosiale dimensjonen og psykologer konsentrerer seg om enkeltmenneskets forutsetninger for problemløsning og sosialt liv. Oppdragsgivere tilordner som regel evalueringsoppgaver til relevante fagmiljøer. Sosiologer blir gjerne tildelt evalueringer av offentlige tiltak som gjelder for eksempel familien eller trygd, geografer evaluerer regionale omstillingsprogrammer og psykologer bringes gjerne inn ved evalueringer av tiltak som gjelder psykisk helse. Miljøer med særlig kompetanse på organisasjon og ledelse benyttes gjerne for problemer av administrativ eller institusjonell karakter. Begrunnelsen fra oppdragsgiver er at disse fagmiljøene har særlig kunnskap om det problemområdet som skal evalueres.

Det er liten tvil om at tildeling av oppdrag etter feltspesifikk kompetanse er en god løsning. Evalueringsforskning er likevel en tverrfaglig virksomhet der innlån fra forskjellige disipliner konstituerer den faglige kjernen. Problemet er derfor gjerne heller *en manglende erkjennelse av at basismiljøene må tilføres kompetanse som er spesifikk for evalueringer.* Det finnes så vidt jeg vet ingen høyere utdanningsinstitusjoner i Norge som innen sine hovedfags- eller doktorgradskurs tilbyr kurs som er direkte rettet mot evalueringsforskning.

Evalueringsresultater bør i størst mulig grad publiseres i journaler med refereeordninger, selv om bare et mindre antall artikler kan forventes å nå frem til publisering. Dette er viktig, både for å kunne styrke evalueringsforskningens anerkjennelse, men også for at politiske aktører skal kunne få tillit til resultater. Referee-ordninger er kostnadsfrie og kompetente styringsgrupper som ikke er ”stake-holders”. Det er trolig en betydelig gevinst i det å kunne trekke veksler på beste tilgjengelige kompetanse for vurdering av evalueringer av offentlige

tiltak. Disiplinforskningen har m.a.o. tilgjengelige ressurser for kvalitetssikring, ressurser som er kostnadsfrie og i tillegg de beste som finnes innen området.

I praksis er det likevel svært lite av resultater fra evalueringer som rapporteres. Selv i spesialtidskrift som *New Directions for Program Evaluation* er det lite rapportering av faktiske evalueringsstudier. Andre emner, som teoridiskusjoner, dominerer. I den senere tid har en imidlertid sett en betydelig vekst i rapportering av evalueringsresultater i økonomiske og økonometriske tidsskrifter, noe som henger sammen med at nyere metodikk for statistisk modellering av intervensjoner har fått mye oppmerksomhet i sammenheng med tildeling av Nobelprisen 2000 til professor James J. Heckman.

3.2.2 Offentlige tjenestemenn/kvinners faglige bakgrunn

Det er et problem at forvaltningen har for liten kapasitet for å absorbere forskningsresultater. De fleste tjenestemenn/kvinner har høyere utdannelse som gir dem en faglig identitet og en måte å se verden på. Utgangspunkt i disiplinforskningen kan gjøre kommunikasjon mer effektiv og senke absorberingskostnader. Det kan også føre til at oppdragsgivere blir bedre i stand til å vurdere kvaliteten på rapporter. At forskjellig bakgrunn kan føre til uenighet om hvilken faglig synsvinkel som er best egnet for det enkelte problem, er trolig bare til berikelse for hvordan problemene defineres i evalueringsoppdrag. Det er trolig en betydelig gevinst å hente i intern skolering av hvordan de ulike profesjonsgrupper oppfatter evalueringer, i den grad slik aktivitet har en plass i de respektive fagmiljøer.

3.2.3 Disiplinenes legitimitet – og begrensninger

Ved å gå nærmere disiplinforningen kan en dra nytte av etablert autoritet. Det er ingen tvil om at evalueringsforskningen trenger slik drahjelp. I fjor hadde *American Journal of Evaluation* et eget nummer viet evalueringsforskningens problemer med dårlig omdømme (Donaldson, 2001). Dette er ganske oppsiktsvekkende ettersom de fleste fag og subdisipliner som har et dårlig rykte sjelden skriver så direkte om det. Erkjennelsen av at feltet har dårlig rykte må anses som et positivt utgangspunkt for debatt. Viljen til problemerkjennelse har likevel trolig sammenheng med evalueringsforskningens fragmenterte karakter. De fleste virker innenfor evalueringsforskningen har gjerne sin faglige identitet knyttet til et *annet* fagfelt eller disiplin. Det kan være lettere å erkjenne feltets feilbarlighet når en ikke er *bare* evalueringsforsker. At evalueringsfeltet er fragmentert, kan også være en fordel i møtet med den praksis forskerne skal betjene.

Det er liten tvil om at evalueringsforskningens forhold til disiplinforskningen er av stor betydning både for legitimitet og kvalitet. Det meste av metodisk og teoretisk utvikling foregår innen basisdisiplinene. For å oppnå respekt som forskningsfelt er det viktig at evalueringsforskere har oppdatert kunnskap og beherskelse av det som foregår innen basisdisiplinene. Fagdisiplinene har likevel er del særtrekk som det er verdt å være oppmerksom på. De enkelte disipliner og subdisipliner kan, sett som systemer, beskrives som autopoetiske[12] (Luhmann, 1990). Systemer preget av autopoesis er selvrefererende systemer som konstitueres ved at de avgrenser seg mot omverdenen ved å blokkere ekstern kommunikasjon. Det legale system blir ofte brukt som eksempel på et slikt sosialt system (Luhmann & Jacobsen, 1992). Anvendt på fagdisipliner betyr

[12] Autopoesis referere til selvgroende systemer. Begrepet er hentet fra biologien der det benyttes om organismer som henter næring fra eget vev.

dette at det er disiplinens interne kommunikasjon via fagspråk og artikler i tidsskrifter[13] som er den karrieredrivende praksis. Kommunikasjon mot den praksisverden en møter i evalueringsforskningen har ingen verdi om den ikke omsettes til denne interne form for kommunikasjon Dette trekket ved disiplinene som systemer forklarer også det som gjerne kalles intellektualisme eller intellektuell skjevhet (Bourdieu, 1996), tendensen til å abstrahere praksiser til idéer verdige for betraktninger snarere enn problemer som skal løses.

Om beskrivelsen over er dekkende, har basisdisiplinen en indre logikk som bryter med policyforskningen krav. Betrakter vi Bourdieu og Luhmanns betraktninger som spissformuleringer av systemtrekk som kan finnes i mer eller mindre utpreget grad, er det lettere å erkjenne at dette er trekk som må tas hensyn til når en forankrer evalueringsforskningen tettere mot disiplinforskningen. Det er likevel liten grunn til å anta at en ved en sterkere profesjonalisering av evalueringsforskningen vil unngå å utvikle uheldige systemtrekk.

3.2.4 Prosessrasjonalitet og seremoniell adopsjon av rutiner

Et gjennomgående trekk ved mange retningslinjer for gjennomføring av evalueringer er tendensen til å regne evalueringer som noe generisk; de fleste typer oppgaver kan gjennomføres innen samme skjema for evalueringsprosedyrer. Oppdragsgiver vil ofte finne at skjemaet passer dårlig i akkurat den saken hun er ansvarlig for. Trolig er det nokså umulig å konstruere et analyseskjema som passer for alle evalueringer. Konstruksjon av slike skjema har også den uheldige side at de oppfordrer til *prosessrasjonalitet* (March, 1988) der gjennomføringen av de foreskrevne prosedyrer blir selve målet for evalueringen, resultatene av evalueringen blir underordnet. Et annet trekk ved mange "retningslinjer

[13] Journaler er gode eksempler på selvrefererende praksis. Referanselistene angir posisjoneringen og refererer nesten utelukkende til andre medlemmer i samme "system".

”er tendensen til å blande sammen programevaluering og ytelsesmålinger. Den amerikanske riksrevisjonen[14] (GAO) advarer sterkt mot slik sammenblanding av ”program evaluation” og ”performance measurement”. Dette er aktiviteter som har ulike foki og ulike mål. Ytelsesmålinger tar sikte på å etablere stabile indikatorer for er eller mindre kontinuerlig overvåking av forhold mellom ressursbruk og måloppnåelse. Programevalueringer derimot, sikter mot vurderinger av enkeltstående, tidsavgrensete tiltak. ”Performance measurement” er noe som benyttes ved vurderinger av effektivitet i permanente institusjoner, men programevalueringer benyttes for å vurdere hvor vellykkede eller hensiktsmessige en kan betrakte de program eller tiltak som gjerne blir iverksatt av de samme permanente institusjoner.

Prosessrasjonalitet (March, 1988) utgjør en betydelig fare ved innføring av faste skjema for hvordan evalueringer skal gjennomføres. Prosessrasjonalitet er nært beslektet med begrepet sermoniell adopsjon - ”formal adoption of practice on the part of a recipient unit’s employees for legitimacy reasons, without their believing in its real value for the organization” (Kostova & Roth, 2002:220).

Det kan likevel tenkes at en kan konstruere en mengde skjema som kan være til hjelp for spesifikke situasjoner og oppgaver. Farene for prosessrasjonalitet eller seremoniell adopsjon vil ikke bli mindre av den grunn, men slike skjema kan ha en retningsgivende funksjon, gitt at en makter å utvikle en typologi for evalueringer som gjør det mulig å kategorisere den enkelte oppgave på en hensiktsmessig måte. En kan likevel risikere at det oppstår miljøer som spesialiserer seg på å gjennomføre evalueringsoppgaver i henhold til slike nokså fastlagte skjema. En gir dermed finansielt rom for en stor mengde ”policy-forskning” som arbeider under andre selvpålagte standarder enn tradisjonelle forsknings-

[14] United States General Accounting Office, Report GAO/GGD-98-26

miljøer. Resultatet vil trolig være en senkning av kvaliteten på det arbeidet som blir utført. Det er derfor ikke tilrådelig å betrakte evaluering som "mer eller mindre forskningslignende" (Sverdrup, 2002:12). En slik beskrivelse kan forståes som et forsøk på å trekke veksler på forskningens legitimitet, samtidig som en unndrar seg forskningens krav.

4 ULIKE RETNINGER INNEN EVALUERINGSFORSKNINGEN

4.1 Skoler, disipliner og kjennetegn ved evalueringer

Det kan være vanskelig å orientere seg med hensyn til hvilket teoretisk perspektiv den enkelte forsker forsøker å formidle i sitt tilsvar på et konkret evalueringsanbud. Ved oppdrag som ikke eksplisitt gjelder evalueringer, er de fleste tilfeller enkelt å skille mellom rådende tenkemåter i et geografisk miljø, mellom sosiologer eller forskere med økonomisk bakgrunn. I de fleste tilfeller vil også oppdragene være rettet mot bestemte fagmiljø utfra oppgavens karakter. Ved anbud som gjelder evalueringer henvender en seg gjerne bredt mot ulike miljø, og får i tillegg tilsvar som indikerer ulike subdisipliner eller distinkt ulike tenkemåter *innen* det enkelte fagmiljø. Slike subdisipliner kan ha forskjellig karakter. Det kan være sosiologer som er prinsipielle tilhengere av kvalitativ metodikk, det kan være organisasjonsteoretikere som er spesialister på transaksjonskostnadsteori og det kan være økonomer som foretrekker ikke-empiriske forklaringsmodeller. Det kan være vanskelig å oppfatte de klare kunnskapsbegrensninger som spesialisering i subdisipliner ofte innebærer.

Karakteristika ved disse subdisiplinene kan være av stor betydning for gjennomføringen av evalueringsoppgaven[15]. Krav om autoritet når det gjelder konklusjoner, kan variere fra beskjedne ambisjoner om at rapporten ikke må forståes som stort mer enn en bemerkning, til krav om at konklusjoner *skal* tas opp og må taes hensyn til. Misforhold mellom det metodiske grunnlaget for konklusjoner og krav om autoritet, kan være problematisk og lede oppdragsgiver inn i situasjoner som gir tap av troverdighet. Sammenfall mellom oppdragsgi-

[15] Det kan være vanskelig å ha noen kvalifisert mening om hvor integrert norske evalueringsmiljøer er i de ulike internasjonale evalueringstradisjoner. For enkelte miljøer er det likevel mulig å registrere at i alle fall ledende enkeltpersoner markerer seg som klare tilhengere av distinkte skoler og retninger.

ver ønske om sikre konklusjoner, selv i situasjoner når slike ikke er enkelt tilgjengelige, og autoritativ rapportering av konklusjoner på sviktende premisser, kan få svært uheldige konsekvenser.

Schriven (Scriven, 1993) identifiserer fem ulike perspektiver eller tilnærmingsmåter som alle har en så stor internasjonal tilhengerskare og er så knyttet opp mot bestemte evalueringsmiljøer at de kan betegnes som *skoler.* For å bedre dekke norske forhold, kan Scrivens perspektiver suppleres med økonomenes perspektiver på evalueringer. For vårt formål er dette en hensiktsmessig inndeling ettersom den gir en klar pekepinn på hvilken type evaluering de ulike skoler tenderer mot å produsere.

4.1.1 Skole 1: Beslutningsstøtteskolen

Beslutningsstøtteskolen betrakter evalueringer som en integrert del av en rasjonell beslutningsprosess for programadministrasjon (Stufflebaum, Guba, & Tyler, 1971). Et kjent konsept fra denne skolen er CIPP-modellen (Context, Input, Process and Product) for evalueringer:

1. **C***ontext evaluation* skal støtte planleggingsbeslutningene. Klarlegging av hvilke behov et program skal rettes mot gjør det lettere å definere programmets målsettinger.

2. **I***nput evaluation* tjener til å strukturere beslutninger gjennom identifisering av tilgjengelige ressurser og klargjøring av alternative strategier. Hvilke planer som har det beste potensiale for programmets målsettinger avgjør design av programutforming.

3. **P***rocess evaluation* skal støtte implementeringsbeslutningene. Hvor godt er planen iverksatt? Hvilke barrierer truer suksess? Hvilke revisjoner er nød-

vendige. Når disse spørsmålene er besvart, kan prosedyrer overvåkes, kontrolleres og justeres.

4. **P***roduct evaluation* skal tjene til å resirkulere beslutninger. Hvilke resultater er oppnådd? I hvilken grad er behovene redusert? Hva skal gjøres med programmet når det har gjort sin misjon? Dette er spørsmål som er viktige ved vurderinger av hva en har oppnådd med programmet.

CIPP- modellen er forførende ved at den etterligner trinnene i en rasjonell beslutningsprosess og på mange måter avspeiler tidsepokens (1971) litt naive tro på mulighetene for implementering av rasjonelle prosesser. Modellen gir ingen retningslinjer for hvordan de enkelte faser praktisk kan gjennomføres slik at de enkelte faser blir logisk sammenlenket. Punkt 4 alene har etter 1970 gitt opphav til en svært omfattende litteratur om hvordan resultater kan vurderes. Oppsummert kan en si at CIPP-modellen stiller en mengde gode spørsmål som ikke har enkle svar og i sin tilsynelatende konsistente form nesten oppfordrer til skinnevalueringer. Begrepet *skinnevalueringer* benyttes gjerne om situasjoner der gjennomføring av evalueringen er det sentrale mens resultater av evalueringen ignoreres, uavhengig om resultatene tilsier handling eller ikke.

4.1.2 Skole 2: Den relativistiske tilnærming

Den relativistiske tilnærmingen hevder at evalueringer bør utføres på oppdragsgivers premisser, uten at evaluerer på noen måte trenger å gi sin tilslutning til de verdier disse premisser representerer. Denne holdningen er trolig utbredt mellom mange evaluerere, og innebærer på et vis at ideen om verdifri forskning ligger til grunn for troen på at en kan benytte de nøytrale metodiske instrumenter en har til rådighet for å besvare de spørsmål som er stilt, uavhengig av spørsmålenes karakter. Begrepet *relativistisk* er hentet fra Scrivens terminologi

og signaliserer den holdning han mener er fremtredende innenfor amerikanske forskningsmiljøer. Så vidt jeg vet er dette Scrivens egen mening mer enn en empirisk gyldig påstand. Det er tvilsomt om denne holdningen er representativ for norske forskningsmiljøer. Trolig ville en slik holdning i norsk sammenheng blitt forstått som opportunistisk og ødeleggende for forskningens integritet. Det har hendt at norske forskningsmiljøer har blitt bedt om å utrede bare de *positive* effekter av et tiltak eller program. Slike henvendelser blir ofte, *men ikke alltid*, avvist av forskningsmiljøene.

Etter Scrivens mening er blant annet Rossi og Freemans mye brukte lærebok fra 1989 (Rossi, Freeman, & Lipsey, 1999) eksempel på en slik relativistisk tilnærming. Rossi og Freemans posisjonen innebærer at evalueringer betraktes som anvendt samfunnsforskning og at forskerne som nøytrale ikke trenger å ta stilling til de verdispørsmål som mange evalueringer uunngåelig stiller. Det er trolig ikke uvanlig i norske miljøer å betrakte evalueringer som anvendt samfunnsforskning. Det kan likevel argumenteres for at evalueringer i de fleste tilfeller krever metodiske kunnskaper som ikke inngår i standard hovedfagsundervisning ved de fleste norske høyere læresteder. Om en tror verdispørsmål kan takles på ryggmargsrefleks, uavhengig av formell opplæring, er det likevel tvilsomt om fraværet av systematisk opplæring i evalueringsarbeid kan rettferdiggjøre at evalueringer i Norge betraktes som anvendt samfunnsforskning.

Et sannsynlig resultat av samspillet mellom manglende metodisk skolering og problemer med å takle verdispørsmål, er konklusjonsvegring. Uviljen mot å trekke evaluerende konklusjoner kan for bli et problem for oppdragsgiver. Debatten om *bruk* av evalueringer reiser mange spørsmål om hensikten med evalueringer. Det er likevel ikke hensiktsmessig å knytte spørsmålet om evalueringers anvendelse opp mot problemet med konklusjonsvegring. Dersom konklusjonsvegring har sin rot i manglende opplæring i norske evalueringsmiljøer

og benyttes av evaluerer for å skjule egen usikkerhet, er dette lite tilfredsstillende. Om ingen konklusjon kan trekkes, må dette begrunnes. Mangelfull opplæring må ikke returneres oppdragsgiver som en vegring mot bruk av usikre konklusjoner. For oppdragsgiver er det ikke nødvendigvis noen fordel at objektivitet etterstrebes med den følge at programmets mål tas som gitt mens konklusjoner uteblir. Manglende opplæring i evalueringsmetodikk kan ikke forståes som respekt for basisdisiplinene, men kan tolkes som manglende respekt for oppdraget.

4.1.3 Skole 3: "Rich description" tilnærming

Rich description tilnærming innebærer en journalistisk eller etnografisk tilnærming. Denne tilnærmingen forbindes gjerne med "the North Dakota School" (Kemmis & Stake, 1988; Stake, 1975; Stake, 1986a; Stake, 1986b; Stake, 1995; Stake, Easley, & Anastasiou, 1978) Rich description tilnærmingen er også utbredt blant en del engelske forskere og er preget av at forskerne stiller seg utenfor og observerende til evalueringsproblemet. De pretenderer å rapportere nøytralt det de observerer, uten å trekke konklusjoner. *"The rich description school"* har mange felles trekk med *"the relativistic approach"*. Begge skoler pretenderer nøytralitet og objektivitet, unngår verdispørsmål og vegrer seg eller sørger for å unngå konklusjoner som gjelder evalueringsoppdraget.

Rich description tilnærming kan være interessant som meta-analyse av et program eller tiltak, men må ellers betraktes som skritt bort fra programevaluering, slik termen benyttes her. Slike refleksjoner over et tiltak eller program kan være sentrale for overordnede policy-vurderinger. For spørsmål som gjelder justeringer av fremdrift i eksisterende tiltak eller hvorvidt et tiltak/program gav resultater i samsvar med uttrykte målsettinger, er slike analyser trolig likevel av mindre verdi. For oppdragsgiver kan en slik tilnærming fort fortone seg som

kommunikasjonsbrist. Resultatet av tilnærmingsmåten er gjerne at en opplever å få svar på spørsmål en ikke har stilt, og ingen svar på de spørsmål en har spesifisert i oppdraget.

4.1.4 Skole 4: Konstruktivistisk eller fjerdegenerasjons tilnærming

Det siste tilskuddet til tilnærminger til evalueringer er den konstruktivistiske eller "fjerdegenerasjonstilnærmingen" som den gjerne kaller seg selv. Denne retningen som gjerne forbindes med Egon Guba og Yvonna Lincoln (1989). Det spesielle med "fjerdegenerasjonstilnærmingen" er at evaluering som søken etter kvalitet, resultat og verdi av tiltak eller program ikke lenger er et sentralt anliggende ved evalueringer. Evalueringer sees på som *forhandlinger* mellom parter med ulike interesser knyttet til evalueringsspørsmålet. Kjernen i en fjerdegenrasjonsevaluering kan beskrives som konstruktivistisk basert. Hensikten er å forhandle de ulike virkelighetsoppfatninger som finnes blant parter med ulike interesser knyttet til evalueringsspørsmålet. Tilnærmingen er relativistisk i den forstand at en aksepterer at evaluerer ikke a priori kan definere hvilken oppfatning av evalueringsspørsmålet som er den objektivt mest riktige. Oppfatninger av problemer og resultater kan variere mellom involverte parter og det finnes ingen objektiv måte for å avgjøre hvilken oppfatning som er den beste eller den rette. Hva som har kommet ut av et program eller tiltak avgjøres ut fra forhandlinger mellom de ulike oppfatninger av hva en har gjort, og hva en har fått igjen for innsatsen.

Dette er en radikal oppfatning av evalueringsoppgaven som endrer evaluerers rolle fra en forskerrolle til en dommerrolle. Tenkemåten, som innebærer en fullstendig rekonseptualisering av evalueringsfeltet, bygger likevel på anerkjente teoretiske posisjoner (Glaser & Strauss, 1967) og kan gi nye innsikter. For oppdragsgiver kan en slik tilnærming være problematisk dersom det ikke er

gitt i utgangspunktet at en ønsker ekstern bistand til forhandlinger mellom involverte parter. Et annet problem er at den eksterne forsker (dommer) sin rolle blir vanskelig å skille fra den rollen en internt evalueringsansvarlige skal ha. En må også kunne anta at det kan bli problematisk å vurdere hvilken politisk legitimitet denne type evalueringer vil ha. Sannsynligvis en fjerdegenerasjonsevaluerings legitimitet være bestemt av i hvilken grad forhandlingsgruppen får status som ekspertpanel. Denne statusen ville per definisjon måtte oppfattes som en sosial konstruksjon, men unntatt fra forhandlinger dersom prosjektet skulle kunne gjennomføres innenfor rådende praksis.

4.1.5 Skole 5: Sosial prosess tilnærming

Sosial prosess tilnærmingen forbindes gjerne med Lee J. Cronbach (Cronbach, 1980b) og Stanford miljøet. Denne tilnærmingen forkaster eksplisitt beslutningsstøttemodellen og representerer tenkemåter som ligger nærmere hovedtrenden innenfor kvantitativt orienterte kretser innen *American Evaluation Association* og også nærmere de tenkemåter som er representative for denne rapporten. Tilnærmingen innebærer en større vektlegging av målemodeller, samt avvisning av evalueringer som redskap for etterrettelighet[16]. Det sentrale for Cronbach og hans tilhengere er ikke evaluering i den konsekvensorienterte forstand begrepet benyttes i denne rapporten, men snarere evaluering som *forståelse* av et program eller tiltak. For å benytte en enkel pasientmetafor, for Chronbach er det viktigere å forstå hva symptomene utrykker enn å redusere symptomene. Sosial prosess skolen tar ikke evalueringsspørsmålet som gitt og unngår heller ikke å problematisere verdispørsmål. Den etiske dimensjonen

[16] Etterrettelighet (accountability) er det sentrale fokus for evalueringer forstått som et integrert element i NPM (New Public Management).

blir ofte eksplisitt drøftet og konklusjoner som gjelder evalueringsspørsmålet blir heller ikke søkt unngått.

Sosial prosess tilnærmingen utgjør et betydelig fremskritt i evalueringsforskningen utvikling. Det er likevel over tyve år siden denne tilnærmingen ble lansert, og mye har skjedd innenfor programevaluering når det gjelder kvantitativ modellering.

4.1.6 Økonomenes perspektiver på evalueringer

Evaluering har for økonomer tradisjonelt vært nesten ensbetydende med nytte-kosnads-analyser. Dette er analyser som måler nytten av et tiltak, kvantifisert i pengestørrelser opp mot kostnadene ved tiltaket, eller sagt med en kapasitet på området som Lewis A. Korhauser:

> "Cost-benefit analysis refers to a narrower class of procedures that evaluate policies in terms of the net benefits the policies provide to the individuals. Benefits are then usually defined solely in terms of the change in individual well-being that the policy induces, and costs are generally measured in terms of the monetary costs of resources required to implement the project. Again, typically, individual well-being is understood as satisfaction of subjective preferences; in practice these subjective values are inferred from market choices of individuals or are elicited through survey techniques. Comparison of costs and benefits thus requires that the cost-benefit analysts measure subjective benefits in monetary terms." (Kornhauser, 2000:1039).

I de senere år har dette perspektivet vært ytterligere fremhevet gjennom NOU 1997:27 (Hervik, Hagen, Nyborg, Scheel, & Sletner, 1997) og NOU 1998:16 (Hervik, Hagen, Nyborg, & Scheel, 1998). NOU 1997:27 gjør rede for det teoretiske grunnlaget for nytte-kostnadsanalyser og NOU 1998:16 gir anvisninger for praktiske anvendelser. Til sammen gir de to utredningene et solid kunnskapsgrunnlag for nytte-kostnadsanalyse.

Det er imidlertid ikke uproblematisk å la nytte-kostandsperspektivet være enerådende. Lewis A. Kornhauser utrykker problemet som følger: "Cost-benefit analysis influences, if not controls, many public decisions of great importance," but "its justificatory foundations remain at best suspect and at worse in ruins." (Kornhauser, 2000:1037) Nobelprisvinneren Amartya Sen har karakterisert nytte-kostnadsanalyse som "a daydream" (Sen, 2000:952) og Henry S. Richardson kaller det "stupid" (Richardson, 2000: 971). Blant økonomer som er eksplisitt om de politiske implikasjonene av teorier, er det en klar forståelse for at grunnen til skillelinjene ligger at "…cost-benfit analysis is inseperable from the free-market principle, and further, it attempts to apply these principles to the conduct of public affairs. Like the marketplace, cost-benefit analysis takes individual preferences as its guide, and it assumes that government should serve these preferences, not direct or educate them" (Wolfson, 2001:95).

Til tross for en ganske opphetet debatt blant ledende økonomer, er det likevel stor enighet om at nytte-kostnadsanalyser er et anvendbart, om enn ikke ufeilbarlig, verktøy for kontroll og vurderinger av offentlige utgifter. På konferansen om Cost-Benefit Analysis ved University of Chicago Law School i september 1999 der de fleste av feltets fremste tilhengere og kritikere var samlet, var det klart at de fleste, til tross for betydelig og fundamental uenighet, kunne samles om "The Statement of Principles on cost-benefit analysis" fra the American Enterprise Institute (AEI). I denne uttalelsen, som blant annet er signert Kenneth J. Arrow, Robert W. Hahn og Robert N. Stavins heter det at:

> Benefit-cost analysis should be required for all major regulatory decisions, but agency heads should not be bound by a strict benefit-cost test. Instead, they should be required to consider available benefit-cost analysis and to justify the reasons for their decisions in the event that the expected cost of a regulation far exceed the expected benefits.

Det er verdt å merke seg at uttalelsen har et svært pragmatisk preg og kan sees som en støtte til grunntanken i nytte-kostnadsanalyse, men med en klar melding om at en ikke må gi slike analyser en for streng eller absolutt betydning. Denne pragmatiske tonen innebærer en erkjennelse av metodens begrensninger, men også en innrømmelse av dens fordeler. Ser vi på den norske utredningsinstruksen av 18. februar 2000 fra Arbeids- og administrasjonsdepartement finner vi mye av den samme pragmatiske tonen: nytte-kostnadsvurderinger er nødvendige, men ettersom slike vurderinger bygger på usikre forutsetninger, bør vurderinger gjøres med omhu og ikke ensidig ta utfallet av slike analyser som handlingsdirektiv.

Et særlig problem med nytte-kostnadsanalyser, er at de faktisk ikke relaterer seg direkte til spørsmål om 1) hvilke effekter et gitt tiltak/program har hatt eller om en 2) har truffet målgruppen. Dette innebærer at nytte-kostnadsanalyser strengt tatt ikke har det samme fokus som i summative evalueringer og heller ikke samme fokus som formative evalueringer. Strengt tatt så forutsetter nytte-kostnadsanalyser at det foreligger en summativ evaluering, slik at en kan verdsette effekter i pengeverdier *når effektene er identifisert*. Økonometrikeren og økonomen James J. Heckman var en av de første til å påpeke denne sammenhengen (Heckman & Smith, 1998).

Empirisk evalueringsforskning som baserer seg på å etablere kausale effekter av inngrep i markeder, har tradisjonelt ikke vært et sentralt felt for økonomer. Denne type forskning forutsetter en eksperimentell logikk som gjør det mulig å isolere effekter av inngrepet. Økonometrikere har likevel i lang tid har levert betydelige bidrag til modeller (Roy, 1951), (Quandt, 1972) som kan anvendes for denne typen problemer.

Nyere arbeider innenfor empirisk økonometrisk basert evalueringsforskning (Heckman, Ichimura, & Todd, 1998; Heckman & Smith, 1999; Heckman, 1988; Heckman, 1995; Heckman, 1999; Heckman, Smith, & Clements, 1997b; Heckman & Smith, 1995; Heckman & Smith, 1998) har gitt bidrag som har brakt nye muligheter for forståelsen av hva som menes med summative evalueringer. Disse arbeidene er de viktigste pilarene for evaluering av effekter av tiltak/programmer og vil bli videre drøftet i kapitlet om evalueringens metodikk (kapittel 8).

4.2 Betydningen av skoler og retninger for oppdraget

Som vist i Tabell 1 kan en ut fra beskrivelsen av særtrekk ved de ulike skoler si en del om hva oppdragsgiver kan forvente seg av den evaluering som bestilles.

Tabell 1 ”Skoler” og trekk ved etablering av agenda for evalueringen

	Agendasettende temaer			
”Skoler”	Konklusjoner om resultater	Forholdet til interessenter (stakeholders)	Forholdet til oppdragsgiver	Basis for validitet
Besluningsstøtteskolen A	*Ja*	*Skjevt til fordel for oppdragsgiver*	*Innenfor*	*Rational prosess forståelse*
Den relativistiske tilnærmingen B	*Ja*	*Nøytral, men lydhør overfor oppdrags-giver*	*Utenfor, men lydhør overfor oppdrags-giver*	*Fagets metodiske verktøy*
‘Rich description’ tilnærming C	*Nei*	*Uavhengig*	*Utenfor, med, overblikk*	*Forståelse*
Sosial prosess tilnærmingen D	*Ja*	*Uavhengig*	*Uavhengig forsker*	*Design og målemetodikk*
Fjerdegenerasjonsevaluering E	*Nei*	*Forhandler*	*Uavhengig dommer*	*Vektet kompromiss*
”Økonometrisk evaluering” F	*Ja*	*Uavhengig*	*Uavhengig fosker*	*Design og målemetodikk*

Tabellen viser at en for de fleste skoler kan få konklusjoner om resultater av tiltaket. Det er bare *'rich description'* tilnærminger og *fjerdegenerasjonstilnærminger* som systematisk ikke leverer konklusjoner om resultater. Halvparten av de nevnte skolene har en grunnidé om at forskeren må forholde seg uavhengig i forhold til interessenter. Denne idéen er så grunnfestet innenfor mange forskningsmiljøer at det er overraskende at så mange evalueringsretninger avviker fra dette prinsippet. Innen *beslutningsstøtteskolen* blir forskeren nærmest en del av oppdragsgivers stab. Oppdragssgiver oppfattes som en kunde og kundens premisser er ikke diskutable[17]. Litt av den samme holdningen finner en innenfor den *relativistiske* tilnærmingen, men her er det likevel rom for en viss kritisk avstand til oppdragsgiver. Den mest atypiske måten å forholde seg til både interessenter og oppdragsgiver finner en innenfor *fjerdegenerasjonstilnærmingen*. Her har evaluerer en *meklerrolle* i forhold til motstridende interesser.

Offentlige oppdragsgivere er godt i stand til å skille mellom en forskerrolle og en konsulentrolle når en evaluering skal gjennomføres. Meklerrollen hos tilhengerne av *fjerdegenerasjonevaluering* vil nok virke forvirrende og fremmed. Det er likevel grunn til å anta at oppdragsgivere vil forvente mer nøytralitet og avstand hos er forskningsmiljø enn i et konsulentselskap og dermed oppleve *beslutningsstøtteskolen* som bærere av mer av en konsulentrolle enn en forskerrolle. Poenget her er at de ulike skoler viser stor variasjon i måten de forholder seg til interessenter og oppdragsgivere. Om en forventer en tradisjonell forskerrolle, kan det være forvirrende om en møter noe som bryter med etablerte forventninger.

[17] Kunden har alltid rett.

Når det gjelder de ulike skolers standarder for hva som regnes som basis for gyldige konklusjoner, viser dette også betydelige variasjoner. *Beslutningsstøtteskolen* bygger på en forståelse av evaluering som en tilnærmet rasjonell prosess, dvs. at evaluering i all hovedsak er en prosess der informasjon om utfall av tiltak gir handlingsalternativer som begrunner beslutninger om tiltakets videre skjebne. Slike forutsetninger om involverte aktørers evner og muligheter for konsistente og rasjonelle handlinger er som regel lite fremtredende innen de fleste andre skoler. Et unntak fra denne regelen er nyere *økonometrisk evaluering* som har sitt utgangspunkt i økonomisk teori. Basis for validitet innenfor *økonometrisk evaluering* er likevel ikke de teoretiske modellenes prediksjoner, men de økonometriske modellenes estimater. *Økonometrisk evaluering* har følgelig mer til felles med den *relativistiske* tilnærmingen og *sosial prosess* tilnærmingen enn med *beslutningsstøtteskolen* ettersom de tre førstnevnte i all hovedsak henter sin legitimitet for konklusjoner fra egenskaper ved design og statistisk modellering.

For *rich description* og *fjerdegenerasjonstilnærmingen*, blir den metodiske tilnærmingen svært ulik de andre skolenes. Når det gjelder *rich description* tilnærmingen er det *forståelse* av det tiltak som evalueres som gir grunnlaget for vurderinger. For *fjerdegenerasjonstilnærmingen* bygger vurderinger på et vektet kompromiss mellom motstridende interesser. Hvilke konsekvenser disse to tilnærmingene har for metodikk, er ikke helt klart formulert i de publikasjoner som må regnes som representative for disse skolene. Det er likevel klart at det legges mindre vekt på kvantitativ empirisk modellering og mer vekt på kvalitative metoder.

For oppdragsgiver er de ulike skolenes basis for gyldige konklusjoner av betydning. Ved offentlige evalueringer vil mekling mellom motstridende interesser som regel være av mindre interesse ettersom utfall av tiltak også vurderes av politiske myndigheter. Lokal enighet blant involverte parter som et resultat av mekling er neppe tilstrekkelig grunnlag for politiske vurderinger av et gitt tiltak. Det er mer trolig at motstridende politiske interesser vil føre til at det reises tvil om konklusjoner, noe som øker nødvendigheten av en legitim basis for konklusjoner. Konkret betyr dette at kravene til etterprøvbarhet og etterrettelig metodikk kan bli *høyere* for evalueringer enn ved andre former for forskning og utredning. Konflikt kan utløse kontrollkrav som krever etterprøvbar metodikk og design som kan gi grunnlag for de konklusjoner som fremsettes. I de tilfeller *uavhengige* evalueringer av evalueringen etter nøye gjennomgang av metodikk og prosedyrer bekrefter konklusjoner, styrkes troen på forskningsbaserte evalueringer. Dersom etterprøving av gjennomførte evalueringer underkjenner konklusjoner, eller finner at den benyttede metodikk gjør forskerens subjektive skjønn så dominerende at etterprøving ikke kan gjennomføres, så svekkes troen på forskningsbaserte evalueringer. Når tiltroen til forskningsbaserte konklusjoner eroderer, svekkes offentlige evalueringer som styringsinstrumenter. For oppdragsgivere i det offentlige styringsverket er dette negativt, ettersom det i mange tilfeller ikke finnes alternative kilder til kunnskap om de forhold en ønsker å undersøke. Det kan også hevdes at når troen på forskningsresultater svekkes, da styrkes tilliten til a priori oppfatninger om sammenhenger og tilstander ettersom det ikke foreligger mer grundige forklaringer. En slik utvikling svekker på sikt det offentlige styringsapparatet i forhold til omskiftelige politiske vurderinger som ikke er undergitt de samme krav til realitetsforankring. – Ved vurderinger av hvilke evalueringer de ulike skoler og retninger kan forventes å levere, må både kortsiktige og langsiktige vurderinger over hva en ønsker med evalueringer, tas med.

5 ORGANISERING AV EVALUERINGSAKTIVITETEN

5.1 Perspektiver fra agentteori

Organiseringen av evalueringspraksis er problematisk. Den omfatter parter med interesser knyttet til utfallet av evalueringen, involverer kontrakter mellom oppdragsgiver og evaluerer og forutsetter faglige vurderinger på områder der konkurrerende skoler innen evalueringsforskningen kjemper om hegemoni.

Økonomisk agentteori kan bidra til å systematisere problemene. Kort fortalt omhandler teorien forholdet mellom prinsipal og agent. Disse to partene er gjensidig avhengig av hverandre, men kan ha ulike interesser i samarbeidet. En tradisjonell tenkemåte tilsier at agenten, som ideelt sett skal handle på vegne av prinsipalen, ønsker å minimalisere sin egen arbeidsinnsats og risiko. Prinsipalen kan ikke utføre oppgaven selv, og må gjennom forhandlinger tilby kompensasjon som gjør agenten villig til å påta seg oppdraget. Agentteorien antar *opportunisme* i forholdet mellom prinsipal og agent, begge parter motiveres primært av egen nytte. Teorien forutsetter ikke antakelser om hierarki eller formell autoritet, men at relasjonen mellom prinsipal og agent er styrt av implisitte eller eksplisitte forståelser, kalt kontrakter. I analyser av agentproblemer er det denne forståelsen, eller kontrakten, som regulerer forholdet mellom prinsipal og agent.

En slik skjematisk tenkemåte avdekker flere problemer:

a) *Agentproblemet*. Agenten og prinsipalen har ulike ønsker og interesser i situasjonen

b) *Problemet med risikodeling*. Agenten og prinsipalen kan ha ulike holdninger til risiko i situasjonen

c) *Informasjonsasymmetri*. Prinsipalen har ikke mulighet til å overvåke agenten eller løpende kontrollere at forpliktelser overholdes

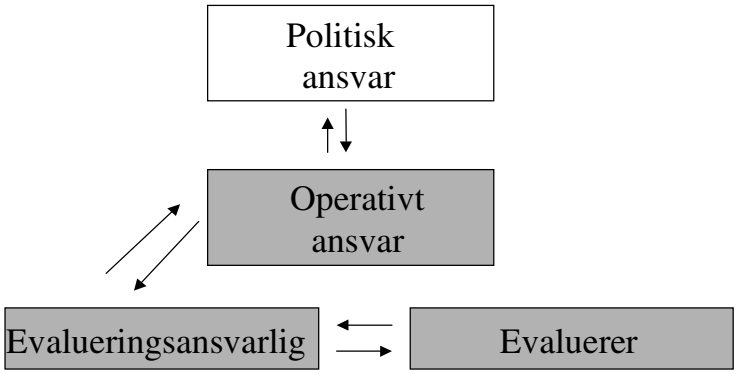
Agentteorien er nær beslektet med transaksjonskostnadsteorien, der en organisasjon betraktes som et sett av kontraktsrelasjoner mellom individer eller koalisjoner. Ettersom oppgaver og ansvar etableres, oppstår nye prinsipal-agent relasjoner med mer eller mindre vel kommuniserte kontrakter. *Agentproblemer* kan oppstå, for eksempel når agenten gjennom forpliktende avtale påtar seg oppgaver som vurderes som ugunstige i forhold til egne interesser. Om oppdraget innebærer *risiko* for agentens egen karriere, er det sannsynlig at agenten har en annen holdning til slik risiko enn prinsipalen. Informasjonsasymmetrien gjør at prinsipalen ikke kan kontrollere i detalj hvorvidt agentens handlinger er i samsvar med prinsipalens ønsker. *Begrenset tillit* oppstår når prinsipalen på grunn av mangelfull informasjon ikke kan kontrollere at prinsipalen virkelig gjør som avtalt. Begrenset tillit medfører *agentkostnader,* i litteraturen gjerne spesifisert som :

- *Styringskostnader* – kostnader knyttet til det prinsipalen må gjøre for å sikre at agenten utfører oppdraget som avtalt
- *Forpliktelseskostnader* – den kompensasjon som er nødvendig for at agenten skal sette prinsipalens interesser foran egne interesser
- *Residualkostnader* – kostnader som påløper dersom resultatet ikke blir som prinsipalen ønsker, dvs. prisen for differansen mellom faktiske og ønskelige resultat

Praktisk anvendt kan vi forstå relasjonen i figur 1 som prinsipal-agent relasjoner. For å holde resonnementet innenfor en enkel økonomisk tenkemåte og

holde politiske teorier utenfor i denne omgang, kan vi begrense resonnementene til å gjelde tre aktører, *operativt ansvarlig*, *evalueringsansvarlig* og *evaluerer*. I relasjonen *operativt ansvarlig – evalueringsansvarlig* er operativt ansvarlig prinsipal og evalueringsansvarlig agent. I relasjonen *evalueringsansvarlig – evaluerer* er evalueringsansvarlig prinsipal og evaluerer agent.

Det overordnete problemet er residualkostnader. De to skisserte agent-prinsipal- relasjonene vil med høy sannsynlighet gi kostnader som følge av at evalueringen ikke gir en riktig tilbakemelding, for eksempel ved at effekter av tiltak overvurderes eller undervurderes. Ser vi på relasjonen evaluerer – evalueringsansvarlig er de *styringskostnader* som påløper i all hovedsak utgifter til referansegruppe og almen møtevirksomhet knyttet til evalueringsansvarlig sitt behov for kontroll over evaluerer.



Figur 1. Hovedaktørene i evalueringsprosessen

Faren for opportunistisk adferd hos evaluerer knyttet til kjennskap til relasjonen mellom evalueringsansvarlig og operativt ansvarlig, blir sjelden vurdert, og er også vanskelig å kontrollere. Operativt ansvarlig vil i de fleste tilfeller *både* ha interesse av at evaluering blir forsvarlig utført og at evalueringen er positiv. Denne tvetydigheten gir grobunn for opportunisme i relasjonen mellom opera-

tivt ansvarlig og evalueringsansvarlig. Evaluerer, som har interesser knyttet til nye evalueringsoppdrag, kan enkelt utnytte dette potensialet for opportunisme, særlig når analysemetodikk er valgfri og lite kommunisert og begrunnelser for tiltaket som skal analyseres er uklare eller tvetydige. Innenfor statlig virksomhet er det få eller ingen betraktninger over den risiko evalueringsansvarlig utsetter seg for ved å være den som administrerer evalueringsoppdraget. Det finnes ingen påviselige styrings- eller forpliktelseskostnader som skal kompensere denne risikoen, snarere er det et inntrykk at prinsipalen, den operativt ansvarlige, lar disse kostnadene hvile på agenten i håp om en viss opportunistisk adferd som kan være til gunst i forhold til overordnet prinsipal, den politiske ledelse.

Denne type risiki knyttet til organiseringen av evalueringspraksis kan naturligvis utbroderes videre. Poenget er at lite er gjort for å minimalisere risiko for opportunistisk adferd. Et annet poeng er den åpenbare *informasjonsasymmetri* som ligger i forholdet mellom evaluerer og evalueringsansvarlig i og med at evalueringsansvarlig i alt for liten grad får anledning til å sette seg inn i de teorier og metoder evaluerer benytter i sitt arbeid.

5.2 Institusjonelle perspektiver på organiseringen

Ser en på veksten i omfanget av evalueringer på ulike nivåer i offentlig sektor de siste ti år, kan en få inntrykk av at det er ønskelig med mange og hurtig gjennomførte evalueringer. Ideen om at de fleste saker og ting kan evalueres og helst bør evalueres så ofte og hurtig som mulig, kan minne om en tvangstanke, om den bare eksisterte i ett hode. Når tanken ikke finnes uten forbehold i noe enkelt hode, mens praksis tilsier at en slik idé må eksistere et sted, snakker vi om *institusjonalisering* av rutiner og handlemåter. Selv om evalueringer av offentlige tiltak og programmer etter hvert er blitt nokså vanlig i norsk for-

valtningspraksis, er vi fortsatt i en fase der rutiner og prosedyrer ennå ikke helt har festnet seg. I denne fasen er det viktig at vi tenker gjennom hva vi ønsker å oppnå med evalueringer, og om de rutiner som er i ferd med å etableres, er tjenlige for våre formål.

Det eksisterer så vidt jeg vet ingen skriftlig instruks for *hva* som skal evalueres eller *hvordan* evalueringer skal gjennomføres. Likevel gjennomføres evalueringer av programmer, hele etater og enkle dokumenter som kommunale reiselivsplaner har blitt gjenstand for evalueringer. Dette betyr at det har funnet sted en *institusjonaliseringsprosess,* en prosess der mange parter har blitt innforstått med at evalueringer, det er noe vi har begynt med nå. Trekk ved institusjonaliseringsprosessen har åpenbare konsekvenser for hva som ”går seg til” som den valgte organisasjonsmodell for evalueringsvirksomheten, samt for hvilke skoler og retninger som blir toneangivende. Når en organisasjonsform, tenkemåte eller metodikk har fått sin plass i etablerte rutiner og prosedyrer, kan det være vanskelig å endre dette, selv om en kan ha gode argumenter.

Institusjonsbegrepet har blitt komplisert fordi samme begrep benyttes i en rekke forskjellige betydninger. Begrepet, slik det benyttes i denne rapporten er godt forklart av sosiologen Dag Østeberg:

”Når samkvemmet mellom mennesker foregår i henhold til normer som mer eller mindre ligger i luften, uuttalt, underforstått, danner samkvemmet en struktur eller et system. Derfra til fastsatte normer og regler for den innbyrdes atferd er det et steg, en overgang — overgangen til *institusjon* — en sosiomateriell struktur med gjerne skrevne regelverk, egne tilholdsrom eller bygninger og ofte egne ansatte (spesialister). Instituering er faguttrykket for det som løselig kalles å bringe noe "i fastere former". (Østerberg, 1994:85).

Denne måten å forstå begrepet er også den mest vanlige innenfor administrasjonsfagene. Stikkord for å bestemme en institusjon, er ut fra dette *permanens* ("faste former"), *formelle organisasjons- prosedyrer* funksjoner og praksiser

nedfelt i regler og rutiner) og *profesjon* (hierarkier av ansatte som behersker og ivaretar ulike sider ved institusjonens virksomhet). Institusjoner representerer immaterielle trekk, normer, tolkninger, verdier, diskurser, ideer og tanker som sirkulerer i og rundt en bestemt sosial praksis, samt de materielle og fysiske manifestasjoner som gir rammene for praksis (Syvertsen, 1999). En institusjonaliseringsprosess blir, ut fra dette, gangen mot fastere former for praksis for samhandling

En slik institusjonaliseringsprosess kan, litt enkelt sagt, oppstå på tre måter (Kvitastein, 2000):

i) Den kan starte som en tilfeldighet, ved at noen som er i posisjon til det finner de for godt å imitere noe som er observert et annet sted.

ii) Den kan oppstå som en prosess der deler av rutiner i en avdeling, etat eller virksomhet kopieres til en annen avdeling, etat eller virksomhet og blir en ekspanderende standard operasjonsprosedyre.

iii) Den kan bli etablert med hensikt for å tjene spesifikke formål.

I det første tilfeller blir det som imiteres det toneangivende, selv om de prosedyrer som etterlignes ikke alltid er like velegnet for de nye oppgavene. I det andre tilfellet kan det være av stor betydning om de oppgaver som finnes der en kopierer *fra*, ligner på de oppgaver som finnes i den avdeling eller etat en kopierer til. Det siste tilfellet, den bevisste tilpasning av rutiner tilpasset avdelingen eller etatens spesifikke oppgaver, er selvsagt den mest ønskelige løsningen, selv om en som regel trenger rom for justeringer.

Det finnes en omfattende litteratur om hvilke mekanismer som tjener til å forsterke og opprettholde institusjoner. Teorier om konsekvenser av den menne-

kelige tendens til vanedannelse (Peirce, Cohen, & Dewey, 1923, Veblen, 1899) på varige, men merkelig og lite tjenlige ordninger for samhandling er ikke noe nytt. De siste årene har det likevel funnet sted en betydelig systematisering av begrep og tenkemåter kring hvilke fenomen som forsterker eller opprettholder institusjoner. En slik systematisering (Scott, 1995) er gitt i tabell 2.

Tabell 2 Institusjonenes tre bærebjelker Tilpasset etter Scott (1995)

	Regulerende	*Normativt*	*Kognitivt*
Grunnlag for opprettholdelse	Hensiktsmessighet	Sosial pliktfølelse	Tas for gitt
Mekanisme	Tvingende	Normativ	Etterligning
Logikk	Instrumentalitet	Passende	Ortodoksi
Indikatorer	Regler, lover	Sertifisering	Vanlig forekomst
Legitimitets-basis	Legalt sanksjonert	Moralsk styrt	Kulturelt støttet og språklig korrekt

Med utgangspunkt i hvilke fenomener som gir grunnlag for opprettholdelse og forsterkning av institusjonaliserte former, kan de tre bærebjelkene betegnes som det regulerende, det normative og de kognitive element. De regulerende prosessene omfatter kapasitet til å etablere rutiner og regler, belønninger og sanksjoner som gjør det mulig å vokte og påvirke konformitet. Den måten å se institusjoner på er den mest vanlige og moderate. Den er også fullt forenlig med idéen om at aktører har interesser som de ivaretar rasjonelt. Den normative måten å se institusjoner på hevder at institusjoner påtvinger aktørene både normer og verdier som forsterker og opprettholder de foretrukne og tolererbare standarder for adferd. Normene spesifiserer hvordan ting bør gjøres og definerer hva som er legitime midler for å oppnå ønskede mål. Det kognitive synet på institusjoner vektlegger de aspekter ved institusjoner som preges trekk ved individuell informasjonsprosessering. Denne måten å se institusjoner på er den

dominerende innenfor sosiologi og organisasjonsteori (Powell & DiMaggio, 1991).

Sett i forhold til den beredskap som er nødvendig for å unngå å stivne i arbeidsmåter for evalueringspraksis som er lite tjenlige, må en ha et blikk for at alle de tre bærebjelkene er aktive og virker konserverende på eksisterende praksis, uavhengig av hvordan praksis har oppstått. Sett i forhold til den kognitive bærebjelke er det enkelt å se at det krever minimal mobilisering av kunnskap å ta eksisterende ordninger for gitt, bare fordi de fremtrer som de aksepterte rutiner og prosedyrer. I forhold til implementeringen av det rasjonale som ligger i NPM som innebærer et premiss om at administrasjon i det offentlige stor sett er lik administrasjon i det private, så er den individuelt sett minst krevende løsning å akseptere etterligninger av prosedyrer og rutiner, selv om en faktisk oppfatter at dette premisset kan være problematisk. Den logikk som begrunner slik tenkevegring kan karakteriseres som ortodoksi. En reagerer ikke på eksisterende rutiner, rett og slett fordi de er vanlige og omsluttet av den begrepsbruk som oppfattes som korrekt og gyldig for den type oppgaver eller problemer som diskuteres. Den språklige binding som konstituerer hva som er gyldig kommunikasjon innenfor den gjeldende organisasjonskultur forsterker på denne måten eksisterende institusjoner. Samtidig gjør denne språklige bindingen kommunikasjon enklere, gitt at implisitte premisser aksepteres.

5.3 Institusjonalisering av kontroll

Sammenhengen mellom agentproblemer og institusjonaliseringsprosesser kan synes åpenbar: Kontrollproblemene blir styrende for valg av organisasjonsmodeller. De valgte former ”går seg til” i en form nær det intenderte, over tid blir innarbeidede rutiner og standard operasjonsprosedyrer vanskeligere å forandre. Slike sekvenser er i samsvar med teori, men hviler likevel på en del kritiske

utestbare forutsetninger. De mest kritiske forutsetningene gjelder de antakelser om *adferd* som begrunner kontrollproblemene.

En mye sitert artikkel (Goshal & Moran, 1996) som gjelder kritikk av transaksjonskostnadsteori (Williamson, 1985) innleder med historien om to økonomer som våknet en natt av en tiger som gikk og lusket utenfor teltet. Den ene grep etter skoene sine, klar til å løpe. Da kameraten gjorde han oppmerksom på at han ikke ville kunne løpe fra tigeren, svarte han at det var heller ikke nødvendig, den eneste han måtte løpe fra var kameraten. Historien er en god påminnelse om forskjellen på biologisk og økonomisk konkurranse. Har et tenkesett eller en teori festnet seg, for eksempel ”survival of the fittest” så kan en handle på refleks uten tanke for alternativer og konsekvenser. I denne i historien ville selvsagt bare tigeren overleve. På litt sikt, for eksempel 10 minutter, kunne tigeren sikkert godt tenke seg å spise dem begge. Ved å samarbeide eller å klatre opp i et tre kunne kanskje begge ha overlevd.

Agentteori og transaksjonkostnadsteori (TCE) er nært beslektet og bygger på svært parallelle adferdsforutsetninger, særlig antagelsen om opportunisme. I agenteorien fremstår denne antagelsen som et eksplisitt premiss, mens TCE fremstiller opportunisme som en sterkere form for egeninteresse som igjen er en del av ”den menneskelige natur” (Williamson, 1993). Opportunisme i betydningen sterk egeninteresse blir hos Williamson selve grunnen til svikt i markedene og følgelig også grunnen til at organisasjoner i det hele finnes (Williamson, 1993b). Uten dette menneskelige trekk kunne de fleste transaksjoner vært gjennomført i autonome kontrakter. For TCE blir oppgaven å redusere tansaksjonskostnader som oppstår på grunn av den menneskelige tilbøyelighet til opportunistisk atferd, i agentteorien blir oppgaven å redusere de *agentkostnader* som påføres prinsipalen på grunn av begrenset tillit. Både agentteori og TCE gir premissene for mye av den tankegang som danner basis

for NPM og følgelig for de kontrollkostnader som følger av kravet om etterrettelighet.

Problemet med opportunismebegrepet er at det tas for gitt at "the human condition" gir egeninteresse som grunnholdning, og at holdning entydig predikerer handling. Dette er ikke i samsvar med litteraturen om holdning og handling (attitude-behavior) som sier at det *ikke* er noen ubetinget enkel sammenheng mellom holdning og handling. Hovedtrenden innen området betrakter forskjellig adferd som bestemt av holdninger og subjektive normer. (Ajzen & Fishbein, 1977). Et individs holdninger er påvirket både av egen adferd og av individets oppfatninger av *andres* holdninger og adferd (Eccles & Wigfield, 2002). Det kan derfor være mer naturlig å behandle opportunisme ikke som noe gitt, men som en *variabel.* Ser en opportunisme som en variabel, som noe en i mer eller mindre grad kan være disponert for, betinget av forhold som kan forsterke eller svekke tendenser til slik adferd, åpner dette for innsikt om konsekvenser av bruk av kontrollsystemer.

Opportunisme kan sees som påvirket av tre faktorer (Goshal & Moran, 1996):

1. På forhånd individspesifikke betingelser som holdninger, verdier og tidligere erfaringer som preger en, bevisst eller ubevisst.

2. Følelser for enheten, dvs. avdelingen, institusjonen eller oppgaven, samt ens kolleger.

3. Opportunistisk adferd.

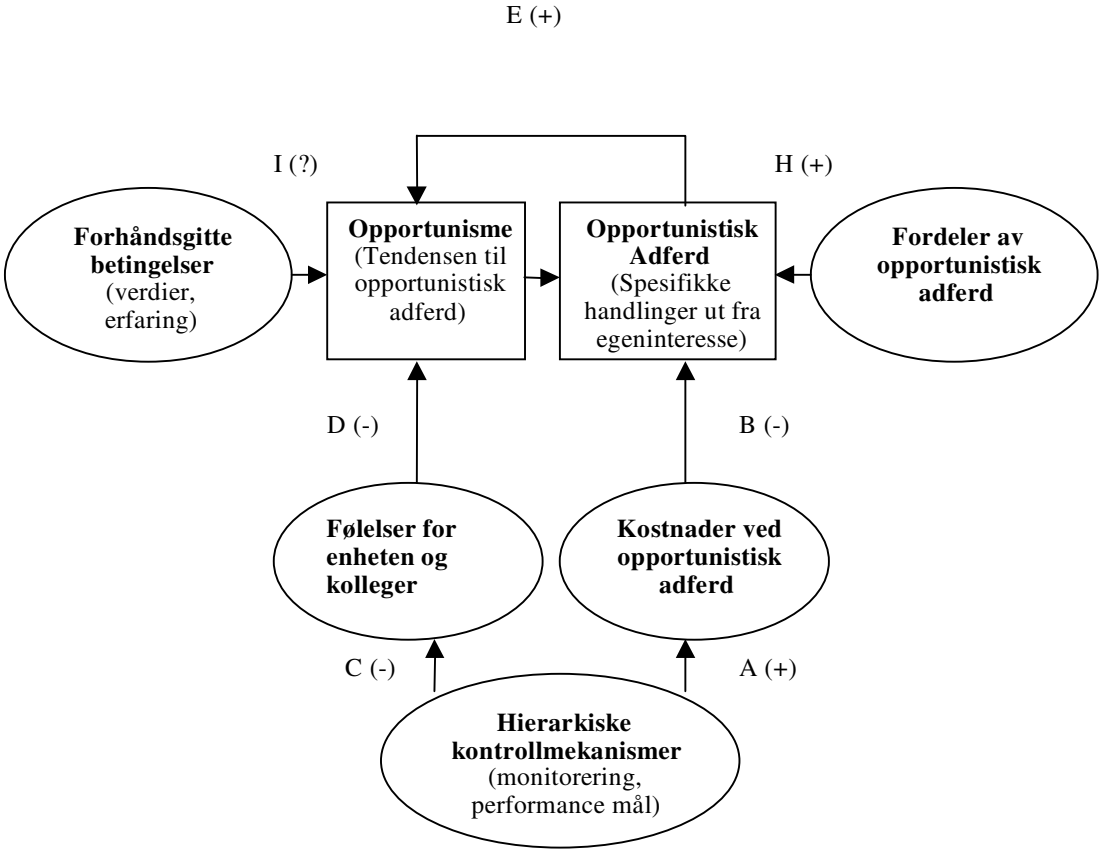
Disse tre faktorene danner et samspill mellom (1) mennesker med ulike erfaringer og tilbøyeligheter til å trekke større fordeler enn legitimt i typer situasjoner der dette er mulig, følelser (2) av at ens arbeid blir verdsatt og at oppgavene er

meningsfulle, samt (3) i hvilken grad en tidligere har handlet på måter som kan karakteriseres som opportunistisk.

Problemet er at disse tre faktorene kan fungere sammen på måter som kan gi selvoppfyllende profetier. Spissformulert kan en hevde at om en *forutsetter* opportunisme så organiserer en gjerne kontrollsystemer som *skaper* opportunisme. Som vist i Figur 2 kan subjektive holdninger, verdier og erfaringer (I) som er etablert før inntreden i en gitt stilling regnes som eksogent gitt. Etablerte individuelle disposisjoner for opportunisme må antas å variere mellom personer og det må antas at læring som både gir mer og mindre opportunisme vi forekomme. Det er likevel klart at individuelle disposisjoner er en ukjent størrelse som virker inn på tilbøyeligheten til opportunistisk adferd. På samme måte må en anta at *utbyttet* av opportunistisk adferd (H) må virke inn på opportunistisk arbeid og da slik at jo større utbytte, dess mer opportunistisk adferd. Ethvert kontrollsystem har kostnader (A) og må forventes å redusere opportunistisk adferd via sanksjoner (B). De kritiske punktene i modellen er stien fra kontrollsystem til følelser for egen enhet og kolleger (C) og stien fra denne boksen til opportunisme (D). Som påvist av flere forskere (Ajzen & Fishbein, 1977, Eagly & Chaiken, 1992) virker en positiv følelse for egen enhet, oppgaver og kolleger til å redusere tilbøyeligheten til opportunisme, mens omfanget av kontroll bidrar til å redusere positive følelser for egen enhet.

Dersom følelser for egen enhet reduseres som følge av for sterkt opplevelse av kontroll, for eksempel ved an en føler at en ikke blir verdsatt for sin innsats, kan dette øke tilbøyeligheten til opportunisme. Gitt at dette i sin tur fører til opportunistisk adferd, tilsier kognitiv dissonansteori (Festinger, 1957) at holdninger justeres i retning av *mer* opportunistiske tilbøyeligheter for å redusere kognitiv dissonans (E). Avstanden mellom handling og holdning gir dissonans.

Ettersom handling er gjort og bare holdning kan endres, oppstår et behov for justering for å finne balanse mellom holdning og handling.



Figur 2 Sirkelen for selvoppfyllende profetier

Stien C -> D -> E skaper en vond sirkel der større vekt på kontroll skaper større behov for kontroll. Sirkelen gir en selvoppfyllende profeti; den som ikke gis tillit må forventes å oppføre seg som en som ikke har tillit. For utforming av kontrollregimer har dette stor betydning. Ettersom agentperspektivet er et grunnleggende premiss innen NPM og kontrollsystemene er selve kjernen i det

system som skal gjøre et ikke-hierarkisk, autonomt beslutningssystem etterrettelig, er dette et betydelig problem. Institusjonell teori tilsier at det vil være vanskelig å endre på rutiner og praksis som har festnet seg, selv om mange kan se at initiativ og virkelyst blir skadelidende. Det er en betydelig oppgave å finne den balanse mellom kontroll og respekt som gir et fleksibelt of velfungerende administrativt system.

Det synes å være et potensiale for legitimitetsgevinst i reorganiseringer som reduserer mistanker om opportunistisk adferd. Beredskap for justeringer og reorganisering tolkes ofte enten som et behov for resultatmål, et behov for monitorering av innsats eller et behov for informasjonssystemer som gir informasjon om hva som gjøres i organisasjonen. Nyorienteringen i retning av NPM innenfor norsk offentlig administrasjon synes nettopp å være drevet av en sterk vilje til å påta seg styringskostnader. Det kan virke som prinsipalens muligheter til å kontrollere agenten nesten gjøres til en hovedsak i styringsverket, uten særlig blikk for styringskostnader. Det kan også synes som om den desentralisering av beslutningsmyndighet som er et sentralt trekk ved NPM, i stor grad begrunner denne viljen til å påta seg styringskostnader. Dette monofokus på kontroll synes å hvile på to bristende premisser, nemlig at: a) utstrakt implementering av kontrollsystem gir en legitimitetsgevinst og at b) kontrollsystemene faktisk virker som forutsatt.

Troen på at utstrakt bruk av kontrollsystemer i offentlig sektor automatisk gir legitimitetsgevinst undergraves fort om folk oppfatter det som at ressurser flyttes fra innsats for å få oppgaver gjennomført til innsats for å kontrollere at oppgavene er gjort. I tillegg styrkes populistiske oppfatninger av offentlig ansatte som potensielt utro tjenere. Mye tyder på at på at grunnpremisset fra NPM om at kontrolloppgaver stort sett er like i offentlig og privat virksomhet, halter. For

privat sektor forventes ikke kontrollsystemene å gi legitimitetsgevinst, de oppfattes mest som nødvendige.

Når det gjelder troen på at kontrollsystemer for kontinuerlige overvåking av at agenten handler på vegne av prinsipalen, gir nyere forskning tvetydige svar. Standard svar innen den eldre litteraturen er slike kontrollsystemer sett som informasjonssystemer vil redusere informasjonsasymmetri og følg minske prinsipalens styringskostnader. Nyere forskning tyder på at eksistensen av kontinuerlige virkende kontrollsystemer gir agenten incitament til å *utnytte* informasjonsasymmetri for å redistribuere kostnader i egen favør. I tillegg viser forskning at fokus mot mer eksterne kontrakter gjerne gir tap av fokus på mer interne prinsipal -agent relasjoner. Dette betyr stort sett at fokus på ekstern kontroll gir mer intern opportunisme (Jacobides & Croson, 2001). Generelt så antyder forskningsresultater at det er bedre å maksimere interne fellesverdier enn å søke å minimalisere prinsipalkostnader. Dette betyr at mer kontroll og informasjonssystemer ikke nødvendigvis er svaret. Utbredt bruk av slike systemer kan både skade og omfordele organisasjonens evne til verdiskaping.

6 FREMVEKSTEN AV EVALUERINGSPRAKSIS

6.1 Fra skoleforskning til GPRA

Innenfor skoleverket i USA vokste de i perioden 1920 til 1945 frem løst organiserte aktiviteter som arbeidet med målinger av elevers evner og prestasjoner. Disse *førstegenerasjons* evaluerere hadde en klar teknisk orientering. Målet var gjerne utvikling av målemodeller. Fra 1940 til 1960 finner vi et skift fra ingeniørrollen mot en opplyserrolle. For denne *andregenerasjonen* av evaluerere var oppgaven gjerne å beskrive hvilke tiltak som gav gode resultat og hvilke som gav mindre gode utfall. For *tredjegenerasjons* evaluerere i tiden etter 1960 var det *dommerrollen* som gav evalueringsaktiviteten legitimitet. I denne perioden var det gjerne måling mot standard som karakteriserte evalueringer av skoleprestasjoner. I sammen periode vokste troen på samfunnsforskningen, særlig økonomisk forskning, mot sine største høyder. Modeller fra *education* ble adoptert og videreutviklet både metodisk (Campbell & Stanley, 1963) og konseptuelt (Jonassohn, Coleman, & Johnstone, 1961), særlig av sosiologer i det som gjerne regnet som evalueringens gullalder. Utover 70-tallet svant mye av optimismen og troen på samfunnsvitenskapene som problemløsere og med Reaganadministrasjonen tørket finansieringskildene for evalueringsforskningen nesten ut. Når (Guba & Lincoln, 1989) lanserer sin bok for *Fourth Generation Evaluation* gode 10 år etter slutten for samfunnsvitenskapens gullalder, er navnevalget nettopp med referanse til den inndelingen i epoker som er nevnt over. Med innføringen av *the Government Performance and Results Act of 1993* får evalueringsforskningen ny giv. Loven pålegger offentlige institusjoner årlige utviklingsplaner og krever evalueringer av måloppnåelse for spesifikke tiltak over en viss størrelse. De idéer som danner basis i denne andre fase av evalueringens historie er likevel forskjellig fra første fase. Nå er det tankesettet fra New Public Management (NPM) som er det styrende.

NPM som ideologi er nettopp mye preget av idéer fra Reagan og Thatcher-perioden.

6.2 Forskning, tillit og politikk

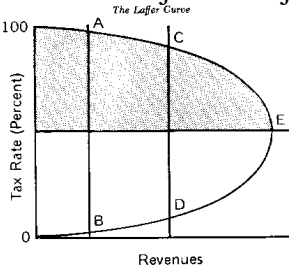
Det kan virke paradoksalt at enkeltforskere og forskningsinstitusjoner skal stå som garantister for politikkens legitimitet gjennom evalueringer, samtidig som samfunnsvitenskapens egen status har vært jevnt fallende de siste tretti år. Samfunnsvitenskapens fallende status har vært påpekt av flere, bl.a. av nobelprisvinner i økonomi i 2000, James J. Heckman. For over 10 år siden, i en artikkel der fem blant verdens ledende samfunnsforskere kommenterte han, sammen med flere andre (Aaron, Gramlich, Hanushek, Heckman, & Wildawsky, 1990), Robert Haveman og Richard Nathan sine store oppsummeringsverk (Haveman, 1987, Nathan, 1988) om samfunnsforskningen bidrag til offentlig politikk og problemløsning. De to bøkene gir en grundig oversikt over hva som har kommet ut av policy relevant forskning, i all hovedsak i USA, fra midten av 60 tallet til slutten av 80 tallet, og gir konklusjoner som trolig er like gyldige for Norge som for USA. Det er en samstemmig konklusjon hos de to forfatterne at viljen til å finansiere samfunnsforskning har vært den avgjørende faktor for hva en har fått ut av samfunnsforskningen, og at denne viljen har variert med skiftende tro på verdien av kunnskap og utdanning. Perioden under Kennedy- Johnson- administrasjonen var preget av stor tiltro til samfunnsforskningen, særlig til utviklingen innenfor økonomifaget. Perioden 1965-1980, The Great Society perioden, gav en formidabel vekst i satsingen på samfunnsvitenskap som etablerte en betydelig andel av dagens sentrale begreper og metoder. Mot slutten av perioden frem mot 1980 begynte både tiltroen og finansieringen å bli borte. Nå begynner Reagans og Thatchers tidsepoke der nøkkelbegre-

pene[18] for næringspolitikken er *"suppy-side economics"* og deregulering (Derthick & Quirck, 1985). Det er i denne perioden grunnlaget for The New Public Management (NPM) etableres. Det er med andre ord i en tid hvor samfunnsvitenskapens legitimitet og finansielle grunnlag er svekket, at tankene om at administrasjon i det offentlige ikke er vesensforskjellig fra administrasjon i det private, etableres.

Det kan også bemerkes at på tross av den manglende suksess for tankene om "supply-side economics" fra Reagan perioden[19] fikk disse idéene sin renessanse midt på 1990-tallet med introduksjonen av *"flat-tax"* som var et sentralt poeng i nominasjonsprosessen foran presidentvalget i 1996, særlig hos kandidaten Steve Forbes. Også etableringen av The Supply-Side University med mer enn 3000 studenter fra 45 forskjellige land siden 1997, viser at dette er idéer som i høyeste grad er levende, til tross for de ikke blir tatt alvorlig av økonomer flest.

---

[18] Det kanskje mest kjente begrepet fra perioden med "Reaganomics" er den såkalte Laffer-



kurven, (figuren over) som indikerer at en under visse forutsetninger kan øke statens inntekter ved å senke skattesatsen. Kurven som er oppkalt etter økonomen Arthur Laffer er mye diskutert blant økonomer, men har hittil ikke fått empirisk støtte, og er stort sett tilbakevist som ideologisk betinget teoretisering.

[19] USA hadde betydelig økonomsik vekst fra 1983 til 1989, men spareratene falt fra 7.8% i perioden 1973-1980 til 6.9% i 1986. Offentlige underskudd økte som prosent av GDP fra 26.1% i 1979 til 41.2% i 1986. En mer grunnleggende kritikk av "supply-side economics" er gitt i Paul R. Kugmans bok, "Peddling Prosperity" (Krugman, 1994).

Denne kortfattete historiske oversikten indikerer at evalueringer som administrativ praksis i næringspolitikken har bestemte idéhistoriske forankringer. Det kan være diskutabelt hvorvidt den tiltakende bruken av evalueringer i næringspolitikken innebærer fornyet tillitsvotum for samfunnsforskningen eller om utviklingen har andre forklaringer.

7 EVALUERINGENS RETORIKK

7.1 Effektbegrepets plass i den politiske diskurs

Den politiske diskursen[20] rundt det enkelte program eller tiltak finner vi gjerne oppdelt i flere faser, men med mest oppmerksomhet knyttet til oppstart og tilbakeblikk. Ved oppstart av nye programmer er det gjerne de positive forventninger som gir diskursen en avgjørende betydning for budsjettdebatter. I den omtale av program/tiltak som kommer i etterkant av gjennomføring, er det gjerne hele den institusjonen som står ansvarlig for det enkelte prosjekt som rammes av diskursens budsjettkonsekvenser. Begge deler fortjener omtale, men i denne omgang vil jeg konsentrere meg om sammenhengen mellom evalueringens retorikk og hvilke konsekvenser denne har for valg av evalueringsmetodikk.

Det er en tese her at den diskurs som kommer i etterkant av gjennomføring av program/tiltak, *tvinger frem effektanalyser*, ettersom effekter av tiltak blir det som styrer diskursen, uavhengig om det finnes grunnlag for konklusjoner om effekter eller ikke. - Om ikke metodisk pålitelige effektanalyser blir gjennomført, blir debatten om effekter likevel gjennomført, men på falske premisser. Om en påberoper seg effekter av et tiltak uten at det er gjennomført analyser som kan dokumentere slike påstander, kan dette åpenbart ha uheldige konsekvenser.

---

[20] Diskurs kan defineres som *"alla slags användande av språk i muntliga och skriftliga sociala sammanhang, dvs utsagor och skriftliga dokument. En diskurs är en social text."* (Alvesson & Skoldberg, 1994:281).

En del trekk ved programevaluering som gjelder kontekst og rutiner i handlingsverdenen, har klare konsekvenser for språkbruk. Retorikken i evalueringsdiskursen preges av to distinkt forskjellige formidlingsmåter: I den tradisjonelle evalueringsrapporten som leses av de få, finner vi den apologetiske formen som utrykker avstand og forsiktighet. Ofte er reservasjonene flere enn konklusjonene. I den videre formidlingen av budskapet fra evalueringsrapporten, sekundærmateriale som stortingsmeldinger og andre viktige dokumenter som ligger nærmere den diskursen som har betydning for politiske vedtak, er språkbruken en annen. Forskernes apologetiske stil er byttet ut med en retorikk som utrykker sterkere påstander. Diskursen starter altså med forsiktige påstander, uklare posisjoner og få deltakere, og endrer karakter mot å involvere flere deltakere med sterkere befestede posisjoner og klarere konklusjoner.

Det er enkelt å finne eksempel på uheldige konsekvenser av denne retorikkens sekvenser. I en mye distribuert publikasjon fra EU-kommisjonen heter det at:

"*Reviews of past projects five year after programme completion report that*

- *43% of participating SMEs had increased their turnover,*
- *53% had accessed new markets and*
- *42% had created new jobs*" [21]

Dette virker som opplysninger som forteller at en har lyktes med de prosjektene som omtales. En fester lett tillit til påstander fremsatt i en slik seriøs publikasjon, og en tror gjerne at denne form for presentasjon av fakta må være basert

[21] European Commission, 2000. SME: Taking the opportunity. In Community Research (Ed.), *Innovation & Research - European support for small companies* . European Commission SME Helpdesk: Office for Official Publications of The European Communities, L-2985 Luxembourg

på pålitelige effektanalyser. - Det er dessverre lett å påvise at så ikke kan være tilfelle:

Når en rapporterer at 43% av deltakerne i et SMB- program har økt omsetningen, er dette å forstå slik at *andre*, lignende virksomheter *ikke* har økt omsetningen innen samme 5-års periode?

Når en hevder at 43% av deltakerne har økt omsetningen, er dette en stor økning, eller er økningen et tall like over null, for eksempel bare et positivt, lite tall, for eksempel 0.05%? *Slik påstanden står alene, er den meningsløs.*

Når en hevder at 43% av deltakerne har økt omsetningen, har de andre 57% av deltakerne fått redusert sin omsetning?. I så fall, er denne reduksjonen i omsetning i sum større eller mindre enn summen for de 43% som har fått økt omsetningen?

Tilsvarande resonnement kan en gjøre for de to andre påstandene. Om 53% deltakerne i programmene har nådd frem på nye markeder og 42% av deltakene har skapt nye arbeidsplasser over femårs- perioden, har utviklingen i andre virksomheter stått stille? Er det grunn til å tro at andre virksomheter *ikke* har endret seg over perioden? Hva om tilsvarende bedrifter som *ikke* deltok i programmet har skapt flere arbeidsplasser enn de som deltok?

Det er åpenbart at EU-kommisjonens publikasjon fremsetter meningsløse påstander, påstander som er positive, men meningsløse. Det synes å være et konsistent trekk ved det meste av den videre rapportering av utdrag fra evalueringsrapporter at den som skriver, har lest rapportene på jakt etter det som kan tolkes som *effekter,* og fletter dette inn i ei ny språkdrakt. Dette kan kalles å tolke rapporter i "beste mening". Konsekvensen er likevel gjerne at de som er ansvarlige for næringspolitiske avgjørelser blir ført på villspor. Det kan være vanskelig å se at slike implisitte påstander om effekter som disse fra EU sine

program for små- og mellomstore virksomheter, er uten informasjon og svært vanskelig tolkbare.

Det ligger ingen vond vilje i en slik "implisitt" bruk av formuleringer som leder leseren til å konkludere med at effekter er påviste. Egen evne til logisk tenking kan være like lite utviklet som en håper andres er. Om noen skulle hevde at en ikke har dekning for påstandene, kan en med en viss rett hevde at dette bare er en vanlig uttrykksmåte, en mener ikke faktisk å hevde at disse gode tallene *bare* skyldes de virkemidler som blir omtalt. – Dessuten, slike uttrykksmåter er de eneste gangbare, politiske myndigheter forstår bare dette språket. En *må* snakke om effekter, det er bare det som blir forstått. Alternativer er å uttrykke seg på måter som kan virke programmatisk kritisk og dermed ikke kommuniserbart.

Bruk av begrepet *effekter* om målbare forhold som slett ikke er effekter etter vanlig språkbruk, er ikke spesifikt for EU-systemet, men vanlig utbredd, også i det norske administrative språket.

Problemene med bruk av begrepet effekt er flere. *For det første* er det et sterkt sammenfall mellom betydningen av begrepet i normalspråk og vitenskapelig bruk: I normalspråk blir begrepet effekt oppfattet som ordet som sier at årsak og virkning er kartlagt. Som normert term i vitenskapelig arbeid som er forankret i empiri, er begrepet effekt som regel probabilistisk kvalifisert, en snakker om at effekter i større eller mindre grad er *sannsynlige*. Dette er uvant for folk flest. Den folkelige oppfatningen av effekt et mer lik det en finner innen 1.ordens predikatlogikk eller matematiske[22] utrykk for deterministiske sammenhenger. Har en sagt effekt, har en også pekt på årsak. Slik er den retoriske

[22] Etter Wittgenstein er de fleste enige om at matematiske systemer er lukkede (tautologiske) i den forstand at de bare refererer til seg selv, ikke til den empiriske "ytre" verden.

virkningen av begrepet nokså entydig; leseren blir ledet til å tro at de utfall en regner som effekter stort sett[23] har sitt opphav i det tiltaket eller virkemiddel som evaluerer. *For det andre* fins de mange, ofte implisitt gitte eller bevisst utelatte bindestreksformer som kvalifiserer begrepet effekt. En snakker gjerne om flere typer effekter i undersøkinger:

- *Halo-effekt* – reaksjon på et stimuli smitter over på reaksjonen på andre stimuli
- *Hawtorne effekt* – reaksjoner hos respondenten som skyldes at en vet en blir observert
- *John Henry effekt* – reaksjoner hos respondenten som skyldes at en vet en er kontrollgruppe
- *Placebo effekt* – reaksjoner som skyldes forventning om virkning (uten årsak)

Slike bindestrekskvalifiseringer av begrepet effekt er stort sett kjent av fagfolk, de er *ikke* en del av den folkelige forståelsen. De har likevel en funksjon innenfor den apologetiske retorikken i forskningsrapporter: Effekter kan være så mangt, og vi mente ikke å påstå effekter i *streng betydning*.

[23] Den folkelege varianten av *sannsynlighet* blir gjerne *unøyaktighet*.

7.2 Konsekvenser for evalueringers legitimitet

Evalueringens retorikk har konsekvenser for evalueringens legitimitet. Ettersom det i mange fora som er avgjørende for ressursfordelinger ofte fremsettes påstander som godtas av lekfolk og gjennomskues av fagfolk, må dette få uheldige følger. For den forsker som opplever at fremlagte funn får sitt eget liv og fortolkes langt utover det som det som er rimelig, kan det oppleves som pinlig. Forskeren kan forklare overfor kolleger at dette slett ikke er en riktig tolkning, men han kan vanskelig fri seg fra mistanke om opportunistisk adferd, for eksempel ved at tvetydige formuleringer har tilrettelagt for overtolking i samsvar med etterspørsel etter resultater. Slik bidrar enkelte presentasjoner, som for eksempel de EU-kommisjonen hadde om effekten av sine SMB-program, til å svekke evalueringers omdømme blant forskere. Det er likevel verre at denne type kommunikasjon svekker befolkningens allmenne tiltro til offentlige evalueringer som hensiktsmessige virkemidler for å sikre etterrettelighet i offentlig forvaltning. Når uetterrettelige påstander om effekter av tiltak benyttes for markedsføring av samme tiltak, tildeles evalueringsforskningen omtrent samme status som filmstjerner som bruker Lux, ukjente "leger" som garanterer for effekten av slankekurer, eller som for en tid tilbake, eksperter som går god for den helbredende effekten av askeavkok.

8 EVALUERINGENS METODIKK

8.1 Formative versus summative evalueringer

Vi har valgt å trekke et skiller mellom formative og summative evalueringer som trolig er noe skarpere enn det en finner i andre fremstillinger. Begrunnelsen for dette er at sammenblandinger av disse vesensforskjellige formene for evaluering kan gi uheldige konsekvenser i form av bevisst eller ubevisst overtolking av resultater fra formative evalueringer. Begrepene formative og summative evalueringer refererer likevel ikke entydig til distinkte former evaluering, men viser til evalueringens *formål* . Mens formative evalueringer har til formål å gi støtte og korreksjon til tiltak under oppstart eller underveis, har summative evalueringer ambisjoner om å dokumentere effekter av gjennomførte tiltak.

Det finnes ulike varianter av formativ evaluering. For det første har vi de det som omtales som *"proactive evaluation"* (Owen & Rogers, 1999). Dette er en type evaluering som gjøres ved etableringen av et program. Evaluator skal her fungere som en rådgiver som skal bidra til å designe programmet. Han skal blant annet vurdere i hvilke grad det er behov for et program, hvilke relevant programteorier som finnes, hva som er hensiktsmessig organisering av programmet og hvordan man kan sikre at tiltakene er treffsikre.

Den andre varianten er *"clarificative evaluation"*. I denne evalueringsformen rettes hovedfokuset mot tiltakets rasjonale, eller det som også omtales som programteori. I evalueringen rettes det søkelyset mot hva som er intensjonene for programmet, holdbarheten i det rasjonale programmet baserer seg på og hvilke tiltak som kan gjennomføres for at programmdesignet skal være mest mulig i tråd med programteorien (Owen & Rogers, 1999).

En tredje variant er *"monitoring evaluation"* (Rossi et al., 1999:192) som beskrives som *"…the systematic documentation of key aspects of program performance that are indicative of whether the program is functioning as intended or according to some appropriate standard."*. Slike formative evalueringer vil være opptatt av hvordan programmet eller tiltaket implementeres og organiseres, ulikheter i implementeringen mellom programmer som inngår i samme tiltak, i hvilken grad programpraksisen er i tråd med det som er intensjonene for programmet og hvordan organiseringen og gjennomføringen av programmet kan styrkes. Langt på vei så dreier det seg om å kvalitetsikre det arbeidet som gjøres i de ulike programmene.

På samme måte finnes det flere former for summative evalueringer. Felles for disse er at de har ambisjoner om å oppsummere *effekter* av tiltak. Det avgjørende kriterium for summative evalueringer blir derfor i hvilken grad de er gjennomført på måter som kan sannsynliggjøre at effekter blir dokumentert.

8.2 Kvalitative versus kvantitative metoder

Spørsmålet om kvantitative eller kvantitative metoder som grunnlag for evalueringen skaper som regel større engasjement blant forskere enn blant oppdragsgivere. Sinnsbevegelse fører ikke sjelden til anklager om manglende integritet eller mangel på kunnskap. Litt standardisert kan en si at mens tilhengerne av den kvalitative tilnærming anklager tilhengerne av kvantitative metoder for manglende integritet og manglende vilje til en dypere forståelse av fenomener, så svarer tilhengere av kvantitative metoder den annen leir med anklager om manglende kunnskaper. En av de mer spissformulerte kommentarer til fremveksten av kvalitative metoder innen evalueringsforskninger finner vi i det anerkjente *Annual Review of Sociology* fra 1984: "The need for evaluations that could be carried out by the technically unsophisticated person and that would be timely and useful to program administrators fueled a strong interest in qualitative approaches to evalua-

tion research."(Rossi & Wright, 1984:336). I den videre diskusjonen av forholdet mellom kvalitativ og kvantitativ metodikk, har den naive oppriktighet forrang. Jeg tiltror alle evnen til å misforstå av et godt hjerte, selv når forståelsen faller sammen med egne interesser. Alternativet til en slik fortolking innebærer mistankens hermeneutikk (Gilje, 1987) som setter *motivene* for metodevalg i fokus. I det videre er det metodevalgene ved evalueringer i seg selv som er det sentrale.

8.2.1 Hva er kvalitative metoder og hva er kvantitative metoder

Avstanden mellom teori og metodikk er et problem som anerkjennes av de fleste samfunnsforskere. Forsøkene på redusere denne avstanden har tradisjonelt gitt seg utslag i utvikling av kvantitative metoder for testing av teori. Utviklingen av kvalitativ metodikk kan sees som et forsøk på å nærme seg empirien fra teorisiden, i motsetning til fra metodikksiden. Mens kvantitativ metodikk har et etablert fundament for verifikasjon og generaliseringer basert på statistisk teori, er disse aspektene ved kvalitativ metode mer problematiske.

Det vil være en stor oppgave å gi en dekkende beskrivelse av hva som faller inn under begrepet kvalitative metoder. På sammen måte vil det være svært omfattende å beskrive hva som kan regnes som kvantitative metoder. Grovt sett kan en si at mens statistisk teori utgjør den fundamentale basis for kvantitative metoder, har slike teorier en mer underordnet rolle innenfor kvalitativ metodikk. På sammen måte som kvantitativ metodikk er preget at et stort antall skoler og retninger, har også kvalitative metoder klare inndelinger i subdisipliner med dedikerte tilhengere. Det er også klart at kvalitative metoder står sterkere innenfor fag som sosiologi, psykologi og antropologi enn for eksempel innenfor økonomi og administrasjonsfag, selv om det innenfor sistnevnte publiseres en økende andel artikler basert på kvalitative metoder.

Det foreligger et betydelig antall lærebøker i kvalitativ metodikk, for eksempel (Finch, 1986; Kopala & Suzuki, 1999; Richardson, 1996; Sankar & Gubrium, 1994, Flick, 2002). Enkelte lærebøker har undertitler som indikerer at dette handler om *ny* metodikk (Miles & Huberman, 1984). Dette kan virke litt søkt ettersom "the Chicago tradition" innen sosiologi fra 1920 til 1950 nettopp var preget av kvalitative metoder, i motsetning til datidens andre dominerende sosiologimiljø ved Columbia University som var preget at Paul Lazarsfelds kvantitative metodikk. Det skal heller ikke underslås at premisser om teoriers forrang kan bli mer problematisk ved kvalitative metoder enn ved kvantitative metoder. Mens kvantitative metoder baserer seg på at testbare hypoteser avledes av teori og testes innenfor et statistisk rammeverk, er det vanskelig å unngå at tilnærminger til empirien fra teorisiden modifiserer teorier underveis snarere enn å teste avledede hypoteser. Det er likevel ikke slik at kvantitative metoder unngår problemer med teoriladet empiri. Teorier styrer oppmerksomheten mot enkelte deler av empirien og unngår derfor ikke at oppmerksomheten ledes bort fra andre deler. Slik sett er det trolig riktig å hevde at kvalitative metoder, som har en tendens til å modifisere teorier direkte i observasjonsprosessen, gir større rom for å avdekke trekk ved den empiriske virkelighet som ikke er spesifisert på forhånd av teori. På den annen side har kvantitative teorier sin styrke i evnen til å teste gyldigheten av spesifiserte, teoriavledede hypoteser.

8.2.2 Valg av metode ved formative og summative evalueringer

Ved evalueringer er skillet mellom kvalitative og kvantitative metoder kontroversielt og relevant. Det er en tese her at kvalitative metoder kan være en velegnet strategi for formative evalueringer, mens kvantitative metoder er bedre egnet for summative evalueringer. Den minst egnede strategi for evalueringer

er survey-metodikk der en ikke skiler mellom formative og summative evalueringer og der survey-metodikk benyttes for å begrunne kausale påstander uten at de foreligger et forskningsdesign som kan rettferdiggjøre slike påstander.

Det er flere grunner til at kvalitative metoder kan være en egnet strategi ved formative evalueringer. For det første er krever evalueringer underveis i et prosjekt/tiltak at evaluerer setter seg raskt inn i forhold han som regel ikke har veldefinerte teorier for. For det andre er begrunnelsen for formative evalueringer som regel at evalueringen skal gi grunnlag for støtte og korrigeringer ved implementering og gjennomføring av tiltaket. Det er derfor en forutsetning at evaluerer makter å sette seg inn i hvordan enkeltpersoner og organiserte grupper reagerer på utforming av tiltak. Dette er former for informasjon som best fremskaffes gjennom personlig kontakt. Samtidig er det informasjon som ofte er bedre tolkbare uten tallfesting.

Ved summative evalueringer derimot, setter kravet om dokumenterbare *effekter* grenser for hvilke metoder som kan benyttes. Dette innebærer at summative evalueringer forutsetter spesifikke forskningsdesign og analysemåter som kan begrunne kausale konklusjoner.

8.2.3 ”Context of discovery” og ”context of justification”

Skillet mellom oppdagelse og bekreftelse av vitenskapelige påstander (Reichenbach, 1938) er mye diskutert. I en del litteratur om forskningsmetodikk finner vi igjen skillet i form av distinksjonen mellom ”exploratory” og ”confirmatory analysis”. Blant de mest kjente innvendinger er kritikken av skillet som for skarpt og absolutt og at Reichenbach skiller mellom deskriptive og normative undersøkelser av påstander om kunnskap (Kuhn, 1962). Denne siste påstanden er trolig feilaktig ettersom Reichenbachs skille mellom ”context of discovery” og ”context of justification” er begrunnet i hans beskrivelser av

epistimologiens tre oppgaver; en *beskrivende*, en *kritisk* og en *rådgivende* oppgave. Uavhengig av debatter om hvor skarpt et skille mellom "context of discovery" og "context of justification" kan trekkes, er et slikt skille av åpenbar relevans for evalueringsforskningen.

Underveisevalueringer eller formative evalueringer har trekk av en søkende prosess der en håper å finne trekk ved det tiltak som gjennomføres og måten de gjennomføres på som gjennom læring kan lette gjennomføringen av andre, lignende tiltak. Slik foregår gjerne den teorioppbygging som kan gi grunnlag for senere teoridrevne evalueringer. Det er imidlertid usikker om de karakteristiske trekk en finner ved å se på et enkelt tiltak, også gjenfinnes i neste, lignende tiltak. Gevinsten av læring er avhengig av at det en observerer har en viss *permanens* og *regularitet*. Undersøkelser av permanens og regularitet krever imidlertid en annen undersøkelseslogikk enn ren leting og vurdering ex ante. Skillet mellom det søkende ved formative evalueringer og det bekreftende ved summative evalueringer kan forstås som nært beslektet med Reichenbachs skille mellom "context of discovery" og "context of justification". Parallellen er særlig nærliggende når en ikke bare undersøker hvorvidt et tiltak har gitt det ønskede resultat men også har ambisjoner om å forklare hvilke mekanismer det er som virker og gir det ønskede resultat. For Reichenbach er "context of discovery" og "context of justification" begge integrerte deler av en kunnskapsproduserende prosess der en i første fase oppdager eller ser noe som kan være av verdi, mens en i den andre fase søker å bli sikrere på at det en oppdaget faktisk også kan være av mer generell betydning. På tilsvarende måte kan det være av betydelig verdi at kvalitative metoder benyttes ved evalueringer for å forstå og sette begrep på prosesser og fenomener som observeres, mens en i neste fase søker å verifisere observasjoner ved hjelp av kvantitative metoder. Sett på denne måten, kan kvalitative og kvantitative metoder ideelt sett kombineres.

8.2.4 Verdikonflikter eller kunnskapskonflikter?

Kvalitative metoder har i betydelig grad blitt utviklet som en kritikk av kvantitative metoder. Begrepet *kvalitativ metode* er retorisk potent i den forstand at det automatisk gir assosiasjoner til *kvalitet.* Utviklingskonteksten, kritikken av kvantitative metoder, sammen med dette retoriske elementet, gir det forføreriske resultat at det virker som en stadig diskuterer et *bedre* alternativ til eksisterende etablerte, kvantitative metoder. Den store og økende mengden av innføringsbøker i kvalitativ metode insisterer stort sett på at kvantitativ metodikk har en dominerende posisjon innenfor samfunnsvitenskapelig forskning. Ved en rekke høyere norske læresteder som ikke vektlegger kunnskaper innen kvantitative metodikk, gir slike innføringsbøker lett studenten en følelse av deltakelse i en heroisk kamp mot det ukjente. Dikotomien kvalitativ - kvantitativ metodikk[24] har således gitt en debatt på villspor, en debatt som nettopp gir næring til den variant av "mistankens hemeneutikk" der egne, edle *motiver* og intensjoner fremheves mens andres motiver mistenkeliggjøres. Debatten fremmer ikke nødvendigvis hensikten med forskningen, å komme frem til ny eller sikrere kunnskap.

En håndverker har mange redskaper. Han har gjerne sitt yndlingsverktøy, men det krever ingen dypere erkjennelse å akseptere at flere redskaper gir større fleksibilitet og muligheter for å løse flere oppgaver, mens færre redskaper innsnevrer mulighetene. Forskere minner ofte mer om barn med lærevegring; har man lært å bruke en hammer, skal alt bankes (Coser, 1975). Analogien halter

[24] Det kan også innvendes at idéen om teoriavledede hypoteser som testes mot empiri, vanligvis et viktig poeng for den kvantitative orienterte forsker, synes å være mer fremtredende innen anglo-amerikans forskning en innenfor fransk samfunnsvitenskapelig forskning. Korrespondanseanalyse (Hjelbrekke, 1999) bygger på en langt mer induktiv logikk, der kvantitative teknikker benyttes for å finne strukturer som kan gi støtte i teoriutviklingen underveis. Slik sett har korrespondanseanalyse mer til felles med kvalitative metoder enn med mange kvantitative metoder, selv om det definitivt er kvantitativ metodikk.

selvsagt, ettersom forskerens beherskelse av en metode gjerne krever en betydelig investering av tid og ressurser. Egen læreinvestering er likevel *ikke* er et gyldig argument mot at valg av metode bestemmes av oppgaven som skal løses. Tvert imot må oppgavens karakter være avgjørende for metodevalg. Om forskeren ikke har gjort den læreinvestering som er nødvendig for å løse oppgaven, bør oppdragsgiver overlate oppgaven til den som er kompetent.

For evalueringsforskningen er det viktig at oppdragets og oppgavenes karakter styrer metodevalg, ikke den tilfeldige fordeling av kompetanse blant forskere og forskningsmiljøer.

8.3 Design, kausalitet og effekt

8.3.1 Design og kausalitet

Begrepet kausalitet har lenge vært et kontroversielt tema innen samfunnsvitenskapene. Innen mange forskningsmiljøer har en systematisk søkt å unngå begrepet. En finner mange fantasifulle hjelpebegreper som er innført med det ene formål for øyet; å unngå all tale om årsak og virkning. En finner erstatningsbegrep som ”føringer” og ”innvirkning”, gjerne fulgt av en forsikring om at en slett ikke mener kausalitet ”i streng betydning”. Manglende evne til å tale klart om årsak og virkning kan ha svekket den troverdighet evalueringsforskningen er avhengig av.

Mange forskningsmiljøer har internalisert denne vegringen som en forsikring mot forhastede konklusjoner. Resultatet har stort sett vært en konsekvent underlæring av metoder og teknikker som er hensiktsmessige for analyser av virkninger av offentlige intervensjoner. I enkelte, mer sofistikerte miljøer benyttes Karl Pearsons uttrykte frykt fra 1911 ”*Beyond such discarded fundamentals as ’matter’ and ’force’ lies still another fetish amidst the inscrutable arena of modern science, namely, the category of cause and effect*” (Pearson,

1911) side iv som unnskyldning. I 1979 skriver Donald T. Campbell at " *The epistemology of causation, and of scientific method more generally, is at the present in a productive state of near chaos*" (Cook & Campbell, 1979:10). I 2001, 90 år etter Karl Pearsons formulering ble trykket, fikk Judea Pearl Lakatosprisen[25] fra The London School of Economics and Political Science (LSE) for sitt bidrag til vitenskapsfilosofien for sin bok om kausalitet (Pearl, 2000).

Med sin klarlegging av kausalitetens logigg, bringer boken debatten ut av det kaos Donald T. Campbell mener eksisterte for litt over type år siden, og bringer også klarhet i et annet begrep som er sentralt for evalueringsforskningen, nemlig det kontrafaktiske. Det er i dag alminnelig enighet om at nøkkelen til å forstå effekter av inngrep er kausalitetsbetraktninger som gjelder målinger mot den kontrafaktiske situasjonen, dvs. tilstanden en hadde hatt om inngrepet *ikke* hadde vært gjennomført, (Heckman, 1999, Heckman, Tobias, & Vytlacil, 2000, Mohr, 1995, Grasdal, 2001, Bratberg, Grasdal, & Risa, 2000, Kvitastein & Hungnes, 2001).

Etableringen av denne tenkemåten har er langt forspill. Begrepet *kvasi-eksperiment* fikk almen utbredelse på midten av 60-tallet med Campbell og Stanleys bok om eksperimentelle og ikke-eksperimentelle design (Campbell & Stanley, 1963).

Empirisk evalueringsforskning som baserer seg på å etablere kausale effekter av inngrep i markeder, har tradisjonelt ikke vært et sentralt felt for økonomer. Denne type forskning forutsetter en eksperimentell logikk som gjør det mulig å isolere effekter av inngrepet. Økonometrikere har likevel i lang tid har levert

[25] Lakatorprisen er en pris til minne om vitenskapsfilosofen Imre Lakatos (1922 – 1974) som var professor i logikk med særlig ansvar for matematisk filosofi ved LSE fra 1969. Formann i komiteen som deler ut prisen er professor Anthony Giddens, direktør ved LSE.

betydelige bidrag til modeller (Roy, 1951, Quandt, 1972) som kan anvendes for denne typen problemer. Innenfor annen samfunnsvitenskap er det særlig med boken til Campbell og Stanley fra 1963, "*Experimental and Quasi-Experimental Designs for Research*" som markerer gjennombruddet for bruk av eksperimentets logikk i felt-sammenheng, dvs. utenfor det rene eksperimentets strengt kontrollerbare omgivelser. Utgangspunktet er egenskaper ved randomiserte felteksperimenter. Ved tilfeldig tildeling av eksponering for tiltak, er det mulig å måle effekter av tiltaket, ettersom *randomisering* gjør at tilfeldige forskjeller jevner seg ut og muliggjør sammenligninger mot en tilfeldig trukket, sammenlignbar kontrollgruppe. Dessverre er slik randomisering, for eksempel ved næringspolitiske tiltak, i praksis sjelden politisk eller etisk mulig å gjennomføre. Forståelse av *kvasi-eksperimentet* ble lansert som analysemåte for å studere effekter av gjennomførte sosiale og næringspolitiske tiltak. Tenkemåten var at bruken av ulike panelmodeller kombinert med sjekklister for potensielle trusler mot valide konklusjoner, var en farbar vei. I ettertid har det vist seg at tiltroen både til komplekse panelmodeller og strukturligningsmodeller sitt potensiale for kausale slutninger fra kvasi-eksperimentelle design, har vært betydelig overdrevet (Davis, 1978). Alt i 1975 utrykte imidlertid Campbell en viss bekymring for at hans arbeider kunne bli brukt til å rettferdiggjøre kvasi-eksperimenter der et randomisert eksperiment faktisk hadde vært mulig (Campbell & Boruch, 1975).

I 1979 kom *boken Quasi-Experiments: Design and Analysis Issues for Field Settings* (Cook & Campbell, 1979). Boken ble en suksess og etablerte *design* av undersøkelser som nødvendig kunnskap ved de fleste institusjoner som ga kurser i samfunnsvitenskapelig metode.

Utover sytti og åttitallet vokste evaluering som eget forskningsfelt, både i USA og Europa. *The American Evaluation Association* og *European Evaluation Society* ble etablert. I 1977 ble Donald Campbell tildelt "*The Myrdal Science*

Award" fra *The American Evaluation Association.* I 1993 kom den mye omdiskuterte *Government Performance and Results Act* (GPRA) i USA., "*Gepra*" som den kalles (Radin, 1998) og lovfestet evaluering for en rekke offentlig finansierte tiltak. I de 25 årene evalueringsfeltet har vokst fra å være en særdisiplin innen utdanningsforskningen til å bli en aktivitet som gjelder svært mange politikk-områder, har det trukket til seg forskere fra mange fagfelt. Både filosofer og psykologer, så vel som sosiologer og økonomer har vært innom. Forsøk på profesjonalisering av feltet har bare delvis lykkes, så selv om feltet har etablert en viss egen identitet, er det fortsatt de etablerte fagdisiplinene som er de styrende.

Dette har vært svært heldig for utviklingen av feltet. Det er lite sannsynlig at en slik kraftanstrengelse og intellektuell utfoldelse som tilfellene Heckman og Rubin kunne ha skjedd innenfor evalueringsfaget. Evalueringsforskningens status som ny og ikke veletablert, ville trolig fungert som en effektiv seleksjonsmekanisme. Nobelpriser og den oppmerksomhet en slik tildeling gir, kunne heller ikke funnet sted innenfor evalueringsforskningen om den hadde vært etablert som en egen fagdisiplin.

8.3.2 Eksperimentelle design

Design er en spesiell måte å se et fenomen på, en måte som tillater oss å analysere eller teste årsak og virknings- forhold. I følge Daryl Bock (1975) er hensikten med eksperimentelle design å demonstrere at, ved å manipulere de forhold respondenter reagerer på, kan vi endre deres adferd på en forutsigbar måte. Den grunnleggende idéen bak eksperimentelle design er kontroll. Dess mer kontroll den som utfører eksperimentet har over tilordning av respondenter til forskjellige forhold, dess mer informasjon kan eksperimentet gi oss.

Diskusjoner av kausalitet og design forutsetter en spesiell terminologi som refererer seg til det klassiske eksperimentelle designet. Vi vil benytte denne terminologien, også i de tilfeller de ikke er snakk om rene eksperimenter. Dette betyr at vi utrykker oss i litt "medisinske" termer og snakker om "behandlingsgruppe" eller "testgruppe", som vi ofte gir fotskriften *t* (treatment group) og kontrollgruppe, som vi gir fotskriften *c* (control group). Vi benevner observasjon med bokstaven *O*, og intervensjon med bokstaven *X*. Når vi snakker om personer eller bedrifter som har blitt eksponert for et tiltak eller et program omtaler vi gjerne disse som medlemmer i "behandlingsgruppen" mens personer/bedrifter som benyttes for sammenligninger benevnes som medlemmer i "kontrollgruppen": Det ligger ingen "terapeutisk" mening i dette, begrepene blir benyttet kun fordi de er hensiktsmessige for diskusjoner av de ulike forskningsmetoder.

Det er primært to forhold som skaper behov for å utvikle eksperimentell design: For det første ønsker vi å sikre at de utfall vi observerer, skyldes våre manipulasjoner og ikke en eller annen *spuriøs* faktor. Dette er nødvendig for å kunne gi de utfall vi observerer en kausal fortolking, dvs. at vi kan si at vi kjenner de faktorer som gir de effekter vi observerer. For det andre ønsker vi å sikre oss at de enheter vi har inkludert i eksperimentet ikke varierer systematisk på andre variable enn de vi ønsker å studere. Den beste strategi for å sikre seg mot slike effekter, er randomsering, dvs. tilfeldig tilordning av enheter til behandlingsgruppe og kontrollgruppe. Randomisering sikrer at effekten av variable som ikke er relevante for eksperimentet, jevnes ut og fremstår som "støy", som ikke virker inn på eksperimentet.

Tabell 3 Noen eksempler på eksperimentelle design for felteksperimenter

Design	**Skjematisk**	**Effektsammenligninger**
"One-shot" case study	X O_{1t}	O_{1t} - k
En gruppe pretest/posttest	O_{1t} X O_{2t}	O_{2t} - O_{1t}
Pretest/posttest med ikke- Ekvivalent kontrollgruppe	O_{1c} X O_{2t}	O_{2t} - O_{1c}
Pretest/posttest med kontrollgruppe	O_{1t} X O_{2t} O_{3c} O_{4c}	$(O_{2t} - O_{1t}) - (O_{4t} - O_{3t})$

Tabell 3 viser noen enkle design som er nokså vanlige. Det første, "the one-shot case study", eller det Campbell (Campbell & Stanley, 1969) kaller statisk gruppesammenligning er svært vanlig innenfor evalueringsstudier og har åpenbare svakheter. Her har det vært en eller annen intervensjon (tiltak/program) og en har gjort en undersøkelse i ettertid. Resultater fra undersøkelsen sammenlignes ikke med noen annen gruppe. Designet gir ingen mulighet for kausale konklusjoner. Det andre designet, en gruppe med pretest og posttest er også mye brukt ved evalueringer. Her har en gjort undersøkelser før og etter intervensjon. Intensjonen ved slike undersøkelser er gjerne å finne forskjeller fra før-tilstanden til etter-tilstanden. Problemet er at en ikke undersøker om resten av verden har forandret seg, for eksempel på samme måte som de enheter som inngår i undersøkelsen. Designet kan gi et skinninntrykk av kausalitet, men utelukker faktisk kausale slutninger. Det tredje designet, pretest – posttest med

ikke-ekvivalent kontrollgruppe, er heller ikke uvanlig. Her er bare kontrollgruppen undersøkt før intervensjon, mens testgruppen bare er undersøkt etter intervensjon. Dette skaper åpenbare svakheter. En er utelukket fra pretest i testgruppen og en mangler muligheter for å sammenligne testgruppen mot kontrollgruppen i tilstanden etter intervensjon.

Det finnes mange typer eksperimentelle design, spesialtilpasset fysiske eksperimenter, eksperimenter innen psykologi og andre tilfeller der en søker etter kausale sammenhenger. De få vi har nevnt over, er trukket frem fordi de ofte forekommer i *felteksperimenter*, eksperimenter der det er mennesker, bedrifter eller andre relevante enheter som er gjenstand for undersøkelser. Dette er tilfeller der en har begrensede muligheter for kontroll, situasjoner der randomisering av tilordning til testgruppe og kontrollgruppe er det som gjør eksperimentet mulig.

8.3.3 Ikke-eksperimentelle design – kvasi-eksperimentet

For næringspolitiske tiltak og andre tiltak i offentlig regi, er det i de fleste tilfeller små muligheter for å kunne gjennomføre eksperimentelle design[26]. I de fleste tilfeller står en overfor situasjoner der bestemte grupper/bedrifter/personer deltar i tiltak/programmer av bestemte grunner. Det er ingen tilfeldighet at de deltar, så tilordning til testgruppen innebærer ingen randomisering. For slike situasjoner, der en har ”treatment”, resultatmål og ”eksperimentelle enheter”, men tilfeldig tilordning til testgruppen som mulig-

[26] Det kan likevel være verdt å minne om at det finnes tilfeller da rene eksperimenter faktisk er gjennomførbare. Trolig er kunnskapsgrunnlaget for næringspolitikken så svakt, at flere randomiserte studier burde vært gjennomført, selv om slike tiltak ville støte mot egalitære verdier.

gjør sammenligninger som kan gi kausale konklusjoner kalles gjerne kvasi-eksperimenter (Stouffer, 1950, Campbell, 1957). For slike studier er det sentrale spørsmålet hvilken tiltro en kan feste til slike studier. En skiller gjerne mellom to klasser av kriterier for troverdighet: Den interne validiteten og den eksterne validiteten(Cook & Campbell, 1979). Den interne validiteten gjelder tiltroen til studien i seg selv, herunder tiltroen til den statistiske konklusjonsvaliditeten, dvs. hvilken status påviste, statistiske sammenhenger skal gis. Den eksterne validiteten gjelder i hvilken grad studien gir grunnlag for generaliseringer. Slik forutsetter ekstern validitet intern validitet, mens den eksterne validiteten som skal danne grunnlag for generaliseringer av for eksempel næringspolitiske virkemidlers virkemåte, forutsette at bl.a. de resultatmål som benyttes også kan benyttes i andre studier. Resultatmålene må derfor være repliserbare og stabile over tid og mellom undersøkelser.

Den interne validiteten til en kausal inferens er proposjonal til antall plausible, alternative forklaringer for den samme effekten. Campbell og Stanley beskriver åtte generelle kategorier for alternative forklaringer – eller *trusler mot den interne validiteten* – og viser hvordan enkle designegenskaper kan brukes for å luke ut en del alternative forklaringer. De åtte kategoriene er (1) historie (2) modning (3) testing (4) instrumentering (5) regression (6) mortalitet og (7) seleksjon (8) seleksjon-interaksjon.

1. **Historie** som trussel mot intern validitet innebærer at pre-post-forskjeller ikke skyldes intervensjonen, men historiske sammenfall av hendelser. Et eksempel kan være et tiltak for å øke trafikksikkerheten der antall ulykker *før* tiltaket sammenlignes med antall ulykker etter tiltaket. Anta videre at før- målingen av antall ulykker er gjort om sommeren og etter- målingen er gjort om vinteren. Den nedgang i antall ulykker som registreres som for-

skjellen mellom pre og post er i dette tilfellet ikke nødvendigvis et resultat av tiltaket. Det kan like gjerne skyldes sesongvariasjoner i antall ulykker.

2. **Modning** som trussel mot intern validitet impliserer at pre- post – forskjeller ikke skyldes intervensjonen, men endringer i den enhet som er gjenstand for analyse. Skal vi undersøke effekten av for eksempel et kostholdsprogram gjennomført i skolen, er det ikke tilstrekkelig å måle barnas høyde før og etter tiltaket for deretter å konkludere at barnas økte gjennomsnittshøyde skyldes bedret kosthold. Barna ville trolig ha vokst uavhengig av tiltaket. I andre tilfeller kan det være vanskeligere å se at enheten som undersøkes har gjennomgått endringer, uavhengig av det tiltak som studeres.

3. **Testing** som trussel mot intern validitet sier at pre- post – forskjeller ikke skyldes intervensjonen, men at det er avhengighet mellom score på pretest og score på posttest. Slike "praktiske effekter" er nokså vanlige, for eksempel når barn gis en skriftlig prøve før og etter et tiltak, kan forbedret score på testen direkte skyldes erfaringene fra den første testen. Slik blir resultatene ikke en effekt av tiltaket, men av den første testen som effekter skulle måles ut fra.

4. **Instrumentering** som trussel mot intern validitet innebærer at pre- post – forskjeller ikke skyldes intervensjonen, men en endring i de måleinstrumenter som anvendes. I litteraturen kalles dette "instrumentation decay". Instrumentering er en potensiell trussel mot gyldige konklusjoner når måleinstrumenter på en eller annen måte er endret i perioden mellom pretest og posttest.

5. **Regresjon** er en trussel mot den interne validiteten når pre- post- forskjeller ikke skyldes intervensjonen, men for eksempel abnormalt høye pretest scorer. I eksperimenter som gjelder terapeutiske inngrep er det for eksempel

ikke uvanlig med svært høye verdier i pretest. Høye verdier tolkes som symptomer på behandlingstrengende tilstand og kan være grunnen til at en pasient søker hjelp. Symptomer kan variere i styrke over tid, og det er naturlig at slike høye verdier blir lavere, uavhengig av intervensjon. Reduksjonen i score fra pretest til posttest kan i slike tilfeller feilaktig bli tolket som resultat av intervensjon.

6. **Mortalitet** som trussel mot intern validitet innebærer at pre-postest- forskjeller ikke skyldes intervensjonen, men naturlig nedgang som følge av avgang av analyseenheter. Måler en skolekarakterer i for eksempel siste året på ungdomsskolen og deretter i første år i videregående skole, kan en få en økning i gjennomsnittskarakterer som skyldes at de elevene med dårligste resultater, av ulike grunner har blitt borte. Slike effekter er kjente bl.a. fra studier som viser at gjennomsnittskarakterer i high school i enkelte amerikanske stater har gått opp, når offentlige bevilgninger til skolesystemet har blitt redusert. Slike resultater har i flere tilfeller fått store mediaoppslag og har blitt tolket som elevene skjerper seg når ressursene blir knappere. Nærmere ettersyn har vist at effekten kun skyldes at de svakeste elevene faller ut slik at gjennomsnittet på karakterene stiger.

7. **Seleksjon** innebærer skjevheter som skyldes at undersøkelsesenheter ikke er tilfeldig tilordnet. Bedrifter som aktivt velger å delta i et program kan være svært ulike bedrifter som velger å *ikke* delta. Slik kan effekten som skyldes at bedriftene er ulike i utgangspunktet, feilaktig bli gjort til effekter av intervensjonen.

8. **Seleksjon-modning-sammenheng** eller seleksjon-interaksjon kan utgjøre en trussel mot den interne validiteten til en undersøkelse ved at for eksempel

vekstsannsynligheten er forskjellig mellom den gruppen som har valgt å være med i et program/tiltak og de som har valgt å avstå. Undersøkelser har vist at bl.a. folk som søker høyere utdanning i seg selv er mer produktive en de som ikke søker slik utdannelse. En tenkt produktivitetsundersøkelse ville derfor lett konkludere med at forskjeller i produktivitet skyldes utdannelse, mens det faktisk er slik at utvelgelsen i forkant av utdannelsen gjør at en ville ha funnet betydelige forskjeller også før disse individene hadde gjennomført noen utdannelse.

For den eksterne validiteten har Campbell en lignende sjekkliste for potensielle trusler (Cook & Campbell, 1979), (Campbell & Stanley, 1969). De seks mest fremtredende er: (1) Interaksjonseffekter av testing (2) interaksjon mellom seleksjon og eksperimentell behandling (3) reaktive effekter av eksperimentell arrangementer (4) Interasjon mellom flere behandlinger (5) irrelevant respons på måleinstrumenter og (6) irrelevant replisering av målerinstrument.

1. **Interaksjonseffekter av testing**, særlig i forhold til pretest, kan skape problemer i forhold til generaliseringer. I en tenkt studie av mobbing på arbeidsplassen er det ikke usannsynlig at oppmerksomheten rundt problemet som skapes av pretesten kan få effekter av tiltak til å bli vanskelig målbare i posttest.

2. **Interaksjon mellom seleksjon og eksperimentell behandling** er ikke usannsynlig i de tilfeller undersøkelsesobjektene får kjennskap til at de er enten i testgruppen eller i kontrollgruppen. Personer/bedrifter i kontrollgruppen kan bli demoralisert (vi er i gode nok til å være i testgruppen, vi skal vise disse forskerne…) til å respondere på måter som svekker eksperimentet. I rene eksperimenter løses dette problemet ofte med såkalte ”blindtester”. I kvasieksperimenter i felt, er en som regel avskåret fra slike løsninger.

3. **Reaktive effekter av eksperimentell arrangementer**, den såkalte *Hawtorne-effekten*, dvs. det at de som er gjenstand for eksperimentet endrer adferd på grunn av at de er oppmerksomme på at de blir undersøkt, og derfor gjerne gjør en ekstra innsats, kan variere etter hvilket opplegg en har for kvasi-eksperimenter. For bedrifter er for eksempel ikke usannsynlig at varierende grad av mediaeksponering ved deltakelse kan svekke generaliserbarhet ved at resultater like godt kan tilskrives eksponeringen som selve det tiltaket en ønsker å måle effekten av.

4. **Interaksjon mellom flere behandlinger** kan være et problem når for eksempel en bedrift deltar i flere programmer/tiltak. I slike tilfeller kan det være vanskelig å skille ut effekten av det ene tiltak en studerer.

5. **Irrelevant respons på måleinstrumenter** kan være et problem, for eksempel når et resultatmål er vel tilpasset for en type bedrifter og deretter anvendes på virksomheter der de er mindre relevante. Ved standardisering av resultatmålinger i bedrifter, kan dette gi spuriøse effekter som kun skyldes at måleinstrumentene er dårlig tilpasset.

6. **Irrelevant replisering av målerinstrument** er beslektet med problemet over (5) og innebærer at ønsket om sammenlignbare undersøkelser kan være en trussel mot ekstern validitet.

8.4 Observasjonsstudier

Sjekklistene fra Donald T. Campbells sjekklister for kvasi-eksperimentet er svært nyttige redskaper for vurderinger av hva en studie er verdt. De er derfor

av stor verdi for oppdragsgiver ved vurderinger av studier der intensjonen er dokumentasjon av effekter. Imidlertid har det skjedd stor utvikling på feltet siden 1979, og det er nå etablert mer pålitelige metoder for beregning av effekter av den type situasjoner Campbell kaller kvasieksperimenter. Terminologien for diskusjoner er fortsatt basert på eksperimentets logikk, med testgruppe og kontrollgruppe. Det nye er at det legges mindre vekt på at kvasieksperimentet egenskaper som eksperiment og mer vekt på at manglende randomisering gjør at eksperimentets muligheter for kausalslutninger ikke er til stede. Det nye i terminologien er at alle ikke-randomiserte eksperimenter betraktes som *observasjonsstudier,* uavhengig om tilgjengelig datamateriale er arkivmateriale eller ulike former for suveyundersøkelser. Fortsatt er randomiserte eksperimenter målestokken for vurderinger av resultater og kravene til pålitelige observasjoner og stabile målemodeller de samme.

Problemene med å trekke kausale slutninger på grunnlag av observasjonsdata er betydelige (Lieberson, 1985). I de siste to tiår er det likevel gjort betydelige fremskritt når det gjelder metoder for å analysere slike data. Utviklingen er drevet frem av statistikere som Paul R. Rosenbaum og Donald B. Rubin (f. eks. (Rosenbaum, 1995, Rubin, 1997) samt av økonometrikere som James J. Heckman og Robert J. Lalonde (f.eks (LaLonde, 1986, Heckman & Smith, 1995) som til sammen har utviklet et konsistent begrepsapparat. Denne utviklingen representerer en forskningsretning som er distinkt forskjellig fra utviklingen innenfor stianalyse og kovariansanalyse (f.eks (Jöreskog & Sörbom, 1993, Bollen, 1989) selv om begge retninger bruker begrepet kausalanalyse.

Det er særlig to kilder til skjevheter som opptar forskere som jobber med observasjonsdata: Resultater fra testgruppen og kontrollgruppen kan være ulike

selv uten intervensjon og påvisbare effekter av intervensjon kan være *forskjellige* for testgruppen og kontrollgruppen (Winship & Morgan, 1999).

Vi vil i det følgende gi en enkel og kortfattet oversikt over noen sentrale problemstillingen ved analyse av observasjonsstudier. Litteraturen er svært omfattende. Vi gir derfor bare noen enkle skisser av hvordan sentrale problemer oppfattes for å kunne relatere dette til tidligere forståelse av kvasieksperimentet.

8.4.1 Seleksjonsproblemet

De fleste som er opptatt av analyser av kvasi-eksperimentet er enige om at nøkkelen til beregninger av effekter av tiltak ligger i konstruksjonen av den kontrafaktiske situsjonen, dvs. den hypotetiske situasjonen en kunne ha sammenlignet mot, gitt at tiltaket ikke hadde blitt gjennomført. Problemet er at den kontrafaktiske, per definisjon, ikke eksisterer. Den må altså *konstrueres* på beste måte. Dette betinger, som et minimum, at vi er i stand til å skille mellom de enheter (for eksempel personer eller bedrifter) som blir eksponert for et tiltak, og de som ikke blir det. Gitt at vi er i stand til å skille, må vi også innse at vi ikke kan observere samme enhet i to tilstander, som eksponert for tiltak og som ikke eksponert for tiltak. Vi er avhengige av å sammenligne en enhet som er eksponert for tiltaket med *en annen* enhet som ikke er eksponert for tiltaket.

La oss anta at en enhet enten kan være i tilstand ”eksponert for tiltak” , kalt tilstand ”1” eller i den ikke-eksponerte tilstand, kalt ”0” og at utfall Y_1 og Y_0 kan tilordnes hver tilstand. Effekten av å være eksponert for tiltaket kunne da bergnes som forskjellen $\Delta=Y_1 - Y_0$. Ettersom vi ikke kan beregne effekten av tiltaket for den enkelte enhet, er vi avhengige av å stole på fordelingen, $F(\Delta)$ av effekter over enheter. Den forventede effekten for en tilfeldig valgt enhet i populasjonen, benevnt $E(\Delta)=E(Y_1 - Y_0)$ refererer derfor bare til forventningsver-

dien eller gjennomsnittet for hele populasjonen, dvs. både de som har vært eksponert for tiltaket og de som ikke har vært eksponert for tiltaket. Dette utrykket, som ofte kalles den naïve estimatoren, kan for mange formål gi nyttig informasjon. For tiltak som tar sikte på å realisere målsettinger for utvalgte enheter, for eksempel bedrifter, er det likevel av større verdi å finne ut hva som skjedde med de som faktisk deltok i tiltaket. Kaller vi deltakere d=1 og ikke-deltakere d=0, kan vi skrive *fordelingen av effekter* for deltakere som $F(\Delta|d=1)$ og effekter for deltakere som $E(\Delta|d=1)=E(Y_1 - Y_0|d=1)$. Problemet nå er at vi ikke kjenner $E(Y_0|d=1)$. Utfallet i deltakergruppen om de *ikke* hadde deltatt må beregnes, og det er ikke helt selvsagt hvordan dette skal gjøres. Vi kan ikke uten videre bruke gjennomsnittlig utfall blant ikke-deltakerne som et tilnærmet mål for $E(Y_0|d=1)$. Dette kan vi se ved å trekke gjennomsnittsutfallet blant ikke-deltakere fra gjennomsnittsutfallet blant deltakerne, $E(Y_1|d=1) - E(Y_0|d=0)$. Dette gir i Heckmans notasjon (Heckman, 1992)

1) $\{E(Y_1|d=1)-E(Y_0|d=1)\}+\{E(Y_0|d=1)-E(Y_0|d=0)\}$

Den *første* delen av utrykket innenfor klammeparantesene er utrykk for gjennomsnittlig effekt av deltakelse for de som deltar, tillegget i den *andre* delen av utrykket (innenfor klammeparanteser) er et utrykk for *seleksjonseffekter*, skjevheter som skyldes at ikke-deltakere er forskjellige fra deltakere i den ikke-deltakende tilstand. Utrykket demonstrerer er rekke problemer: Uttrykket for det kontrafaktiske, utfallet blant deltakerne om de ikke hadde deltatt, $E(Y_0|d=1)$, inngår i både den første og den andre klammeparantesen. I tillegg ser vi at utrykket for den kontrafaktiske må justeres for utfallet i den gruppen som ikke deltar, noe som er mindre problematisk.

Heckman har siden syttitallet levert flere interessante forslag til korreksjoner for seleksjonseffekter. Det mest kjente er trolig hans forslag fra 1979 (Heckman, 1979) der skjevheter som skyldes ikke-tilfeldig seleksjon blir fortolket som et "utelatt variabel"-problem. En tostegs-prosedyre som først benytter en probit-modell for seleksjonsmekanismen og deretter modellerer korrigerte effekter ved hjelp av minste kvadraters metode, er implementert i kjente økonometriske programpakker som Shazam og Limdep. Problemet med denne tilnærmingen er at den bygger på strenge fordelingsforutsetninger, og at problemet med å finne gode instrumentvariable som ikke er relatert til utfallsvariablene, fortsatt er uløst. Av større relevans for oss er hans arbeider om "Matching As An Econometric Evaluation Estimator" der han stiller spørsmålet om det er *mulig å konstruere ikke-eksperimentelle prosedyrer som gir effektestimater og inferenser om virkninger av tiltak som er svært nær det som ville vært resultatet i randomiserte eksperimenter* (Heckman, Ichimura, & Todd, 1997a:605). I disse arbeidene, som klart er inspirert av hans tidligere kollega Donald B. Rubins arbeider, gjennomgår Heckman en rekke forslag til prosedyrer for ikke-eksperimentelle data, og gir en systematisk oversikt over sterke og svake sider ved ulike estimatorer.

8.4.2 Matching som forskningsstrategi

Tidlig på åtitallet kom statistikerne Paul R. Rosenbaum og Donald B. Rubin med en oppsiktsvekkede artikkel i *Biometrika* (Rosenbaum & Rubin, 1983). Problemet som tas opp gjelder bruk av "tilpasnings"-teknikker ved effektanalyser. I observasjonsstudier (Rosenbaum, 1995) bl.a. innen medisinsk statistikk, har en tradisjonelt benyttet tilpasninger basert på data fra sykejournaler for å konstruere par av "like" pasienter. For hver enkelt pasient som har fått en type medisinsk behandling, identifiseres en "tilsvarende pasient" som ikke har fått slik behandling. På denne måten konstrueres en kontrollgruppe. Ett av pro-

blemene med en slik fremgangsmåte er det er vanskelig å finne to like pasienter når en utvider kriteriene for samsvar. Samsvar på enkle variable som kjønn, alder, kroppsvekt og høyde, krever i seg selv tilgang på et betydelig utvalg av sykejournaler. Utvides kriteriene med relevant informasjon om sykehistorie, må utvalget være svært stort. En risikerer likevel fort at to nøyaktig samsvarende pasienter ikke lar seg identifisere.

Rosenbaum og Rubin foreslår at såkalte "*propensity scores*" benyttes for denne type studier. *Propensity scores* er uttrykk for *sannsynlighet* for medlemskap i en gruppe, gitt en rekke variable. Kaller vi gruppen som får medisinsk behandling, d=1 og gruppen som ikke får slik behandling d=0, kan sannsynligheten for å være en pasient som får behandling uttrykkes som $e(\boldsymbol{X})=prob(d=1|\boldsymbol{X})$**,** der X for hver enkelt person kan representere en lang rekke relevante opplysninger, for eksempel fra sykejournaler. Prediksjon av disse sannsynlighetene for medlemskap kan gjøres ved hjelp av en enkel logistisk regresjonsmodell. De predikerte sannsynlighetene blir deretter benyttet som kriterium for grad av samsvar mellom par av pasienter.

Resonnementet som begrunner prosedyren er enkelt: I det enkleste randomiserte forsøk kan en benytte seg av myntkast; krone er behandling, mynt, du går til kontrollgruppen, slik at $e(\boldsymbol{X})=prob(d=1|\boldsymbol{X})= ½$ for hver $\boldsymbol{X}$**.** En slik prosedyre innebærer at individer med ulike mønster av kovariater kan ha samme sannsynlighet for å havne i behandlingsgruppen eller i kontrollgruppen. I observasjonsstudier derimot, er det normale at noen enheter er mer tilbøyelige enn andre til å havne i behandlingsgruppen, m.a.o. $e(\boldsymbol{X}) \neq ½$ for enkelte enheter, men mønsteret av kovariater kan ofte gi gode prediksjoner på hvilken gruppe en enhet vil havne i. Sett at vi setter sammen to og to enheter som har *samme* sannsynlighet for å havne i behandlingsgruppen. To enheter med samme *propensity score* kan være svært forskjellige med hensyn til $\boldsymbol{X}$, men vektoren av gjennom-

snitt for $\boldsymbol{X}$ vil bli nokså lik både kontrollgruppen og i behandlingsgruppen. Samsvar på *propensity score* balanserer m.a.o. som regel de to gruppene bedre på relevante kovariater enn ved randomisert tilordning. Randomsering tar ikke hensyn til observerte kovariater, men har den egenskap at en også balanserer *uobserverte* kovariater. Prosedyren til Rosenbaum og Rubin muliggjør samsvar mellom behandlingsgruppe og kontrollgruppe basert på en enkelt variabel, nemlig på propensity score, og har egenskaper som reduserer skjevheter i effektanalyser. Prosedyren har ikke helt de samme egenskaper som en ville få ved perfekt randomisert tilordning av behandling, men er det beste substitutt når det ikke foreligger slik tilfeldig tilordning.

Forutsetter en at prosedyren gir som resultat at betinget av $\boldsymbol{X}$, er (Y_1, Y_0) og d uavhengige, dvs.

2) $(Y_1, Y_0) \perp\!\!\!\perp d|\, \boldsymbol{X}$, der "$\perp\!\!\!\perp$" betyr "uavhengig av", innebærer dette at, betinget av $\boldsymbol{X}$, har

utfall i gruppen av ikke-deltakere den samme fordeling som utfall blant deltakere, gitt at de *ikke* hadde fått behandling (deltatt i programmet, blitt eksponert for samme næringspolitiske tiltak). Går vi tilbake til Heckman's notasjon i utrykket **1),** innebærer dette at

3) $E(Y_0|\boldsymbol{X},d=1) = E(Y_0|\boldsymbol{X},d=0) = E(Y_0|\boldsymbol{X})$ dvs. at den manglende

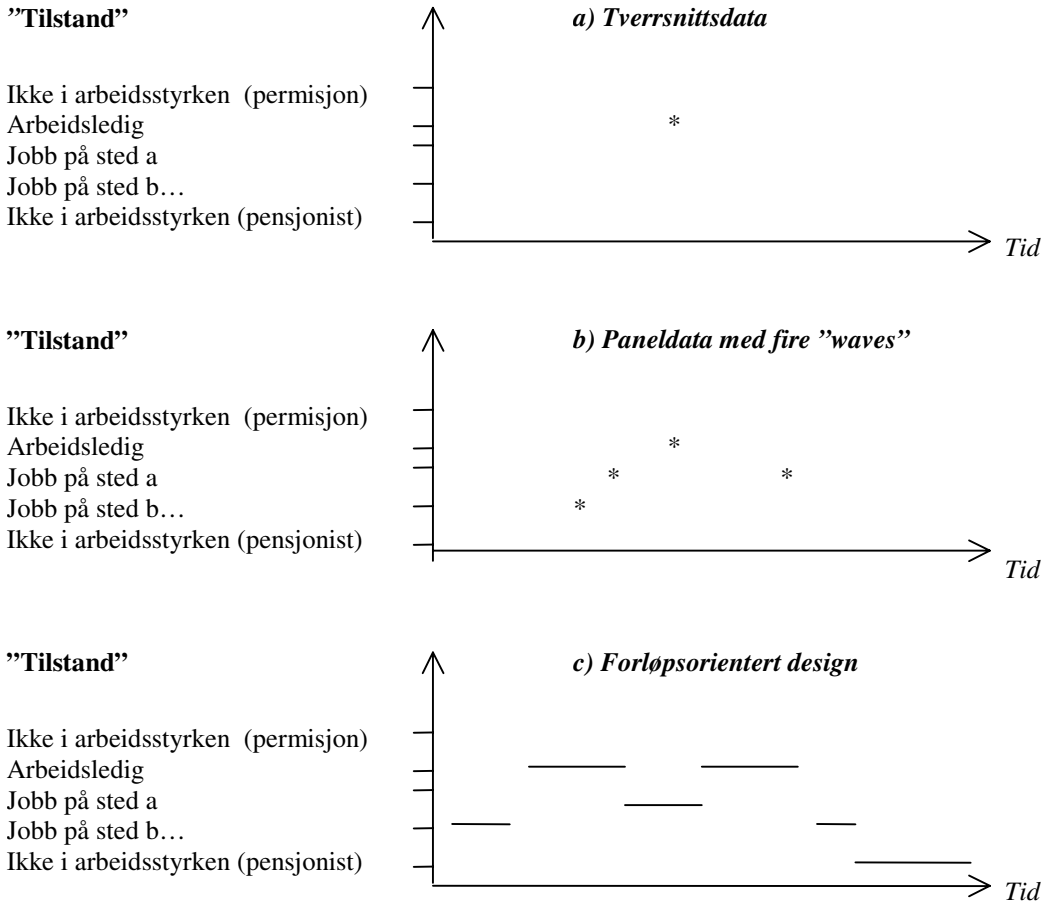
kontrafaktiske kan beregnes ut fra utfallet blant ikke-deltakere. Dette er et svært viktig funn. Det åpner for en mer forsvarlig bruk av dataregistre for effektanalyser gitt at i) *propensity scores* balanserer observerte kovariater, at ii) det er tilstrekkelig å justere for kovariatene i $\boldsymbol{X}$ og dette kan oppnås ved "matching" på *propensity scores,* og at iii) det ikke finnes andre betydelige kilder til skjevheter i effektestimater.

8.5 Analyser av langtidseffekter

8.5.1 Problemer med "timing" av observasjon av effekter

Bruk av spørreskjema og/eller former for intervju er et element som finnes i et betydelig antall norske evalueringer. Det er gjerne formulert som et eksplisitt krav i evalueringsspesifikasjonen at en empirisk undersøkelse skal gjennomføres Det er en erkjennelse at mange (de fleste norske) empirisk baserte evalueringsstudier kan plasseres i to kategorier; svake studier basert på design som ikke gir grunnlag for de konklusjoner som fremsettes og studier basert på design som gir for stor tiltro til de konklusjoner som fremsettes. I den første kategorien faller stort sett alle tverrsnittstudier basert spørreskjema eller former for intervju, i den andre kategorien faller nyere studier basert på kvasi-eksperimentelle design. I tillegg er det et problem at det trekkes for vidtgående konklusjoner av den enkelte studie. Det er en fare for at oppdragsgiver kan fungere som pådriver for uheldige konklusjoner dersom oppdrag spesifiseres på måter som fremskynder rutinisering av forskningsprosedyrer.

Figur 3 viser et enkelt eksempel fra en tenkt studie av arbeidsledighet. En kan tenke seg at en er opptatt av fem ulike tilstander som varsler inn og utgang av arbeidsmarkedet. Problemet med en enkel, for eksempel spørreskjema- basert tverrsnittstudie er illustrert ved at en slik studie bare evner å fange opp *en* tilstand. Om en skal få noe ut av en slik studie så forutsetter fortolkninger at alle tilstander er stabile, dvs. at hver enkelt variabel studien omfatter gjelder en prosess i likevekt slik at statistiske modeller gir et bilde av en slags statisk likevekt. For de fleste sosiale og økonomisk prosesser er slike likevektstilstander lite sannsynlige. I eksempelet under forutsetter mulige konklusjoner at alle endringer mellom tilstander er stabile, dvs. at inn og- utgangsrater mellom de fem ulike tilstandene er tilnærmet konstante over tid.



Figur 3 Observasjonsplaner for datainnsamling

Ser vi f. eks. på endring i antall arbeidsplasser en person har vært innom for å kunne være i arbeid, ser vi fort svakhetene med de tradisjonelle tverrsnittsstudier basert på spørreskjema. Problemet et at en person bare kan observeres i *en* tilstand. Bruk av spørreskjema før og etter intervensjon vil muliggjøre observasjon av *to* tilstander, men gir ingen informasjon om hva som skjer i mellomtiden og ingen informasjon hva som har skjedd med vedkommende rett etter siste observasjon. Vi har med å gjør et observasjonsregime som forutsetter en rekke uspesifiserte likevekter.

Panelstudier er litt bedre, men har likevel den svakhet at de prosesser vi er opptatt av, bare lar seg observere på fire tidspunkter, tidspunkter som er valgt av den som gjennomfører undersøkelsen. Forskeren har ingen over kontroll over hvilke endringer som har skjedd mellom observasjonspunktene. Om er person går fra arbeid til ledighet flere ganger mellom observasjonspunktene men tilfeldigvis observeres i arbeid på de fire tidspunkter panelundersøkelsen omfatter, er dette en kilde til betydelige skjevheter.

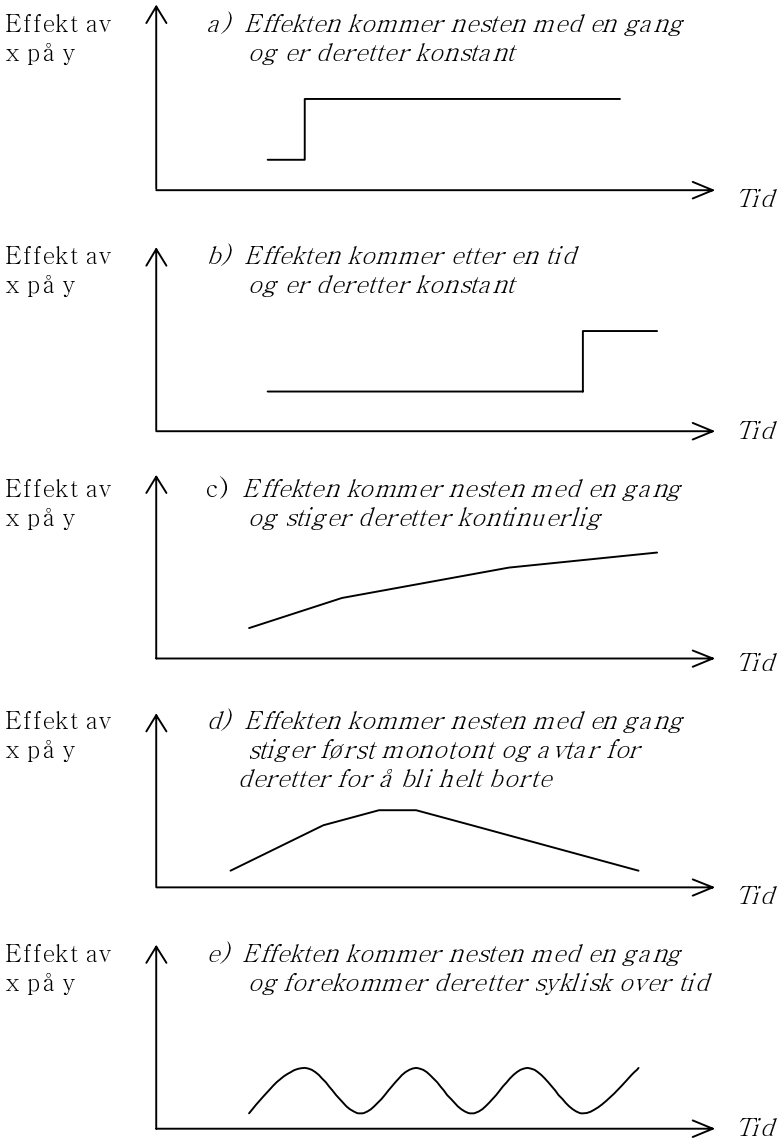
Studier av forløp er åpenbart den beste løsningen. Her observeres tidspunkt for alle tilstandsendringer som har skjedd over observasjonsperioden. I dette tilfellet kan vi faktisk se om arbeidsledige har kommet i arbeid, hvor lenge de har forblitt i den nye jobben osb. Kombinert med design for matching av sammenlignbare enheter kan slike observasjonsplaner gi gode svar. Denne type data kan for eksempel fremskaffes gjennom samarbeid med Arbeidsmarkedsetaten.

Et annet problem som er særlig aktuelt ved vurderinger av effekter, er ”timing”. Som vist i Figur 4 er der ingen grunn til å anta at effekter oppstår når en ønsker å observere dem. Det kan tenkes flere konfigurasjoner av effekter av inngrep. Det er særlig uheldig om det blir gjennomført undersøkelser med sikte på å avdekke effekter rett før (*panel a*) de er observerbare, eller rett etter de var observerbare (*panel b*). Det kan også være uheldig om man er for tidlig ute i situasjoner der effekter tiltar over tid (*panel c*). Effekter som først stiger, og deretter er fallende er også sannsynlige (*panel d*) det samme er sykliske mønster som for eksempel kan skyldes at effekter er sesongbetonte.

Jeg anbefaler derfor at observasjonsmåter som ikke er strengt avhengig av når undersøkelser gjennomføres som det beste for analyser av effekter av tiltak el-

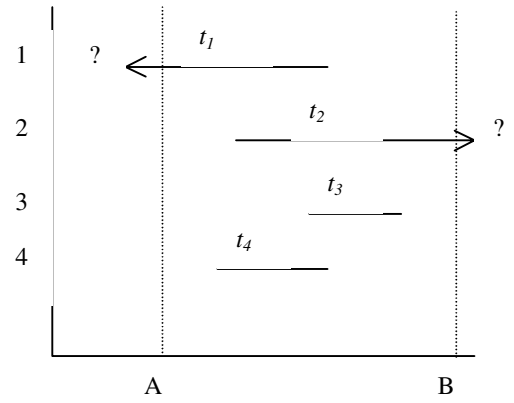
ler programmer. Slike observasjonsplaner sammen med *matchingteknikker* som sammenligner *sammenlignbare* enheter er sannsynligvis beste løsning.

Slike observasjonsplaner er likevel ikke uten problemer. Det mest fremtredende er at de skaper trekk ved data som gjør at vi må benytte andre statistiske modeller enn våre tradisjonelle, lineære regresjonsmodeller. Går vi tilbake til panel *c)* i Figur 3 kan en del av problemene illustreres: Tidsaksen må åpenbart ha en begynnelse og en slutt. *Vi kan ikke observere komplette forløp for alle enheter* uansett tid og ressurser til råde. Folk blir født og dør, nye individer kommer inn i arbeidsstyrken og forlater arbeidsstyrken. Dynamisk betraktet er de fleste interessante fenomener som kan begrunne offentlige inngrep slik: Bedrifter etableres og nedlegges, pasienter blir innlagte og utskrevet. Slike fenomener kan analyseres ved hjelp av forløpsteknikker, men de kan bare observeres i et bestemt tidsintervall, det som skjer utenfor dette tidsintervallet må betraktes som ikke observerbart.



Figur 4. Hypotetisk utvikling av effekter over tid

Problemet er vist i Figur 5. Tenker vi oss et vindu i tid fra A til B der vi observerer for eksempel fire bedrifter eller personer i fire mulige tilstander, vil vi ikke nødvendigvis ha komplette observasjoner for alle enheter.



A: Start på observasjonsperioden
B: Slutt på observasjonsperioden
t_1: Venstresensurert forløp
t_2: Høyresensurert forløp
t_3, t_4: Komplette forløp

Figur 5. Sensurerte og komplette forløp

For den første enheten mangler vi sikker observasjon av startpunkt. Det ligger på et eller annet ukjent sted *før* starten av vår observasjonsperiode. Dette kalles *venstresensurering*. For den andre enheten har vi sikker observasjon for startpunkt, men mangler observasjon av tidspunkt for potensiell tilstandsendring ettersom dette ligger etter avslutningen av vår observasjonsperiode. Dette kalles *høyresensurering*. For enhet 3 og fire kan vi observere på start og slutt for enheten, vi har komplette enheter. I mange tilfeller kan en enhet observeres i mange tilstander. I Figur 3 har vi skisser fem tilstander der en enkelt person kan forventes å bevege seg over alle tilstander unntatt den siste (pensjonist). En kan gå fra permisjon til jobb, fra ledighet til arbeid osv. En tilstand en ikke kan vende tilbake fra, kalles en absorberende tilstand (for eksempel død). Det er også slik at enkelte tilstander forutsetter andre tidligere tilstander. For eksempel kan en ikke komme inn i tilstanden skilt uten å være gift. Det enkleste

modellene, for eksempel en del modeller for konkursprediksjon, har gjerne bare to mulige tilstander; en bedrifter etablerer og finnes i markedet eller den forlater markedet, for eksempel gjennom konkurs. Poenget med de fleste forløpsanalyser er å studere hvilke kovariater som kan påvirke tilstandsendringer. En skiller gjerne mellom tidskonstante og tidsvariate kovariater. Ved analyser av arbeidsledighet vil kjønn være et slikt tidskonstant kovariat. For den enkelte person vil variabelen ha den samme verid uavhengig av når den observeres. Ved konkurspredikskjoner der relevant kovarieter kan være bedriftens nøkkeltall vil disse være forskjellige fra år til annet og må oppfattes som tidsvariate kovariater, altså en enkelt variabel som vil ha forskjellig verdi avhengig av når den observeres. Sensurereing, tidsavhengige kovariater, samt en rekke andre trekk ved observerte forløp, gjør at spesielle analyseverktøy og statistiske teknikker må benyttes.

8.5.2 Forløpsmodeller

Forløpsmodeller eller hasardratemodeller (Petersen, 1991) gir løsninger både for problemene med *sensorering* og problemene med tidsavhengige kovariater. Den sentrale idé er å dele vartgheten t_k inn i flere segmenter som hver dekker tidsintervallet fra en tilstandsendring til den neste. La T være en tilfeldig varierende variabel som står for tiden som har gått i en tilstand før en tilstandsendring eller sensorering skjer og la t være realiseringen av T. Vi kan da spesifisere sannsynligheten for at tilstanden er forlatt i intervallet fra t_j til t_{j+1} , gitt at tilstanden ikke var forlatt før t_j:

$$P(t_j \le T < t_j + \Delta t \,|\, T \ge t_j), \text{ der } \; t_j + \Delta t = t_{j+1} \qquad \textbf{(1)}$$

eller vice versa, sannsynligheten for at tilstanden *ikke* var forlatt i intervallet t_j til t_{j+1} gitt at den ikke var forlatt før t_j..

$$P(T \geq t_{j+1} \mid T \geq t_j) = 1 - P(t_j \leq T < t_{j+1} \mid T \geq t_j), \text{ der } t_{j+1} = t_j + \Delta t \qquad \textbf{(2)}$$

Ved hjelp av regnereglene for sannsynlighet kan vi nå utlede den grunnleggende ligningen for forståelsen av *overlevelsesraten*, nemlig:

$$P(T \geq t_k) = \prod_{j=0}^{k-1} P(T \geq t_{j+1} \mid T \geq t_j), \text{ der } t_0 = 0 \text{ and } t_{j+1} = t_j + \Delta t \qquad \textbf{(3)}$$

dvs. produktet gir sannsynligheten for å "overleve" etter periode t_4, gitt at en har "overlevd" utover t_3 og slik videre over periodene.

På samme måte kan en utlede den grunnleggende ligningen for *hasardraten*, dvs sannsynligheten for at tilstanden var forlatt mellom t_1 og t_2, t_2 og t_3 og så videre:

$$\begin{aligned} &P(t_k \leq T < t_k + \Delta t) = P(T \geq t_k) \bullet P(t_k \leq T < t_k + \Delta t \mid T \geq t_k) \\ &= \prod_{j=0}^{k-1} P(T \geq t_{j+1} \mid T \geq t_j) \bullet P(t_k \leq T < t_k + \Delta t \mid T \geq t_{k-1}) \end{aligned} \qquad \textbf{(4)}$$

Strengt tatt, så gjelder formuleringene i **(3)** og **(4)** bare for diskrete tidsformuleringer, dvs. situasjoner der tidsintervallene er observert i diskrete intervaller som gir en naturlig inndeling. For kontinuerlig tid er hasardraten definert som:

$$\lambda(t_j) = \lim_{\Delta t \downarrow 0} P(t_j \leq T < t_j + \Delta t \mid T \geq t_j) / \Delta t \qquad \textbf{(5)}$$

Dette er et utrykk for en betinget tetthetsfunksjon og angir muligheten for at tilstanden er forlatt ved varighet t_j gitt at den ikke var forlatt før t_j.

Dette rammeverket gir muligheter for modellere forløp og beregne hvordan ulike kovariater virker inn på hasardrater. Det finnes et stort utvalg av mulige måter for introduksjon av kovariater som tar hensyn til særtrekk ved de proses-

ser som modelleres. Disse mulighetene gjør forløpsteknikker særlig fleksible for estimering av parametre for komplekse prosesser. Slike modeller har fortolkninger som er lignende de vi har i vanlige regresjonsmodeller. Mulighetene for realistiske prosessbeskrivelser er likevel mye større enn i tradisjonelle lineære modeller.

Kombinert med matchingteknikker som gjør det mulig å emulere tilnærmede eksperimenter, gir forløpsteknikker oss en mulighet til å identifisere kausale sammenhenger når de er observerbare. For analyser av langtidseffekter er dette den minst spekulative analysemåte og også den analysemåte som forutsetter færrest ikke-testbare forutsetninger.

9 DESIGN AV PROGRAMMER OG TILTAK

9.1 Utredningsinstruksen og tilrettelegging for evalueringer

Utredningsinstruksen[27] har som formål å sikre god forberedelse av og styring med offentlige reformer, regelendringer og andre tiltak. Instruksens pkt 2.1 sier at "Hver sak skal inneholde en konsekvensutredning som skal bestå av analyse og vurdering av antatte vesentlige konsekvenser av den beslutning som foreslås truffet". Videre i samme punkt heter det at "Konsekvensutredningene skal omfatte konsekvenser for statlig, fylkeskommunal og kommunal forvaltning og for private, herunder næringsvirksomhet og enkeltpersoner". Dette er ambisjoner som for de fleste tiltak sitt vedkommende er vanskelig å oppfylle. I instruksens pkt 2.3.1. heter det at "Det skal i nødvendig utstrekning inngå grundige og realistiske samfunnsøkonomiske analyser".

Dette er ideelle fordringer som krever gjennomtenkte vurderinger ex ante. For evalueringer har utredningsinstruksen relevans på to måter. For det første gjør utredningsinstruksens krav om konsekvensutredninger at målsettinger må spesifiseres *før* iverksetting av tiltak. Dette gir en mulighet for det i ettertiden kan foreligge skriftlig materiale som angir hvilke mål de enkelte tiltak søkte å realisere. For evaluerere gir dette en mulighet for bedre spesifikasjoner av *hvilke* effekter av tiltaket som bør prioriteres ved analyser. For det andre kan utredningsinstruksen begrunne innhenting av data som informasjonsgrunnlag for initialsituasjonen. Om dette kunne gjøres i større omfang enn i dag, ville det være er verdifullt grunnlag for gjennomføring av evalueringer. – Den grunntanke som ligger i utredningsinstruksen ligger nær det som er den bærende ide

[27] Instruks om utredning av konsekvenser, foreleggelse og høring ved arbeidet med offentlige utredninger, forskrifter, proposjoner og meldinger til Stortinget. Vedlegg til Kgl. Res. Av 18. februar 2000 (Sak nr. 99/2881)

for effektevalueringer, men med den forskjell at en i evalueringen ønsker å kartlegge i hvilken grad antagelser om konsekvenser faktisk viste seg å være tilnærmet riktige, eller feilslåtte. Det er nærliggende å hevde at ex ante vurderinger, og dermed utredningsinstruksen, mister mye av sin verdi dersom det ikke også gjøres vurderinger ex post.

9.2 Praktiske løsninger for evalueringsstudier

De foreslåtte løsninger for effektstudier tilsier at det ved analyser av langtidseffekter legges større vekt på bruk av tilgjengelige arkivdata (Kvitastein & Hungnes, 2001). SSB har ansvar for at arkivdata blir innsamlet og tilrettelagt i samsvar med offentlige behov. Design av programmer og tiltak bør gjennomtenkes med sikte å at en alt fra starten legger til rette fremtidig etterprøving av det som er gjennomført. I de få tilfeller det finnes muligheter for randomiserte eksperimenter, bør dette prioriteres. Slike tilfeller er unike muligheter for å øke kunnskapen om effekter av inngrep. I standardtilfellet, når muligheter for randomisering ikke foreligger, vil det være av stor verdi om det gjennomføres initiale matchingprosedyrer slik at en når en har bestemt hvilke enheter (personer, bedrifter) som skal inkluderes i tiltaket også har sikret identifisert en potensiell kontrollgruppe.

9.3 Programutforming og analysemuligheter

Programutformingen setter ofte klare grenser for analysemuligheter. I svært mange tilfeller har manglende forarbeid avskåret en fra innsikt ex post. Konstruksjon av informasjon om initialsituasjonen er i de fleste tilfeller et mindreverdig alternativ til innhenting av data før tiltak iverksettes. Tabell 3 (pkt. 8.3.2.) gir en oversikt over hvilke design som gir hvilke begrensinger. Det er nokså åpenbart at studier av effekter av tiltak uten noen form for informasjon om initialsituasjonen (”The one-shot case study”) må bli av begrenset verdi.

Dessverre må vi gå ut fra at dette er situasjonen for et flertall av evalueringer av offentlige inngrep. I de fleste tilfeller er selvfølgelig konstruksjon av eksperimentelle situasjoner ikke mulig og en må betrakte inngrepet som et kvasieksperiment. Som vist i Tabell 3 er en betraktning av et inngrep som et kvasieksperiment avhengig av det finnes noe å sammenligne mot, dvs en eller annen form for kontrollgruppe. I mange tilfeller er det mulig å legge til rette for målinger av initialsituasjonen utover de personer eller virksomheter som måtte berøres av tiltaket. Dette gjøres i svært få tilfeller. Det er trolig en betydelig gevinst å hente i at en ved programutforming har i mente at en også senere ønsker å finne ut om tiltaket virket etter hensikten. I det fleste tilfeller virker det som om en ønsker at dette på en eller annen måte skal kunne skje, uten at en i utgangspunktet har sørget for at den nødvendige informasjon eksisterer.

10 KRAV TIL EVALUERINGSMILJØ

10.1 Krav til evaluerere

Evalueringer av næringspolitiske tiltak har økt i omfang det siste tiår og utgjør i dag en betydelig andel av omsetningen for en rekke forskningsinstitusjoner og konsulentselskaper. Departementer, Fylkeskommuner og institusjoner som Statens nærings- og distriktsutbyggingsfond (SND) er store oppdragsgivere. Det finnes pr. i dag ingen lovpålagt plikt til å gjennomføre evalueringer.

Den norske modellen tilsier anbudsrunder, der evaluerere, i all hovedsak forskningsinstitusjoner og konsulentselskap, konkurrerer om oppdrag, stort sett på andre forhold enn pris, fortrinnsvis oppdragsgivers oppfatning av *kvalitet* i en eller annen form. Kvalitet uttrykkes skriftlig i evaluerers tilbudsdokument i form av erklært eller formell kompetanse, forståelse av oppdragsgivers formulering av problemstillinger, beskrivelse av forskningsdesign og forslag til løsninger av undersøkelsesproblemet. Oppdragsgiver, som gjerne er nært knyttet til tiltaket som skal evalueres, gjør sine valg av evaluerer, overvåker oppdraget og offentliggjør resultatene.

Det stilles i dag ingen spesielle krav til de miljøer som gjennomfører evalueringer. Det mest spesifikke en finner i mange anbudstekster er krav om at det skal gjennomføres en nytte-kostnadsanalyse. Enkelte anbudstekster sier også at det skal gjennomføres en survey. I slike tilfeller må en gå ut fra at oppdragsgiver tar for gitt at anbud bare besvares dersom utførende institusjon besitter den nødvendige kompetanse.

Troen på at anbudskonkurranse automatisk gir det beste valg av evaluerer er litt naivt, men det finnes få kjente, troverdige, alternative løsninger. Troen på at oppdragsgiver er i stand til å velge mellom innkomne anbud på en måte som sikrer valg av kompetente evaluerere er også naiv, men her finnes det kjente

alternativer. Det mest åpenbare er erkjennelsen av at evalueringer krever særlige kunnskaper som er spesifikke for denne type forskning. For de fleste typer relevante utdanninger blir det i dag bare i beskjeden grad tilbudt relevant undervisning på høyere nivåer. Det er et åpenbart behov for oppdatering av evalueringskompetanse ved de fleste av våre høyere læresteder. Samfunnets nytte av evalueringer er åpenbart avhengig av at noe gjøres.

En økende etterspørsel etter kandidater på doktorgradsnivå med fordypning innen evalueringsforskning har fått flere universiteter, blant annet Western Michigan University, til å vurdere å opprette interdisiplinære Ph.D. program (Stufflebaum, 2001). Et slikt tilbud fantes på 80-tallet ved Stanford University. The Evaluation Consortium ved Stanford[28] under ledelse av professor Lee J. Cronbach er nedlagt og for tiden eksisterer det ikke noe slikt interdisplinært tilbud i USA. De tilbud som finnes, bl.a. ved Claremont Graduate University, Ohio State, Illinois, Virginia, UCLA, Utah State, Minnesota, Syracuse og Western Michigan University er alle begrenset til en enkelt disiplin, som regel "education". I de fleste tilfeller går ikke utdanningen utover master-nivå.

I Europa er det svært mange miljøer som er aktive innen evalueringsforskning uten at det finnes organiserte utdanningstilbud som reflekterer dette. Det finnes forskjeller mellom USA og Europa som kan begrunne dette. Blant annet ser det ut som om en i USA legger større vekt på systematisk metodikk for at noe skal kunne aksepteres som evaluering, mens en i Europa har et mer anarkistisk forhold til hva som er akseptabel metodikk. Som antydet i kapittel 7 er en tydelig metodikk særlig viktig i evalueringer som gjør krav på å ha påvist effekter av gjennomførte programmer.

[28] En beskrivelse av grunnlaget for programmet finnes i (Cronbach, 1980a).

I Norge gjennomføres det mange evalueringer som hevder å kunne påvise effekter av gjennomførte programmer og tiltak, uten at det alltid er like klart om analyser er gjennomført på en måte som kan underbygge slike påstander.

Med den økning som har funnet sted i alle typer evalueringer de siste år, kan Stufflebaums tanker om interdisplinære tilbud på doktorgradsnivå være aktuelt også for Norge. Organiserte utdanningstilbud, for eksempel i regi av Norgesnettet, kan være en vei å gå. Ved våre universitet og vitenskapelige høyskoler ligger forholdene vel til rette for tverrfaglige kurs som kan gjøre doktorgradskandidatene bedre i stand til å lede og gjennomføre evalueringer.

Den eneste studie jeg kjenner til som er gjennomført blant aktive evaluerere (King, Stevahn, Ghere, & Minneka, 2001) rangeres 1) evnen til å definere forskningsspørsmålet, 2) kunnskaper om forskningsdesign, 3) kunnskaper om måleteori og 4) forskningsmetodikk som de fire viktigste når det gjelder essensielle kunnskaper for evaluerere. Innenfor flere av de utdanningsmiljøer som er aktive innen norsk evalueringsforskning gis det lite eller ingen undervisning innenfor flere av disse feltene. Undersøkelsen indikerer også at manualer av typen *Guiding Principles of the American Evaluation Association* (AEA) har liten kvalitetssikrende effekt ettersom det ikke er mulig å avlede spesifikke kunnskapskrav fra slike generelle retningslinjer. I Norge har nettopp en del av debatten dreiet seg om bruk av slike manualer som kvalitetssikringstiltak.

For oppdragsgivere må det være lov å kreve dokumentasjon for at enkeltforskere og forskningsmiljøer har den nødvendige kunnskap for å ta på seg oppdrag. Et enkelt tiltak er nettopp interdisiplinære kurs på doktorgradsnivå som gir en felles forståelse av evalueringsforskningens metodiske krav, på tvers av disiplinenes grenser.

10.2 Kravspesifikasjon

Det er en klar sammenheng mellom kravspesifikasjoner forstått som spesifikasjoner i et anbudsdokument og forskernes faglige kompetanse. Ofte går det bra, forskerne avstår fra å ta på seg oppdrag de ikke kan gjennomføre. Andre ganger går det nokså galt, særlig når oppdragsgiver ber om noe som under alle omstendigheter ikke er gjennomførbart. En vanlig umulighet er forespørsler om å måle effekter av komplekse tiltak som nettopp er iverksatt. En annen umulighet er forespørsler om å gi svar på omfattende spørsmål på i løpet av svært kort tid til en lav pris. Det er ikke enkelt å gi svar på hva som er det rette forholdet mellom oppdragets omfang og varighet og oppdragets pris. Det er likevel svært tvilsom at priskonkurranse gir bedre kvalitet (Sørgard, 1990).

Det åpenbare problemet er informasjonsassymetri eller forholdet mellom kvalitet og usikkerhet (Akerlof, 1970). For oppdragsgiver er det problematisk å vite hva som er god kvalitet, særlig når forskersiden definerer hva som er kvalitet. Det er også vanskelig å vite hvilke spørsmål som kan besvares av forskerne og når de gjør sitt beste for å gi gode svar.

Det norske markedet for forskningstjenester er ikke stort og velfungerende, men preget av noen få større miljøer. For å øke kvaliteten på oppdragene og dermed øke sannsynligheten for å øke kvaliteten på evalueringer er det nødvendig med bedre informasjonsutveksling. Det er liten tvil om at korte, mindre oppdrag med omfattende problemstillinger undergraver forskernes respekt for oppdragsgiverne og, vice versa, at de evalueringer som produseres på slike premisser reduserer oppdragsgivernes respekt for forskerne.

Det er et klart behov for mer utveksling av informasjon for å redusere oppdragsgiveres usikkerhet med hensyn til hva de betaler for. Samtidig er det også nødvendig at forskerne signaliserer klarere hvilke oppgaver de kan løse.

Det er imidlertid ikke særlig heldig om informsjonsutvekslingen får preg av markedsføring fra forskningsinstitusjonenes side. I et anbudssystem som skal konkurrere på kvalitet i tillegg til pris, er dette imidlertid vanskelig å unngå. Den enkleste vei er trolig en større satsing på ferdighetsutvikling blant offentlige tjenestemenn/kvinner som gjør dem bedre i stand til å vurdere den forskning som tilbys.

Det må likevel først bli større rom for å kunne diskutere hva som ligger i et løst formulert evalueringsanbud og hvordan en skal forholde seg til er velspesifisert oppdrag med utilstrekkelige ressurser. En oppdragsgiver som oppfatter seg selv som "krevende kunde" er ikke garantert kvalitet. Når mange institusjoner lar være å svare på et anbud er det ikke nødvendigvis slik at de som svarer har de beste forutsetninger for å gjennomføre oppdraget.

Det er et åpent spørsmål om ikke mange av det mest heseblesende evalueringer som gjennomføres bør gjennomføres. Det er betydelig oppgave å sikre at offentlige evalueringer er basert på spesifikasjoner som gir realistiske og gjennomførbare kontrakter.

11 OPPSUMMERING

Hvilke implikasjoner har så de argumenter som er fremført i denne rapporten? Den kan ikke underslås at jeg stiller meg noe kritiske til gjeldende evalueringspraksis. Det positive i budskapet er at det er et betydelig rom for forbedringer. Skillet mellom evalueringer som implementeringsstøtte og evalueringer som dokumentasjon av resultater (kapittel 2) kan virke provoserende for de som mener et slikt skille er av mindre betydning. Ser vi dette skillet i sammenheng med den praktisk politiske bruk av evalueringer, dvs. evalueringens retorikk, (kapittel 7) kan flere bli enige om at det beste er samsvar mellom fremsatte påstander og forskningsbaserte begrunnelser. Konsekvent mangel på slikt samsvar vil på sikt gi det paradoksale resultat et evalueringer undergraver den legitimitet som er selve fundamentet for bruk av evalueringer. Dess lavere tiltro til samfunnsforskningen, dess mindre verdi har evalueringer.

De mange skoler og retninger som har vokst frem internasjonalt (kapittel 4) kan gi ny innsikt og føre til refleksjon over hva en egentlig vil med evalueringer. På den annen side blir forskere gjerne tilhengere av den skole eller retning som stiller minst krav om de kunnskaper de mangler. Dette er som regel ikke en bevisst prosess, men kan for eksempel arte seg som manglende vilje til oppdatering. En avviser nye teorier og metoder og anerkjenner helst det en gjenkjenner. Oppdragsgivere kan også bli fristet til å velge skoler og retninger som benytter mindre kostnadskrevende metodikk. Enkelte tilnærminger har trekk som fører forskerrollen nærmere ekspertrollen. Ekspertrollen er basert på tiltro til et individ og er som regel ikke etterprøvbar. Over tid kan dette gi ekspertbaserte evalueringer. Slike evalueringer vil ha andre kilder til legitimitet enn forskningsmetodikk og vil bryte med kravet til etterrettelighet. Utviklingen av

skoler og retninger gir klare signaler om i hvilke retninger evalueringspraksis *kan* utvikle seg.

Agency -perspektivene (kapittel 5) er en del av kjernen i New Public Management og gir retning til de nye organisasjonsprinsipper som litt lettvint kan kalles *frihet under kontroll*. Den enkelte offentlige tjenestemann/kvinne skal ha mye selvstendighet og ansvar, men må samtidig underordne seg nye rutiner for kontroll av resultatert. Det problematiske er at den som blir kontrollert i en organisasjonskontekst opplever det som manglende tillit. Den som ikke blir gitt tillit tenderer til å oppføre seg som om han/hun ikke var verdig tillit (pkt 5.3). Slik gir kontrollaktiviteten lett selvoppfyllende profetier som skaper sin egen begrunnelse. Det kan virke som om de historiske pregninger en finner i NPM (kapittel 6) gir en viss forklaring for dette tilsynelatende paradoks: En ønsker trekke ressurser ut av den enkelte tjenestemann/kvinners individualitet ved å gi større frihet og ansvar samtidig som den historiske periodens statsfrykt er styrende for kontrollbehovene. Sammenblandingen av *performance measurement* og *evaluation* en av konsekvensene innveving av evalueringspraksis i idéene fra NPM. Resultatet blir gjerne karikerte evalueringer som vurderer individer i stedet for saksforhold.

De metodiske argumenter som er fremført (kapittel 8) kan mobilisere mange motargumenter. Noen kan mene at metodevalg er forskerens ansvar og antydninger om noen valg er bedre endre, støter mot prinsipper om forskningens frihet. Andre kan mene at jeg har for stor tiltro til probabilistiske argumenter og viser for liten respekt for de innsikter andre tenkemåter kan gi. Mine kommentarer til kvalitative (pkt. 8.2) metoder indikerer at jeg har respekt for flere tilnærminger. Kausale resonnementer, som mange forskere mener er reservert for

kvantitative teknikker, har sin plass innenfor kvalitative metoder (Mohr, 1999) . Svært mange innsikter som *ikke* kan gripes med kvantitative teknikker kan fanges med mer åpne metoder. Det er likevel noen problemstillinger som bedre lar seg gripe i probilistiske modeller. De mest opplagte er de som gjelder påstander om permanens og generalitet. De mest oversette problemer innenfor kvalitativ metodikk er de som gjelder kvalifisering av påstanders sikkerhet. Innenfor klassisk hypotesetesting bruker en begrepene type I feil og type II feil om henholdsvis sannsynligheten for forkaste nullhypotesen når den er sann og sannsynligheten for å akseptere nullhypotesen når den er usann. Når en velger et α-nivå som gir sannsynligheten for type I feil, velger en også sannsynligheten for type II feil. Styrkefunksjoner angir forholdet mellom type I feil og type II feil, gitt utvalgsstørrelsen*, men det er substansielle grunner som bør avgjøre valg av α-nivå.* Det er vel kjent at mange som sverger til de kvantitative skoler velger seg et α-nivå på .05 (95% signifikansnivå) og fortsetter, ikke alltid vel vitende om at de da mener at det er greit å gjøre en type I feil i 1 av 20 tilfeller, dvs de tar et valg om å akseptere en påstand som riktig dersom den er korrekt i 19 av 20 tilfeller. For testing av hvorvidt mennesker er HIV-positive eller ikke er det ikke sikkert at dette er tilfredstillende. En ville kanskje fortrekke type II feil ut fra det resonnement at det er bedre om en som ikke er HIV-positiv tester postivt og følgelig må gjennomgå nye tester, enn at en som faktisk er HIV-positiv får en bekreftelse på å være smittefri. Den mer kvalitative varianten av slike grublerier er spørsmål om hvorvidt det er bedre å dømme en uskyldig en å la en skyldig gå fri. Verken kvantitativt orienterte eller kvalitativt forskere unngår valg mellom usikre beslutninger. I slike situasjoner gir probabilistiske resonnementer en mulighet for å si hvordan de vurderer en slik usikkerhet. Kvalitative metoder gir små muligheter for å kvantifisere hvordan en velger å vurdere en slik usikkerhet.

Når det gjelder den vekt som legges på design av undersøkelser (pkt. 8.3) og seleksjonsproblemer (pkt. 8.4) er dette ut fra følgende begrunnelser:

- Det gjennomføres alt for mange survey-undersøkelser som ikke har utgangspunkt i et undersøkelsedesign som kan begrunne de konklusjoner som trekkes.
- Seleksjonseffektene har større konsekvenser en ofte antatt.

Det første punktet vil de fleste være enige om. Grunnene til at en får mengder av undersøkelser av begrenset verdi for evalueringer kan være mange. Ikke alltid er forskeren alene skyldig. Mangel på tid og ressurser er den vanligste, om ikke alltid den beste begrunnelsen.

Når det gjelder seleksjonseffekter er konsekvensene av å ignorere dette fenomenet mindre kjent. Et eksempel kan være illustrerende. Federal Trade Commission (FTC) i USA var i en periode på 80-tallet pådrivere for å få leger til å annonsere for sine tjenester. En studie av effekten av dette viser at priser hadde gått opp og kvalitet var blitt bedre som følge av annonseringen. Legene hadde rettet annonsekampanjene mot mer velstående om mindre prissensitive pasienter og hadde nådd frem. Dersom en ikke hadde kontrollert for seleksjonseffekter ville studien ha vist at prisene hadde gått signifikant ned. (Rizzo & Zeckhauser, 1991). Ser en bort fra seleksjonsproblemene, kan en altså lett havne opp med en uriktig konklusjon.

Det problemer som er drøftet er få og utvalgte. Gjennomgangen er overfladisk og lang fra komplett. Det er likevel et håp at enkelte av de tema som drøftes, kan skape debatt som bringer feltet videre.

12 REFERANSER

Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*(84), 888-918.

Akerlof, G. (1970). The Market for Lemons: Quality Uncertainty and the Market Mechanism. *Quarterly Journal of Economics, 84*.

Alvesson, M., & Skoldberg, K. (1994). *Tolkning och reflektion - vetenskapsfilosofi och kvalitativ metod*. Lund: Studentlitteratur.

Bock, D. R. (1975). *Multivariate Statistical Methods in Behavioral Research*. New York: McGraw-Hill.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bourdieu, P. (1996). *Homo Academicus*. Stockholm: Brutus Ostlings Bokforlag.

Bratberg, E., Grasdal, A., & Risa, A. E. (2000). *Evaluating social policy by experimental and nonexperimental methods*. Bergen: Department of Economics University of Bergen.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*(54), 297-312.

Campbell, D. T., & Boruch, R. F. (1975). Making the case for randomized assignment to treatment by considering the alternatives: six ways in which quasi-experiments tend to underestimate effects. In C. A. Bennet & A. A. Lumsdaine (Eds.), *Evaluation and Experience: Some Critical Issues in Assessing Social Programs* (pp. 195-296). New York: Academic.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Campbell, D. T., & Stanley, J. C. (1969). *Experimental and quasi-experimental designs for research*. (5th print ed.). Chicago ,.

Christensen, T., Laegreid, P., & Wise, L. R. (2002). Transforming Administrative Policy. *Public Administration, 80*(1), 153-178.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation Design & analysis issues for field settings*. Boston: Houghton Mifflin Co.

Coser, L. S. (1975). Presidential Address: Two Methods in Search of a Substance. *American Sociological Review, 40*, 671-700.

Cronbach, L. J. (1980a). *Toward reform of program evaluation*. San Francisco: Jossey-Bass.

Cronbach, L. J. (1980b). *Towards Reform in Program Evaluation: Aims, Methods, and Institutional Arrangements*. (1st ed.). San Francisco: Jossey-Bass Publishers.

Darcy, R. L. (1981). Value Issues in Program Evaluation. *Journal of Economics Issues, XV*(2), 449-461.

Davis, J. A. (1978). Studying categorical data over time. *Social Science Research, 7*(151-179).

Demsetz, H. (1989). *Efficiency, Competition and Policy*. (Vol. II). Cambridge, MA: Basil Blackwell.

Derthick, M., & Quirck, P. J. (1985). *The Politics of Deregulation*. Washington D.C.: The Brookings Institution.

Donaldson, S. I. (2001). Overcoming our Negative Reputation: Evaluation Becomes Known as a Helping Profession. *Amreican Journal of Evaluation, 22*(3).

Eagly, A. H., & Chaiken, S. (1992). *The psychology of attitudes*. San Diego: Harcourt Brace Jovanovich.

Eccles, J. S., & Wigfield, A. (2002). Motivational Beliefs, Values and Goals. *Annual Review of Psychology*(53), 109-132.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA.: Stanford University Press.

Finch, J. (1986). *Research and policy : the uses of qualitative methods in social and educational research*. London: Falmer Press.

Flick, U. (2002). *An introduction to qualitative research*. (2nd ed.). London: Sage.

Gilje, N. (1987). *Hermeneutikk i vitenskapsteoretisk perspektiv*. Bergen.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory*. Chicago: Aldine.

Goshal, S., & Moran, P. (1996). Bad for practice: A critique of the Transaction cost Theory. *Academy of Management Review, 21*(1), 13-47.

Grasdal, A. (2001). *Evaluation of labour market outcomes with experimental and non-experimental methods*. [Bergen]: Department of Economics University of Bergen.

Guba, E. G. (1990). *The Paradigm dialog*. Newbury Park, Calif.: Sage Publications.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, Calif.: Sage Publications.

Haveman, R. H. (1987). *Poverty Policy and Poverty Research.* Madison: University of Wisconsin Press.

Heckman, J., J. , Ichimura, H., & Todd, P. (1998). Matching As An Econometric Evaluation Estimator. *Review of Economic Studies*(65), 261-294.

Heckman, J., J., & Smith, F., A. (1999). The Pre-Programme Earnings Dip and the Determinants of Participation in a Social Programme. Implications for Simple Programme Evaluation Strategies. *The Economics Journal, July*(109), 313-348.

Heckman, J. J. (1979). Sample Bias as a Specification Error. *Econometrica, 47*, 153-161.

Heckman, J. J. (1988). *The microeconomic evaluation of social programs and economic institutions ; The value of longitudinal data for solving the problem of selection bias in evaluating the impact of treatments on outcomes*. Nankang, Taipei, Taiwan, Republic of China: Institute of Economics Academia Sinica.

Heckman, J. J. (1992). Randomization and Social Program Evaluation. In C. Manski & I. Garfinkel (Eds.), *Evaluating Welfare and Training Programs* (pp. 201-230). Boston: Harvard University Press.

Heckman, J. J. (1995). *Randomization as an instrumental variable*. Cambridge, MA: National Bureau of Economic Research.

Heckman, J. J. (1999). *Casual parameters and policy analysis in economics : a twentieth century retrospective*. Cambridge, MA: National Bureau of Economic Research.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997a). Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *Review of Economic Studies, 64*, 605-654.

Heckman, J. J., Smith, J., & Clements, N. (1997b). Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts. *Review of Economics Studies, 64*, 487-535.

Heckman, J. J., & Smith, J. A. (1995). Assessing the Case for Social Experiment. *Journal of Economic Perspectives, 9*(2), 85-110.

Heckman, J. J., & Smith, J. A. (1998). *Evaluating the welfare state*. Cambridge, MA.: National Bureau of Economic Research.

Heckman, J. J., Tobias, J. L., & Vytlacil, E. (2000). *Simple estimators for treatment parameters in a latent variable framework with an application to estimating the returns to schooling*. Cambridge, MA: National Bureau of Economic Research.

Hervik, A., Hagen, K. P., Nyborg, K., & Scheel, H. H. (1998). *Nytte-kostnadsanalyser. Veiledning i bruk av lønnsomhetsvurderinger i offentlig sektor* (NOU 1998:16). Oslo: Finans- og tolldepartementet.

Hervik, A., Hagen, K. P., Nyborg, K., Scheel, H. H., & Sletner, I.-J. (1997). *Nytte-kostnadsanalyser. Prinsipper for lønnsomhetsvurderinger i offentlig sektor* (NOU 1997:27). Oslo: Finans- og tolldepartementet.

Hjelbrekke, J. (1999). *Innføring i korrespondanseanalyse*. Bergen: Fagbokforlaget.

Jacobides, M. G., & Croson, D. C. (2001). Information Policy: Shaping the Value of Agency Relationships-. *Academy of Management Review, 26*(2), 202-223.

Jonassohn, K., Coleman, J. S., & Johnstone, J. W. C. (1961). *Social Climates in High Schools*. Washington: US government printing service.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8 Structural equation modeling with the SIMPLIS command language.* Chicago: Scientific Software International.

Kemmis, S., & Stake, R. E. (1988). *Evaluating curriculum*. Geelong, Vic.: Deakin University.

King, J. A., Stevahn, L., Ghere, G., & Minneka, J. (2001). Toward a Taxonomy of Essential Evaluator Competencies. *American Journal of Evaluation, 22*(2), 229-247.

Knudsen, K., & Wærness, K. (2001). Kontant evaluering. *Tidsskrift for velferdsforskning, 4*(4), 252-258.

Kopala, M., & Suzuki, L. A. (1999). *Using qualitative methods in psychology*. Thousand Oaks: Sage.

Kornhauser, L. A. (2000). On Justifying Cost-Benefit Analysis. *The Journal of Legal Studies, xxix*(2), 971-1004.

Kostova, T., & Roth, K. (2002). Adoption of an Organizational Practice by Subsidiaries of Multinational Corporations: Institutional and Relational Effects. *Academy of Management Journal, 45*(1), 215-233.

Krugman, P. R. (1994). *Peddling prosperity : economic sense and nonsense in the age of diminished expectations*. New York: Norton.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Kvitastein, O. A. (2000). *Extracting Lessons from Central Debates in order to Improve Policy Implications of Evaluations.* Paper presented at the The Fourth EES Conference, Lausanne.

Kvitastein, O. A., & Hungnes, P. A. (2001). Effektanalyser basert på ikke-eksperimentelle data. *Økonomisk Forum*(7), 25-29.

LaLonde, R. J. (1986). Evaluating the econometric evaluation of training programs with experimental data. *American Economic Review, 76*, 604-620.

Lieberson, S. (1985). *Making it count: The Improvement of Social Research and Theory*. Berkeley: University of California Press.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, Calif.: Sage Publications.

Luhmann, N. (1990). *Essays on self-reference*. New York: Columbia University Press.

Luhmann, N., & Jacobsen, J. C. (1992). *Autopoiesis : en introduktion til Niklas Luhmanns verden af systemer*. K*benhavn: Politisk revy.

Manski, C. F. (1996). Learning about Treatment Effects from Experiments with Random Assignment of Treatments. *The Journal of Human Resources, xxxi*(4).

March, J. G. (1988). Bounded rationality, ambiguity, and the engineering of choice. In J. G. March (Ed.), *Decisions and Organizations* . Oxford: Basil Blackwell.

Miles, M. B., & Huberman, M. A. (1984). *Qualitative Data Analysis. A Sourcebook for New Methods*. London: Sage.

Mohr, L. (1999). The Qualitative Method of Impact Analysis. *American Journal of Evaluation, 20*(1), 69-84.

Mohr, L. B. (1992). *Impact Analysis*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Mohr, L. B. (1995). *Impact analysis for program evaluation*. (2nd ed.). Thousand Oaks, Calif.: Sage Publications.

Nathan, R. P. (1988). *Social Science in Government: Uses and Misuses*. New York: Basic Books.

OECD. (1995). *Governance in transition: Public Management Reforms in OECD Countries.* . Paris: Organisation for Economic Co-ordination and Development.

Owen, J. M., & Rogers, P. J. (1999). *Program Evaluation: forms and approaches*. London: Sage.

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

Pearson, K. (1911). *Grammar of Science*. London: A. and C. Black Publishers.

Peirce, C. S., Cohen, M. R., & Dewey, J. (1923). *Chance, love, and logic; philosophical essays*. New York: Harcourt.

Petersen, T. (1991). The Statistical Analysis of Event Histories. *Sociological Methods & Research, 19*(3), 270-323.

Powell, W. W., & DiMaggio, P. J. (1991). *The New Institutionalism in Organizational Analysis*. Chicago and London: University of Chicago Press.

Quandt, R. (1972). A New Approach to Estimating Swtching Regression. *Jornal of the American Statistical Association, 67*, 306-310.

Radin, B. A. (1998). The Government and Performance Act (GPRA): Hydra-Headed Monster or Fexible Mangement Tool? *Public Administration Review, 58*(4), 307-316.

Reichenbach, H. (1938). *Experience and Prediction: An Analysis of the Foundation and the Structure of Knowledge*. Chicago, Illinois: The University of Chicago.

Richardson, H. S. (2000). The Stupidity of the Cost-Benefit Analysis. *The Journal of Legal Studies, xxix*(2), 971-1004.

Richardson, J. T. E. (1996). *Handbook of qualitative research methods for psychology and the social sciences*. Leicester: British Psychological Society.

Rizzo, J. A., & Zeckhauser, R. J. (1991). Advertising and the Price, Quantity, and Quality of Primary Care Physician Services. *The Journal of Human Resources, XXVII*(3).

Rosenbaum, P. R. (1995). *Observational studies*. New York: Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55.

Rossi, P. H., Freeman, H. E., & Lipsey, M. W. (1999). *Evaluation : a systematic approach.* (6th ed.). Thousand Oaks, Calif.: Sage Publications.

Rossi, P. H., & Wright, J. D. (1984). Evaluation Research: An Assessment. *Annual Review of Sociology*(10), 331-352.

Roy, A. D. (1951). Some Thoughts on the Distribution of Earnings. *Oxford Economic Paper*(3), 135-146.

Rubin, D. B. (1997). Estimating Causal Effects from Large Data Sets Using Propensity Scores. *Annals of Internal Medicine, 127*(8), 757-763.

Sankar, A., & Gubrium, J. F. (1994). *Qualitative methods in aging research*. Thousand Oaks, Calif.: Sage.

Scott, W. R. (1995). *Institutions and Organizations*. Thousand Oaks: Sage Publications.

Scriven, M. (1991). *Evaluation thesaurus*. (4th ed.). Newbury Park, Calif.: Sage Publications.

Scriven, M. (1993). *Hard-won lessons in program evaluation*. San Francisco: Jossey-Bass.

Sen, A. (2000). The Discipline of Cost-Benefit Analysis. *The Journal of Legal Studies, xxix*(2), 931-952.

Stake, R. E. (1975). *Evaluating the arts in education : a responsive approach*. Columbus, Ohio: Merrill.

Stake, R. E. (1986a). *Issues in research on evaluation : improving the study of transition programs for adolescents with handicaps*. Champaign, Ill.: College of Education University of Illinois.

Stake, R. E. (1986b). *Quieting reform : social science and social action in an urban youth program*. Urbana: University of Illinois Press.

Stake, R. E. (1995). *The art of case study research*. Thousand Oaks: Sage Publications.

Stake, R. E., Easley, J. A., & Anastasiou, C. J. (1978). *Case studies in science education*. Urbana

Washington: Center for Instructional Research and Curriculum Evaluation University of Illinois at Urbana-Champaign ;

for sale by the Supt. of Docs. U.S. Govt. Print. Off.

Stouffer, S. A. (1950). Some observetions on study design. *American Journal of Sociology*(55), 355-361.

Stufflebaum, D., Guba, E. G., & Tyler, R. (1971). *Educational Evaluation and Decision Making*. New York: Peacock Publishers.

Stufflebaum, D. L. (2001). Interdisciplinary Ph.D. Programming in Evaluation. *American Journal of Evaluation, 22*(3), 445-455.

Sverdrup, S. (2002). *Evaluering, Faser, design og gjennomføring*. Bergen: Fagbokforlaget.

Syvertsen, T. (1999). Medieinstitusjoner som forskningsfelt: Tendenser i norsk kringkastingsforskning. *Norsk Medietidskrift*(2).

Sørgard, L. (1990). *Privatisering av rengjøring* (SAF arbeidsnotat nr. 51). Bergen: Senter for anvendt forskning.

Veblen, T. (1899). *The Theory of the Leisure Class: An Economic Study of Institutions.* New York: Macmillan.

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

Wallis, J., & Dollery, B. (1999). *Market failure, government failure, leadership and public policy*. Basingstoke: Macmillan.

Williamson, O. E. (1985). *Economic Institutions of Capitalism*. New York: Free Press.

Williamson, O. E. (1993). Calculativeness, trust, and economic organization. *Journal of Law and Economics*(36), 453-486.

Williamson, O. E. (1993b). Opportunism and its critics. *Managerial and Decision Economics*(14), 97-107.

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observationnal data. *Annual Review of Sociology, 25*, 659-706.

Wolfson, A. (2001). The costs and benefits of cost-benefit analysis. *The Public Interest, Fall 2001*.

Østerberg, D. (1994). *Sosiologiens nøkkelbegreper og deres opprinnelse. 4.utg.* Oslo: Cappelen akademisk forlag.

Aaron, H. J., Gramlich, E. M., Hanushek, E. A., Heckman, J. J., & Wildawsky, A. (1990). Social Science Research and Policy. *The Journal of Human Resource, 25*(2), 297-304.