Liv Langfeldt

Evaluering av forskningskvalitet

Et kritisk blikk på fagfellevurdering

NIFU skriftserie nr. 9/99

NIFU – Norsk institutt for studier
av forskning og utdanning
Hegdehaugsveien 31
0352 Oslo

ISSN 0808-4572

Forord

Denne rapporten er utarbeidet under et strategisk instituttprogram for studier av forskningskvalitet. Programmet er finansiert av Norges forskningsråd. I rapporten gjennomgås studier av fagfellevurdering av forskningskvalitet, og begrensningene ved denne type kvalitetsvurderinger diskuteres.

Programmet omfatter ellers en studie av hva sentrale forskere definerer som god forskning (NIFU Rapport 9/97) og av hvordan miljøfaktorer påvirker forskningskvalitet (under arbeid).

Rapporten er utarbeidet av Liv Langfeldt.

Oslo, november 1999

Petter Aasen
Direktør

Egil Kallerud
Seksjonsleder

Innhold

Denne rapporten tar for seg den tradisjonelle og mest utbredte formen for vurdering av faglig kvalitet – bedømmelser foretatt av fagekspertise på feltet. Første del inneholder kort historikk og ser på hva fagfeller legger til grunn for sine kvalitetsvurderinger. De to neste delene tar utgangspunkt i empiriske studier og ser på hvilke svakheter ulike former for fagfellevurdering har. Siste del oppsummerer begrensninger og ser kort på hvordan de kan reduseres.

1 Hva er fagfellevurdering?

«Peer review» er et innarbeidet begrep for evalueringer der fagfolk konsulteres for å gi en *formell vurdering av forskning innen deres eget fagfelt.* På norsk snakker vi om fagfellevurdering. Utgangspunktet er at de som bedømmer den faglige kvaliteten på et (utført eller planlagt) forskningsarbeid, er kompetent til å utføre liknende arbeid og har minst like høy fagkompetanse som den eller de som blir evaluert (Niiniluoto 1987:12). Om mulig benyttes fagfolk som man regner med har høyere kompetanse på feltet enn den eller de som skal evalueres.[1] I enkelte tilfeller kan det imidlertid være problematisk å finne en egnet evaluator som har like høy kompetanse som de som skal evalueres. Forskning er i dag svært spesialisert, og antall fagfolk som er eksperter på et gitt område er begrenset. De fremste ekspertene på et felt vil ofte være enten nære kolleger eller åpenbare konkurrenter, og vil derfor ikke anses som uhildede evaluatorer av hverandres arbeid.

Generell faglig anerkjennelse og en bred kompetanse i den aktuelle fagdisiplinen kan imidlertid også anses som viktigere enn spesialkompetanse på forskningsfeltet som skal vurderes. Problemene med å finne en fagfelle som verken er nær kollega eller konkurrent blir da mindre. I flere sammenhenger konsulteres både spesialekspertise og generell ekspertise for å få et bredest mulig grunnlag for vurderingen.

[1] Et spesielt trekk ved peer review er imidlertid at aktørene kan skifte mellom å være evaluatorer og å bli evaluert. Den som i en sammenheng vurderer f.eks. et manus fra en fagfelle kan i prinsippet ved neste anledning risikere å få sitt eget manus vurdert av denne fagfellen.

1.1 Historie og bruksområder

Bruk av fagekspertise for å bedømme forskning kan føres helt tilbake til enkelte forløpere for vitenskapelige tidsskrift på 1600-tallet (Zuckerman og Merton 1971). Dagens vurderingsformer for tidsskriftartikler er mer formaliserte og regulerte, men det er også store variasjoner mellom tidsskrift. Hvert enkelt tidsskrift har stort sett utviklet sin form for fagfellevurdering ut fra egne behov. Redaksjonsråd som opprinnelig tjente som kanaler til å oppspore aktuelle manus, har med økt tilgang på stoff i stedet fått i oppgave å sortere ut de manus som bør trykkes. Redaktører uten redaksjonsråd har gradvis gitt tapt for den økende spesialiseringen i forskningen og har i større utstrekning bedt kolleger og venner om å vurdere innkomne manus. Uformelle evalueringsordninger er stadig blitt mer formalisert (Burnham 1992). En studie av 156 tidsskrifter på begynnelsen av 1960-tallet viste at 71 prosent brukte en eller annen form for ekspertgjennomgang av manus som skulle trykkes (Zuckerman og Merton 1971:75). I dag har de fleste tidsskrift med akademiske ambisjoner innført formelle ordninger for vurdering av innkomne manus, gjerne «doble blind review» som betyr at evaluator ikke får vite hvem som er forfatteren, samtidig som forfatteren ikke kjenner evaluatorens identitet.

En annen form for fagfellevurdering med lange tradisjoner er bedømmelseskomiteer for vitenskapelige priser. Særlig på 1700-tallet ble det i europeiske land hyppig utlyst priser for det beste essay som bidro med ny innsikt innen oppgitte emner, og ekspertkomiteer ble nedsatt for å vurdere kandidatene (Burnham 1992:56). Et annet område hvor det oppstod ordninger med formelle bedømmelseskomiteer, var ved ansettelser i vitenskapelige stillinger. I etterkrigstiden, da forskningsrådene ble etablert, ble også fagfellevurdering av prosjekt-søknader utbredt. Det er en del faglige og nasjonale variasjoner i slike ordninger, men en vesentlig del av offentlige midler til grunnforskningsprosjekter fordeles på grunnlag av faggruppers og/eller individuelle fageksperters bedømmelse av søknader (Chubin og Hackett 1990; Mazuzan 1992; van de Kaa 1993; Wood 1997).

De siste ti-årene har forskningsråd og myndigheter tatt i bruk fagfellevurdering i flere nye sammenhenger, og vi har fått en bruk av fagfellevurdering som skiller seg vesentlig fra den «tradisjonelle» fagfellevurdering som er knyttet til manus- og individ-nivået.[2] Nå foretas evalueringer av hele forskningsinstitusjoner, fagdisipliner og forskningsprogrammer. Når det er fagfeller som settes til bedømmelsesarbeidet, faller disse evalueringene inn under kategorien fagfellevurdering. Som vi skal se i del 3 skiller slike «meso- og makro»-evalueringer seg fra «mikro»-evalueringer ved at de kan ha mindre konkrete formål og ikke nødvendigvis er knyttet opp mot fordeling av knappe goder.

[2] Ved tradisjonelle fagfellevurderinger er det den enkelte forskers kompetanse, manus eller prosjektsøknad som blir bedømt, og rangert, med henblikk på ledige stillinger, publisering, prosjektmidler eller spesielle utmerkelser.

Fagfellevurdering er for øvrig mest utbredt som evalueringsmetode i offentlig finansiert forskning og særlig i grunnforskning. Næringslivet benytter i mindre grad fagfellevurdering for å bedømme sin FoU-virksomhet (Bozeman 1993). Dette har blant annet sammenheng med forskningens formål og målgruppe. Grunnforskning er rettet mot generell kunnskapsproduksjon og forskersamfunnets kvalitetskriterier, mens næringslivets FoU er rettet mot praktiske anvendelser og vurderes ikke bare ut fra de typiske «fagfellekriteriene» (Gulbrandsen og Langfeldt 1997).

1.2 Grunnlaget for fagfellevurdering

Ideelt sett kan en si at en evaluering bør være basert på klare og eksplisitte kriterier, være etterprøvbar og gi samme utfall uansett hvem som evaluerer. Dette er ikke mulig ved fagfellevurdering av forskning. Fagfellevurdering vil være basert på faglig skjønn og kriteriene kan være implisitte og vage, og hvem som evaluerer kan ha betydning for utfallet av evalueringen. En vitenskapssosiolog beskriver grunnlaget for fagevalueringer ved hjelp av begreper som «subtle» og «tacit judgements» og «intimate craft knowledge», og legger til at på grunn av den raske utviklingen ville eventuelle formaliserte kriterier for hva som har vitenskapelig verdi bli «a blunt and obsolete instrument» så snart det var tatt i bruk (Ravetz 1971:274).

Konklusjonene er med andre ord ikke uavhengig av tid. Hva som er adekvate metoder og verdifulle resultater er ikke noe det finnes gitte svar på, det er noe som defineres blant annet gjennom de mange former for formelle og uformelle bedømmelser som foretas innen et fagfelt,[3] og en vesentlig del av vurderingene er basert på «taus» fagkunnskap som det kan være vanskelig å redegjøre for (Niiniluoto 1987).

Samtidig er fagfellevurdering en konservativ vurderingsform. Vurderingene baserer seg på standarder internalisert i de fagmiljøer som evaluatorene har vært tilknyttet. Disse standardene vil være relatert til tidligere evalueringer (formelle og uformelle) og hva som regnes som de mest prestisjetunge publikasjoner, tidsskrifter, miljøer og personer på feltet (Cole 1983:136-138).

Hvem som har skrevet noe, hvordan det er relatert til tidligere forskning, hvor det er publisert eller hvem som har bedømt det, kan også brukes som et selvstendig argument for at noe er god eller dårlig forskning, uten at de implisitte forutsetningene for resonnementet diskuteres, nevnes eller bedømmes. Dette medfører «taus» konsolidering av normer. Konklusjonene befestes, mens premissene tas for gitt. En som evaluerer forskning det ikke finnes etablerte meninger om, kan på den annen side bidra til «taus» utvikling av normer hvis det ikke redegjøres for grunnlaget for bedømmelsen.

[3] Forskning har blant annet vist at vektleggig av kriterier læres gjennom gruppeinteraksjon i vurderingskomiteer (Ajenstat 1993).

Det er likevel mulig å si en del om hvilke kriterier som vektlegges i ulike typer bedømmelser. En studie fra NIFU basert på intervjuer med 64 norske forskere viser at original og solid (grundig og holdbar) forskning og faglig betydningsfull/faglig interessant forskning[4] er generelle kjennetegn for god grunnforskning på tvers av fag. Disse begrepene utdypes noe forskjellig i ulike fag og de enkelte kjennetegnene har også ulik vekt i ulike fag. Forskning som skårer høyt på enkelte kjennetegn, men lavt på andre, kan likevel være fremragende. Informantene var forbeholdne med hensyn til muligheten for å utvikle detaljerte kriterier eller retningslinjer for bedømmelse av forskning. Det vil uansett være et subjektivt element i vurderingene og helhetsvurdering og avveininger mellom kriterier er nødvendig. Skjønn, og kanskje intuisjon, vil være sentralt (Gulbrandsen og Langfeldt 1997).

Intervjumaterialet inneholder også svar på spørsmål om hva som vektlegges i ulike typer vurderinger: vurdering av prosjektsøknader i forskningsråd, vurdering av manus for vitenskapelige tidsskrifter og vurdering av søkere til vitenskapelige stillinger ved universiteter.[5]

Ved vurderinger av prosjektsøknader ser prosjektets potensial for original og faglig betydningsfull/interessant forskning ut til å være en viktig del av bedømmelsen (hyppigst nevnt). Muligheten for å gjennomføre prosjektet understrekes som viktig å vurdere i alle fag. Hva som vektlegges her ser imidlertid ut til å variere noe mellom fag. Tekniske fag, matematikk og sosialøkonomi skiller seg ut ved å *ikke* nevne prosjektbeskrivelsen som et viktig grunnlag for vurderingen, mens medisinere vektlegger faktorer knyttet til søkerens kvalifikasjoner, tidligere prestasjoner og forskningsmiljø, faglige nettverk mer enn det forskere i andre fag gjør.[6]

Særlig to typer søknader ble sagt å være vanskelige å vurdere: søknader utenfor eller på grensen av eget kompetanseområde og søknader med uklare prosjektbeskrivelser. Det kunne ellers være vanskelig å veie ulike hensyn mot hverandre. Noen sa også at det kunne være problematisk å bedømme prosjektsøknader fra forskere de kjente og hadde (negativ) informasjon om mulighetene for å lykkes med prosjektet utover det som fremkom av søknaden, mens andre derimot nevnte prosjekter der man *ikke* kjente søkeren, som problematisk å bedømme.

Også ved vurdering av manus for vitenskapelige tidsskrifter ble uklare manus og forskning utenfor eller på grensen av eget kompetanseområde hyppig nevnt som eksempler på hva som

---

4 F.eks. forskning med faglige ringvirkninger/bred faglig interesse.

5 Materialet er ikke tidligere presentert. Intervjuundersøkelsen omfatter 64 forskere i ulike fag og sektorer. Her tar vi for oss svarene fra de 30 forskere i universitetssektoren. Fagene som er inkludert er filosofi, fransk, sosiologi, sosialøkonomi, kjemi, matematikk, medisin, bioteknologi og teknisk kybernetikk. Se ellers Gulbrandsen og Langfeldt (1997) for beskrivelse av metode og datamaterialet.

6 Disse fagforskjellene stemmer også med funn i en studie av fagfellevurdering i Norges forskningsråd som blant annet tok for seg hvilke kriterier som ble vektlagt i bedømmelser av søknader om "frie prosjektmidler" (Langfeldt 1998).

var vanskelig å vurdere. Her var klar og god språkføring og argumentasjon også en viktig del av vurderingen. Ulike faktorer knyttet til soliditet ser ut til å være spesielt sentralt ved denne typen vurderinger: etterrettelighet, overbevisende dokumentasjon, solide metoder og holdbare analyser og sammenheng mellom data og konklusjoner. Relevans for tidsskriftets lesere ble også vektlagt.

Da informantene ble spurt om hva de vektla når vitenskapelige stillinger skulle besettes, stod den faglige produksjonen til søkerne og hva de hadde tilført faget av original, spennende forskning sentralt. I denne typen søknader syntes flere også at søkernes personlighet var relevant. Stillingene burde besettes av omgjengelige og motiverte personer med evne til å bygge opp et fagmiljø og med pedagogiske evner. Personlige egenskaper var blant det forskerne syntes det var vanskeligst å bedømme. Som ved vurdering av prosjektsøknader, kunne det ved prioriteringen mellom kandidater også være vanskelig å veie ulike hensyn mot hverandre.

Det er funnet vesentlige fagforskjeller i stillingsbedømmelser. En studie av innstillingsdokumentene for 31 professorater i ulike fag i Sverige, kom blant annet til at det i humaniora og samfunnsvitenskap legges større vekt på teoriutvikling, mens det i de «hardere» disiplinene legges mer vekt på analyseresultater. Det er generelt større variasjon i hva som vektlegges i de «myke» disiplinene enn i de «harde» disiplinene. Uttalelsene fra de «myke» disiplinene er også lengre enn fra de «harde» disiplinene (Montgomery og Hemlin 1991).

Oppsummert er det mulig å sette opp felles og generelle kjennetegn på god forskning, men det er variasjoner mellom fag i hva som vektlegges og også i hva som er viktig avhengig av om det er manus, søknader om midler eller søknader til stillinger som vurderes. At vurdering av forskning utenfor eller på grensen av eget kompetansefelt og avveining mellom ulike kriterier/hensyn kan være særlig vanskelig, bekrefter at skjønn og fagekspertise er sentralt i fagfellevurdering.

Neste del av rapporten tar for seg en annen type studier av denne type vurderinger – studier som ser på samsvar mellom hvordan forskere bedømmer samme objekt og på muligheten for partiske vurderinger.

2 Vurderinger på manus- og forskernivå

Fagfellevurdering av forskning ved publisering i fagtidsskrifter, ved prosjektbevilgninger fra forskningsråd, og ved tilsettinger ved universiteter er den tradisjonelle formen for fagfellevurdering. Dette er vurderinger hvor hver(t) enkelt forsker eller manus bedømmes opp mot hverandre. Vi kan kalle dette fagfellevurdering på mikronivå. Det er en gitt mengde søkere, kandidater eller manus som vurderes opp mot hverandre for å ta beslutninger om fordeling av et knapt gode. Det er én vitenskapelig stilling som skal besettes, et begrenset antall sider i et tidsskrift som skal fylles eller en begrenset pott med forskningsmidler som skal fordeles. Med utgangspunkt i empiriske studier av samsvar i bedømmelser eller mulighetene for partiske vurderinger, skal vi her se på hvilke begrensinger som ligger i fagfellevurdering av denne typen.[7]

2.1 Studier av fagfellevurdering for tidsskrifter

En studie som tok for seg 449 manus til et tysk kjemitidsskrift og to ekspertuttalelser til hver manusene, fant lav grad av samsvar mellom anbefalingen til de to ekspertene (Daniel 1993). Bare i 38 prosent av tilfellene hadde de to ekspertene krysset av i samme rubrikk.[8] Imidlertid var det bare 23,5 prosent som skilte mer enn én kategori og dette tolkes som betydelig enighet i vurderingene (kappa = 0.67, Daniel 1993:26, 72). Det vektlegges at redaktører for vitenskapelig tidsskrift vanligvis velger ut eksperter som komplimenterer hverandre (en spesialist og en generalist) og at ekspertuttalelsene sjelden tar for seg de samme poengene. En høy grad av enighet mellom ekspertene kan derfor ikke forventes.

Også ulike former for partiskhet ble studert i dette materialet: forfatterens akademiske status og nasjonalitet og manuskriptets tema-område. Daniel fant at akademisk status og nasjonalitet hadde noe betydning for om manus ble publisert[9], mens emneområde ikke hadde betydning. Videre ble skjebnen til refuserte manus studert. Manus som ble publisert, ble sitert omtrent dobbelt så hyppig som artikler som ble refusert og publisert andre steder. Daniel tolker dette som «high predictive validity».[10]

---

7 Litteraturen om fagfellevurdering er nokså uoversiktlig med «lokal» litteratur og «lokale» debatter fordelt på en rekke fag og tidsskrifter. For å orientere meg i litteraturen har jeg dels brukt «snøballmetoden» og dels hatt nytte av litteratursøk i bibliotekdatabaser. Spesielt nyttig har den annoterte bibliografien til Speck (1993) vært. Den tar for seg 780 publikasjoner, 664 under «Journal Peer Review» og «Book Peer Review», resten under «Grant Peer Review».

8 Det var fire anbefalingskategorier: publiser uten endringer, publiser etter små endringer, publiser etter betydelige endringer og ikke publiser.

9 Fagfellevurderingen i dette tidsskriftet var ikke «blind», ekspertene kjente forfatterens identitet. U-landsforfattere og ikke-professorer fikk sjeldnere inn sine manus.

10 Han tar imidlertid forbehold om at hyppigere sitering dels kan relateres til at det studerte tidsskriftet hadde høyere «impact factor» enn de tidsskriftene som senere publiserte de refuserte manuskriptene.

Ceci og Peters (1982) foretok en skjult studie av 12 artikler som var publisert i 12 ulike amerikanske psykologi-tidsskrifter. Alle artiklene var skrevet av forfattere fra institutter med høy prestisje, og publisert i tidsskrifter som oppgir navn og institusjonstilknytning på manusforfatterne til ekspertene som får manus til vurdering. Ceci og Peters ga de tolv utvalgte artiklene fiktive forfatter- og institusjonsnavn, og endret noe på overskrifter, sammendrag, figurtyper etc. (rent kosmetisk), for at artiklene ikke skulle være umiddelbart gjenkjennelige. Deretter sendte de artiklene som manus til de respektive tidsskrifter med anmodning om at de skulle vurderes for publisering. Hverken redaktørene eller evaluatorene fikk noen informasjon om at manusene tidligere var publisert i et annet navn. I tre tilfeller ble «testen» oppdaget enten av redaktøren eller en av evaluatorene. Av de ni resterende artiklene ble åtte refusert, og én godtatt for publisering. I mange tilfeller ble «serious methodological flaws» oppgitt som grunnen til refusjon. Ikke i noe tilfelle fikk forfatterne tilbakemelding som tydet på at refusjonen skyldtes at manuset ikke var interessant fordi det var «gammelt nytt». Ceci og Peters mener at dataene tyder på at forskere med tilknytning til lavstatus institusjoner diskrimineres, selv om de medgir at resultatene også kan forklares på andre måter. Tidsskriftene som ble undersøkt hadde en avslagsrate på 80%, noe som generelt gir en lav sannsynlighet for at samme manus skal aksepteres to ganger (Ceci og Peters 1982b:46).

En studie av betydningen av institusjonstilknytning for publisering i et pediatrisk tidsskrift som også oppgir forfatteridentitet til evaluatorene, fant at institusjonsstatus *ikke* hadde betydning for vurdering av «major papers», men at det hjalp å ha institusjonell prestisje i ryggen for å få publisert «brief reports» (Garfunkel et al. 1994).[11] En studie av betydningen av forfatterens kjønn som inkluderte 2160 manus til JAMA (som også oppgir forfatteridentitet til evaluatorene) fant ingen forskjellsbehandling i vurderingene (Gilbert 1994).

Andre studier som ser på forskjellen mellom «blinde» og «ikke-blinde» manusvurderinger finner derimot at evaluatorens kjennskap til forfatterens identitet har betydning. Laband og Piette (1994) tok for seg 1051 artikler til 28 («top») økonomitidsskrifter og fant at artikler til de av tidsskriftene som holder forfatteridentiteten ukjent for evaluatorene blir vesentlig mer sitert enn artikler i tidsskrift som oppgir forfatteridentitet. Laband og Piette mener at dette viser at vurderingene blir bedre når forfatter er ukjent og at evaluatorer som kjenner forfatteridentitet lett kan bytte ut universelle vurderingskriterier med mer partikularistiske kriterier.

I en studie av Fisher et al. (1994) ble 57 manus (i pediatri) sendt til to «blinde» og to «ikke-blinde» eksperter for vurdering. Resultatet ble satt opp mot antall artikler forfatterene hadde publisert tidligere. Fisher et al. fant at evaluatorer som kjenner forfatteridentitet gir manus fra

[11] Alle manus innkommet til redaksjonen i løpet av 7 måneder var med i studien (405). Institusjonene ble rangert etter samlede prosjektbevilgninger fra NIH.

forfattere med flere tidligere artikler dårligere karakter enn det evaluatorer som ikke kjenner forfatteridentitet gjør. Dette mener de kan skyldes at «ikke-blinde» evaluatorer kan være partiske, for eksempel at konkurranse og misunnelse spiller inn i karaktergivningen. (Studien viste imidlertid også at nesten halvparten av de «blinde» evaluatorene gjettet riktig forfatter på manusene).

Også trekk ved selve manuset kan vurderes «skjevt». En skandinavisk studie fant at en engelsk versjon av et manus fikk bedre bedømmelse enn den skandinaviske versjonen (Nylenna et al. 1994). En studie av Mahoney (1977) så på betydningen av resultatene som presenteres. Mahoney fant en sterk tilbøyelighet blant ekspertene til å undervurdere resultater som gikk mot deres teoretiske perspektiv ved å studere 67 ekspertuttalelser om ulike versjoner av et manus. Mahoney kaller dette «bekreftende partiskhet» (confirmatory bias). Mahoney er blitt kritisert for at materialet hans ikke gir godt nok grunnlag for å påvise bekreftende partiskhet, og det er blitt påpekt at det hos Hull finnes grunnlag for å trekke motsatte konklusjoner (Chubin og Hackett 1990:100-102). Hull (1988) finner at fagfeller kan gi en bevisst ukritisk vurdering av manus skrevet innen et konkurrerende paradigme. Hensikten er å bidra til at manus som ikke holder mål blir publisert og derved «blamere» det konkurrerende paradigmet. Det kan imidlertid også tenkes andre forklaringer på at evaluatorer er lite kritiske overfor andre faglige ståsted enn sitt eget. Det kan for eksempel skyldes at de ikke føler seg kompetente til en kritisk gjennomgang, eventuelt kombinert med «generøsitet» eller respekt for pluraliteten i faglige tilnærminger innenfor feltet.

2.2 Studier av fagfellevurdering av prosjektsøknader

I en studie av bedømmelseskomiteer for prosjektsøknader i et britisk forskningsråd skiller Travis og Collins (1991) mellom kognitiv og institusjonell partiskhet. *Kognitiv* partiskhet vil si at evaluatorens vitenskapelige ståsted eller kognitive interesser påvirker vurderingene. *Institusjonell* partiskhet vil si at institusjonell tilknytning, eller sosiale strukturer generelt, har innvirkning på vurderingene.[12]

I Travis og Collins analyse av den muntlige kommunikasjonen som går forut for beslutninger om finansiering av prosjekter, er det kognitiv partiskhet som står i sentrum. De argumenterer for at når det gjelder bedømmelser av forskning, er kognitiv partiskhet langt mer alvorlig enn institusjonell partiskhet. Såfremt ikke kognitive og institusjonelle grenser er overlappende, har kognitiv partiskhet langt større betydning for forskningens innhold og utvikling enn institusjonell partiskhet. Travis og Collins' studieobjekter er fagkomiteer (med ca. 12 medlemmer) som diskuterer og bedømmer prosjektsøknader på basis av skriftlige vurderinger fra tre-fire individuelle evaluatorer («postal referees»). I sin artikkel refererer Travis og Collins eksempler der fagkomiteen diskuterer det «kognitive fellesskapet» mellom forskeren som søker om prosjektmidler og forskerne som har vurdert prosjektsøknaden, som bakgrunn

[12] Av studiene som er referert over, er Mahoney (1977) et eksempel på en studie av kognitiv partiskhet, mens Ceci og Peter (1982) behandler institusjonell partiskhet.

for fagkomiteens egen vurdering. I et referert tilfelle besluttet fagkomiteen at en søknad ikke var støtteverdig på tross av at alle de individuelle evaluatorene hadde gitt meget positive vurderinger av søknaden. På grunn av det «kognitive fellesskapet» mellom evaluert og evaluatorer, mente komitémedlemmene at de heller burde stole på sine egne vurderinger. I et annet tilfelle ble en lignende sak vurdert som så vanskelig at den ble utsatt til et senere møte. Behandlingen av to slike utsatte saker er også beskrevet. Det ble her innhentet nye vurderinger fra fagfeller tilhørende andre «kognitive fellesskap» enn søkerne. Disse var adskillig mer negative enn den første runden der evaluatorene tilhørte samme «kognitive fellesskap» som søkeren, og begge søknadene ble avslått.[13]

Travis og Collins viser at fagkomitémedlemmene har et stort forhandlingsrom når det gjelder tolkningen av de individuelle vurderingene. På den måten kan de jobbe for de søknadene de selv synes er viktige. «Kognitive fellesskap» som ikke er representert i komiteen vil da stå svakt. Forfatterne nevner spesielt nye fagområder, tverrvitenskapelig forskning og kontroversielle områder som utsatte.

En evaluering som General Accounting Office (GAO) i USA foretok av søknads-vurderingsprosessene ved National Institutes of Health (NIH), National Science Foundation (NSF) og National Endowment for the Humanities (NEH) undersøkte blant annet skjevheter («bias») i karaktergiving (GAO 1994). Studien omfattet et utvalg av 254 søknader, spørreundersøkelse til 1370 evaluatorer, intervjuer med forskningsrådspersonale og observasjon av enkelte komitémøter. GAO fant at karakterer for det meste var uavhengig av egenskaper ved evaluatorene eller søkerne. Menn og hvite fikk imidlertid bedre karakterer enn kvinner og minoriteter, og evaluatorens kjennskap til søkeren og oppfatning av søkerens tidligere arbeid ga også utslag på karaktergivingen.[14] Det påpekes at årsaken til disse skjevhetene enten kan være at erfarne, kjente, hvite mannlige forskere skriver bedre søknader enn andre, at de kjenner reglene og normene for søknadsskriving bedre, eller at det er «bias» i karaktergivingen (GAO 1994:4). Evalueringen fant ingen kognitiv partiskhet. Den undersøkte imidlertid bare effekt av forskningsspesialitet, ikke av eventuelle ulike faglige ståsted i form av forskningsretninger eller skoler.

---

[13] Travis og Collins var til stede på ti fagkomitémøter i *Science and Engineering Research Council* i Storbritannia og tok opp dialogen på bånd. Også en studie av bedømmelsesprosessen i det australske forskningsrådet finner at en del av arbeidet til fagkomiteeen består i å vurdere evetuell partiskhet i uttalelsene fra de individuelle ekspertene og kalibrere og eventuelt moderere karakterene deres (Wood 1995).

[14] Andre faktorer som ble undersøkt og som *ikke* viste signifikante effekter, var evaluators faglige nærhet til søknadens forskningsfelt og kjennskap til litteraturen referert i søknaden, søkerens geografiske tilhørighet og akademiske stilling og evaluatorens oppfattelse av prestisjen til søkerens institusjon. En studie av Pfeffer et al. (1976) som brukte mer aggregerte data fra NSF (institutt-representasjon i komiteene som behandlet søknadene og hvor mye midler instituttene fikk som resultat av søknadsbehandlingen), fant imidlertid at representasjon i komiteen ga signifikant utslag på tildeling av midler, kontrollert for instituttstørrelse og «instituttkvalitet», det siste målt med «rankinglister» fra American Council on Education.

En studie av søknadsvurderingen ved tildeling av postdoktorstipend ved det medisinske forskningsråd i Sverige konkluderer adskillig skarpere om favorisering av menn. Ved å sette karaktergiving på faglig kompetanse opp mot søkernes publikasjoner i anerkjente tidsskrifter og hvor hyppig andre forskere siterte arbeidene deres fant de at kvinner måtte ha publisert og være sitert adskillig mer enn menn for å oppnå den samme bedømmelsen. De fant også at inhabilitet hadde en positiv innvirkning på karaktergivingen – i saker der en av komitemedlemmene måtte «på gangen» fikk søkeren bedre karakter enn andre søkere med samme publisering og sitering. Det hevdes at «the system is revealed as being riddled with prejudice» (Wennerås og Wold 1997:341).[15]

I tillegg til GAOs evaluering har National Science Fondation vært gjennom en omfattende studie som ikke finner systematiske «bias». 150 prosjektsøknader til NSF ble sendt til ny vurdering hos nye «referees».[16] Resultatene tyder på at fra 24 til 30 prosent av søknadene ville fått motsatt utfall ved «revurderingen», altså enten gått fra avslått til støttet eller omvendt.[17] Forfatterne finner at en søknads skjebne er halvveis bestemt av karakteristika ved søknaden og halvveis bestemt av hvem som blir satt til å evaluere søknaden. De finner ingen tegn på institusjonell eller annen systematisk partiskhet og konkluderer med at resultatene skyldes «real and legitimate differences of opinion among experts about what good science is or should be» (Cole et al. 1981:885).

Dette er forsåvidt i samsvar med Travis og Collins (1991) sine funn, men vinklingen er en annen. Det Cole et al. nøyer seg med å kalle faglig *uenighet,* kaller Travis og Collins kognitiv eller faglig *partiskhet*. Forskjellen er at Travis og Collins med sin bruk av begrepet kognitiv partiskhet, ser ut til å mene at vurderingene er strategiske med sikte på å fremme et spesielt «kognitivt fellesskap» eller paradigme, mens Cole et al. konstaterer manglende konsensus om vurderingen av vitenskapelig kvalitet og finner ikke begrepet «bias» relevant i den forbindelse.

2.3 Studier av fagfellevurdering ved ansettelser

Det finnes også enkelte studier av bedømmelser for professorater. Disse er mer opptatt av hvilke faglige kriterier som vektlegges i uttalelsene (ref. 2.1). En studie har imidlertid også sett på kvinners muligheter til å nå opp i prosessen.

Fürst (1988) behandler beslutningsprosessen ved ansettelser i vitenskapelige stillinger ved universiteter og høyskoler i Norge. Fürst foretok bl.a. en kvalitativ innholdsanalyse av saksdokumentene i ansettelsessaker ved UiO. Analysen understreker potensialet for sterke –

---

15 Alle søkerene for 1995 er inkludert i studien – 62 menn og 52 kvinner. 16 menn og 4 kvinner fikk stipend.

16 Første del av studien var basert på 1200 søknaders skjebne. Det konkluderes her med at det ikke er systematisk «bias», men at det er moderat korrelasjon mellom vurderingene av søknadene og blant annet akademisk status og tidligere NSF bevilgninger. Betydningen av kjønn ble ikke studert (Cole et al. 1978).

17 Tallet varierer for ulike fagområder og er noe avhengig av hvilket beregningsgrunnlag som velges (gjennomsnittlig rangering i første runde eller faktisk beslutning om støtte).

men til dels ubevisste – innslag av skjønn og interesser i vurderingene. Reglene er åpne for tolkning og det er stor variasjon i hvilke kriterier som vektlegges – og på hvilken måte de vektlegges – i vurderingene. Dette gir seg utslag blant annet i at en kvinnes forskningsproduksjon gjerne karakteriseres som «snever» eller «ensidig», mens de mannlige motkandidaters forskningsproduksjon «går i dybden» eller har «tematisk sammenheng». Motsatt har en kvinne «stor spredning» i sin kompetanse, mens enn mann har «faglig bredde». I sitt forslag til forklaring av forskjellsbehandlingen henter Fürst frem begreper som «familisme» og «paternalisme». Universitetene er dominert av «farsfigurer» og «høvdinger»: det er de som utgjør bedømmelseskomiteene.[18]

2.4 Hva viser studiene?

En gjennomgang, sammenstilling og sekundæranalyse av enighet i fagfellevurdering som ble foretatt for noen år siden konkluderer: «The available data are clear. Quite low levels of chance-correlated inverreviewer agreement are obtained in every area of scientific inquiry» (Cicchetti 1991:126).[19] En annen konklusjon var at det er mer enighet om vurderingen av dårlig forskning, enn av god forskning. Ekspertene er mer enig i bedømmelsen av de som samlet får en dårlig karakter, enn i bedømmelsen av de som får en god karakter (Cicchetti 1991). Hva som er lav eller høy grad av enighet kan imidlertid diskuteres. Cole et al. fikk reaksjoner på sin studie av fagfellevurdering i NSF som gikk både på at enigheten var uventet lav og at den var uventet høy (Cole og Cole 1985:54). De studiene vi har gjennomgått her, viser også at samsvaret mellom ulike eksperters bedømmelser av det samme manuset eller den samme søknaden generelt er lavt, men med et annet utgangspunkt for analysen kan vi få tolkninger som går i en annen retning. Daniel (1993) redefinerte uenighet til å bety mer enn én karakter forskjell og fant da uenighet for 23,5 prosent av manusene, og tolket dette som betydelig enighet.

Vi kan også ta et helt annet perspektiv og vektlegge at fagfellevurdering er en prosess der ulike vurderinger og ulik ekspertise settes mot hverandre for å få et best mulig grunnlag for fordeling av knappe goder i forskersamfunnet. Da er ikke uenighet i seg selv et problem.

> «The problem with dissensus in peer review ... is not, of course, with the fact that scientists disagree. The problem lies in the mythology that they do not or should not disagree» (Cole og Cole 1985).

Når ulike eksperter har ulike vurderinger, kan det imidlertid være avgjørende for skjebnen til det enkelte manuskript, og den enkelte søknad, *hvem* som plukkes ut til å foreta vurderingen. Flere av studiene gjennomgått her, viser systematiske skjevheter i vurderingene. Ikke

[18] Studien, som mente å kunne påvise forskjellsbehandling av kvinnelige og mannlige søkere, avfødte en heftig debatt. Fürst ble kritisert for metodiske svakheter og feil (se Aubert 1989 for en oppsummerende artikkel). Kritikken var sentrert om den kvantitative delen av analysen og ikke den kvalitative innholdsanalysen som er referert her.

[19] En annen sekundærstudie konkluderer med at samsvaret mellom vurderinger av manus for tidsskrifter bare er «a little better than a dice roll» (Lindsey 1988).

overraskende finner flere at det faglige ståstedet til eksperten har betydning for bedømmelsen. Noen finner også forskjellsbehandling basert på søkernes kjønn, akademiske status, institusjonelle tilhørighet eller bekjentskaper.

Både Travis og Collins (1991), Cole et al. (1981), Mahoney (1977) og Hull (1988) mener at evaluatorens faglige ståsted er en viktig forklaringsfaktor for utfallet av evalueringer. At Mahoney bruker faglig ståsted som direkte forklaringsfaktor, mens denne faktoren er indirekte hos Hull (evaluatorene bruker en strategi for å svekke konkurrerende paradigmer som medfører at de kan være ekstra kritiske overfor bidrag innen eget paradigme), endrer ikke det faktum at det første de begge ser etter for å forklare et evalueringsutfall, er evaluatorens faglige ståsted. Travis og Collins finner at man i forskersamfunnet er svært oppmerksom på den virkningen evaluatorens faglige ståsted har på hans/hennes vurderinger, og at fagkomiteer behandler dette som et problem som bør bekjempes. Fagkomiteene som vurderer påliteligheten til evalueringer er imidlertid også satt sammen av fagfolk som har et faglig ståsted, og sammensetningen av fagkomiteen vil derfor kunne være avgjørende for hvilke «kognitive fellesskap» det satses på.

Cole et al. (1981:885) understreker at på grunn av klare uenigheter om vurdering av vitenskapelig kvalitet innen alle fagfelt, vil en prosjektvurdering være sterkt avhengig av hvilke evaluatorer som blir valgt. Også her er altså faglig ståsted den underliggende forklaringsfaktoren, men som nevnt er dette ikke som hos Travis og Collins knyttet til noen form for partiskhet. Prøver vi å se Cole et al. og Travis og Collins i sammenheng, blir spørsmålet hva som legges i begrepet kognitiv partiskhet. Ved en vid definisjon kan det inkludere alle vurderinger som er bestemt av faglig ståsted, også den faglige uenigheten som Cole et al. mener er forklaringen på motstridende utfall av evalueringer. Med en slik definisjon blir alle fagevalueringer i prinsippet «kognitivt partiske» hvis det ikke er full enighet om dem. Med en snevrere definisjon kan kognitiv partiskhet begrenses til tilfeller der favorisering av eget faglige ståsted er en bevisst strategi hos evaluatoren. Når det skal bestemmes om en gitt evaluering er partisk eller ikke, vil en slik definisjon være lite klargjørende såfremt man ikke står overfor utvetydige tilfeller av «faglig nepotisme».[20]

Et tilsvarende problem oppstår også når vi vil undersøke forskjellsbehandling på bakgrunn av institusjonell tilhørighet eller akademisk status. Hva er forskjellsbehandling på grunn av status alene og hva er ulik bedømmelse på grunn av ulik kvalitet? Effekten av berømmelse og status er velkjent i forskningsverdenen og kalles kumulative fordeler eller Matteus-effekten –

[20] I en utredning om habilitetsspørsmål i Norges forskningsråd heter det (når habilitet i fordelingssaker behandles): «Motsetninger som springer ut av ulike forskningsretninger og «skoler», sterkt avvikende vurderinger av problemers verdi, nytten av ulike metoder, tolking av resultatene o.l. vurderes som forhold der forskningsrådet har særlig grunn til å vise *aktsomhet* med hensyn til den måten beslutningsprosessen organiseres og personer velges ut på. I dette aktsomhetsområdet kan det bl.a. være aktuelt å benytte flere (eller andre typer av) sakkyndige enn normalt, stille økte krav til dokumentasjon m.v.» (Norges forskningsråd 1993, s 44). Her gjøres det ingen forsøk på å avklare hva som skal regnes som faglig nepotisme, men det påpekes at dette er et problem en må være oppmerksom på og søke å minske med organisatoriske mottiltak.

allerede berømte forskere og prestisjetunge institusjoner tiltrekker seg en uforholdsmessig stor andel av anerkjennelse, ressurser og lovende rekrutter (Merton 1968 og 1988). Det er ikke gitt at partiskhet er et dekkende begrep for fenomenet. En forskers institusjonelle tilknytning eller akademiske status kan være en til dels ubevisst forutsetning når hans/hennes forskning blir bedømt, en del av evaluatorens «tause» grunnlag for å vurdere faglig kvalitet – basert på «grunnfestede» indikatorer på god og dårlig forskning som det sjelden er naturlig å tenke eksplisitt gjennom eller stille spørsmålstegn ved. I så tilfelle dreier det seg ikke om noen bevisst partiskhet eller strategi for favorisering eller diskriminering, men om et fenomen som heller bør kalles «taus» partiskhet – om partiskhet i det hele tatt skal brukes.[21]

De studiene vi har gjennomgått er ikke entydige når det gjelder denne type skjevheter i bedømmelsene. Ceci og Peters (1982) fant «institusjonell partiskhet» i psykologi ved å bytte ut forfatter og institusjon på manus og sende dem til ny vurdering. Garfunkel et al. (1994) fant ingen forskjellsbehandling av institusjoner ved publisering av «hovedartikler», mens Fisher et al (1994) og Laband og Piette (1994) fant vesentlige forskjeller mellom «blinde» og «ikke-blinde« vurderinger og mener dette kan skyldes ulike former for «bias» når forfatteridentitet er kjent. Daniel (1993) fant favorisering av forfattere med høy akademisk stilling og fra etablerte forskernasjoner i *Angewandte Chemie.* Men samtidig fant han at manus som kom gjennom nåløyet til dette tidsskriftet ble hyppigere sitert enn de som ikke kom igjennom, men ble publisert i andre tidsskrifter. Dette mener han viser at vurderingene var riktige (høy validitet). GAO (1994) fant en viss favorisering av kjente forskere, men ikke av forskere med høy akademisk stilling. Cole et al. (1978) undersøkte betydningen av blant annet søkerenes institusjonelle tilknytning, akademiske stilling og akademiske «alder», samt om søkerne hadde fått midler fra forskningsrådet (NSF) tidligere. De fant at disse karakteristika stort sett hadde liten eller ingen betydning for bedømmelsen av søknadene. (Det var likevel noe variasjon mellom fag. Høyest korrelasjoner fant man i økonomi, lavest i antropologi og økologi.)

Fire av studiene som er gjennomgått, ser på betydningen av kjønn. Tre av dem konkluderer med at kvinner kommer dårligere ut enn menn. En svenske studie fant at kjønn hadde en uavhengig effekt på bedømmelsene av søknadene om postdoktorstipend. Kvinner må ha publisert og være sitert betydelig mer enn menn for å nå opp i konkurransen (Wennerås og Wold 1997). To andre studier fant at kvinner kom dårligere ut enn menn ved vurdering for henholdsvis universitetsstillinger og prosjektsøknader til forskningsråd (Fürst 1988; GAO 1994), mens en studie av fagfellevurdering for et tidsskrift ikke fant forskjellsbehandling (Gilbert et al. 1994).

Det er ikke mulig å gi noen entydig oppsummering av konklusjoner på grunnlag av de ulike studiene. Sprikende konklusjoner kan skyldes både at studiene undersøker ulike objekt og at de har svært ulike design. Studieobjektene er ulike fag, ulike tidsskrifter, ulike kategorier av søknader, ulike forskningsråd med ulike typer prosesser, retningslinjer og karakterskala.

[21] På engelsk løses problemet ved å bruke det flertydige «bias».

Noen studier har en eksperimentell tilnærming og sender manus eller søknader til ny vurdering, noen setter resultatene av fagfellevurdering opp mot andre mål som publikasjoner og siteringer, mens noen bare måler hvilke kategorier av forskere som lykkes eller mislykkes i prosessen. Studiene av enighet mellom evaluatorer bruker ulike korrelasjonsmål og har ulike tolkninger av korrelasjon.

Det er likevel et ganske solid grunnlag for å hevde at vurderinger varierer med bedømmerens faglige ståsted (dvs. ulike retninger eller tradisjoner i faget). Tatt i betraktning den rollen skjønn spiller i vurderingene, blant annet i avveininger mellom ulike kriterier og hensyn, og hvor viktig spesialkunnskapen til evaluatoren kan være, skal det mye til at alle aktuelle evaluatorer bedømmer forskning helt likt (ref. del 1).

Konklusjoner når det gjelder forskjellsbehandling på bakgrunn av institusjonstilknytning, akademisk stilling og andre lignende karakteristika ved den som blir bedømt, er mer uklare. Studiene kommer til ulike konklusjoner og det er dertil prinsipielt vanskelig å påvise hva som er bakgrunnen for at noen grupper kommer dårligere ut enn andre. Studiene som gjelder forskjellsbehandling på bakgrunn av kjønn ved vurdering av prosjektsøknader og søknader til stillinger er mer entydige, men den studien som har de klareste konklusjonene her omfatter bare en begrenset form for fagfellevurdering (postdoktorstipendsøknader) i ett fag. Samlet gir likevel de ulike studiene av forskjellsbehandling argumenter for at faglig irrelevante karakteristika ved forskeren som bedømmes bør holdes skjult for evaluatorene.[22]

[22] Men som vi så i del 1, mener forskere at ved vurdering av *prosjektsøknader* bør blant annet tidligere prestasjoner og institusjonstilknytning tas hensyn til i vurderingen av muligheten for å gjennomføre et prosjekt.

3 Vurderinger på institusjonsnivå og nasjonalt nivå

Denne delen av rapporten ser på hvordan fagfellevurdering fungerer når vurderingsobjektet ikke er enkeltmanus eller enkeltforskere, men hele institutter, programmer eller fagfelt.

3.1 Evalueringsformål

Evaluering av forskningsinstitutt, forskningsprogram og fagområder kan ha en rekke – mer eller mindre klare – formål. Som en fellesnevner kan vi si at krav om dokumentasjon om hvor godt offentlige midler blir brukt som regel er en viktig del av årsaken til at de utføres. De tradisjonelle fagfellevurdering-praksiser har derimot hatt mer vitenskapsinterne formål. De er prosedyrer for å fordele stillinger, bevilgninger og tidsskriftplass[23] til de beste forskerne. Ifølge offisielle dokumenter er de nye evalueringsformene også rettet mot vitenskapsinterne formål. I Hernes-utvalgets innstilling fremheves læring – «veiledning om hva som bør gjøres for å nå bedre resultater» – som hovedformålet med makro- og mesoevalueringer av forskning. I utvalgets forslag inngår imidlertid også andre formål:

> «Utvalget foreslår at forskningsmiljøer, fag og forskningsprogrammer mer systematisk blir gjenstand for evaluering ... både for 1) å fremme læring i miljøene, for 2) å gi premisser for den forskningspolitiske debatt og for 3) å sikre forskere og almenhet bredere innsyn og informasjon som basis for beslutninger» (NOU 1988:28, ss 180-181).

Norges forskningsråds evalueringshåndbok opererer med et tredelt formål for evalueringer av forskning – legitimering/kontroll, læring/motivasjon og styring. Det heter at «Et hensiktsmessig evalueringsopplegg representerer kort sagt både en kontrollfaktor, en lære- og motivasjonsfaktor og et hjelpemiddel til å systematisere foreliggende informasjon og erfaringer i beslutningsøyemed» (Norges forskningsråd 1996:3).

Hvem som initierer eller bestiller slike evalueringer og hva som er hensikten med dem kan også variere betydelig avhengig av nasjonal kontekst. I UK har man et system for nasjonal «rating» av universitetsinstitutter initiert av bevilgende myndigheter, hvor fagfellekomiteer hvert fjerde år bedømmer instituttene på grunnlag av en rekke statistiske data, publikasjonlister, intervjuer med et utvalg forskere og gjennomgang av et utvalg publikasjoner. 94 prosent av forskningsbevilgningene fra Higher Education Founding Council til universitetene fordeles på basis av den oppnådde karakteren. I Nederland har universitetssektoren iverksatt sitt egen kvalitetsvurderingssystem som ikke er direkte knyttet opp mot ressurstildeling. Det er fagfellekomiteer, oppnevnt av det Nederlandske universitets-rådet, som vurderer kvaliteten på bakgrunn av data/dokumentasjon fra instituttene, intervjuer og eventuelt «site visits». Regjeringen har forpliktet seg til å ikke endre bevilgninger før

[23] Blant de ulike formene for fagfellevurdering har vurdering av manus for publisering en særstilling som forskningsintern fagfellevurdering. Vurderingen av manus kan bli en del av selve forskningsprosessen idet evalueringsprosessen som regel søker å forbedre aktuelle manus før publisering.

institusjonen har hatt en reell mulighet til forbedring over 3-5 år. Både i UK og Nederland blir instituttene rangert i forhold til hverandre ved en femdelt karakterskala. I UK gis instituttene bare en samlet karakter, i Nederland får de karakterer og skriftlige kommentarer langs fire dimensjoner[24], men ingen samlet karakter (Hansen og Jørgensen 1996).

I Norge, og Norden for øvrig, har ikke evalueringer av forskning på instituttnivå og nasjonalt nivå vært standardisert på denne måten. Evalueringsrapportene har ikke vært av en slik art at instituttene eller gruppene som blir evaluert blir rangert på noen karakterskala. Evalueringene har i hovedsak vært enkeltstående, ofte *ad hoc* (ref. Brofoss 1997, Christiansen og Christiansen 1989). Mer institusjonaliserte eller «rutinemessige» praksiser, med muligheter for å knyttes opp mot bevilgninger, finnes imidlertid også eller er under utvikling, blant annet for instituttsektoren i Norge. Under skal vi se på karakteristika ved evalueringer som ikke er klart knyttet opp mot bevilgninger – slike evalueringer har utgjort hovedtyngden av faglige evalueringer på meso- og makronivå i Norge.

3.2 Hvordan makro- og mesoevalueringer skiller seg fra mikroevalueringer

I Norge har vi sett flere eksempler på evalueringsrapporter som er mer beskrivende enn evaluerende og som er svært forsiktige i sine konklusjoner. I tillegg til at det ikke blir krevet av evalueringskomiteene at de rangerer eller gir karakterer, og at evalueringene ikke er knyttet opp mot spesifikke beslutninger eller ressurstildelinger (slik mikro-evalueringer alltid er), er det en rekke andre forhold som også tilsier at makro- og mesoevalueringer av forskningskvalitet kan bli vage og innholdsløse. Stikkord er evalueringsmaterialets omfang, evaluatorenes kompetansebegrensninger, kompromisser mellom ulike faglige ståsted og rapportens offentlighet.

Når hele institutt eller fagfelt skal evalueres innebærer det at evalueringen av enkeltprosjekt nødvendigvis blir temmelig overfladisk i forhold til hva som er mulig ved vurdering av et manus eller en forsker. Dertil kommer den faglige bredden på virksomheten som skal evalueres. Selv om man tar sikte på å dekke alle de aktuelle fagområdene når man oppnevner evalueringskomiteen, vil «matchingen» av evaluator og evalueringsobjekt, nødvendigvis bli mye dårligere enn ved mikro-evalueringer. Evaluatorene blir da satt til å evaluere områder hvor de ikke kjenner all litteraturen på forskningsfronten og nyansene i fagdebatten, og bedømmer virksomheten ut fra mer generelle kriterier som relevans i forhold til disiplinen som helhet, om det er brukt allment anerkjente metoder etc. Slike forhold kan beskrives kort og generelt og uten å nevne enkeltprosjekt, og evalueringen blir mer overfladisk i forhold til den konkrete virksomheten som er gjenstand for evaluering. Når evaluatorene beveger seg på et slikt generelt plan, blir resultatet gjerne også mer beskrivende enn evaluerende.

[24] De fire dimensjonene er: vitenskapelig kvalitet, vitenskapelig produktivitet, vitenskapelig relevans og langsiktig levedyktighet i lys av nasjonale og internasjonale konkurranseforhold (Hansen og Jørgensen 1996:40).

En annen årsak til vaghet i evalueringsrapporter kan være uenighet mellom evaluatorene. Det blir som regel forventet at rapportene skal være enstemmige, og ved uenighet blir et (eksplisitt eller stilltiende) forhandlingskompromiss lett løsningen. At hver evaluator får gjennomslag for sin vurdering av den virksomheten som står ham/henne nærmest og at eventuelle kommentarer fra mer kritiske medlemmer av evalueringsgruppen utelates, vil ofte være et tilfredsstillende forhandlingsresultat for alle parter. Slike kompromisser mellom de ulike faglige ståsted som er representert i evalueringsgruppen kan være en sentral kilde til vage og «ufarlige» evalueringer. «Kompromissrapporter» av denne typen kan også bidra til å avdempe faglige kontroverser generelt. Når alle evaluerte får en bedømmelse på sine egne premisser, hindrer man åpne kontroverser som følge av at noen føler seg urettferdig behandlet. Vage rapporter gir også større rom for forhandlinger om tolkningen av rapportens innhold.

At evalueringsrapportene er offentlige kan også bidra til at bedømmelsene blir vage og «ufarlige». Evaluatorene foretar som regel intervjuer i de miljøene som skal evalueres, og evalueringsrapporten er knyttet til evaluatorene som personer, ved at de står som forfattere av rapporten. Samtidig er rapportene som regel offentlig tilgjengelige. Dette er helt forskjellig fra ikke-offentlige mikro-evalueringer hvor evaluatorene er anonyme og de ofte heller ikke kjenner de evaluertes identitet. Makro- og mesoevalueringer gir en kontekst hvor lojalitets- og legitimitets-hensyn lettere kan påvirke bedømmelsene. Kritikk av identifiserbare forskere i offentlige rapporter kan karakteriseres som «offentlig henging» og ha negative sosiale og psykologiske konsekvenser (Luukkonen 1995).

Når evaluator og evaluert kjenner hverandres identitet, og møtes for å diskutere den evaluertes arbeid og rammebetingelser, blir lojalitetsbånd mer åpenbare enn ved «dobbelt blind» fagfellevurdering. I en slik kontekst vil det være vanskeligere for evaluatorene å skrive rett ut hva de mener. Å kritisere fagfeller i en offentlig rapport vil også kunne oppfattes som brudd på kollegiale normer. Når slik kritikk i tillegg kan bidra til å true ressurssituasjonen til fagfeller vil evaluatorene ofte være forsiktige med hvordan de formulerer seg.

Legitimitets-hensyn peker i samme retning. Når de evaluerte kjenner evaluatorenes identitet, har evaluatorene større grunn til å produsere en evaluering som kan stå for enhver kritikk, enn når de er anonyme. Å sikre at en evaluering oppfattes som legitim kan være vanskelig hvis det ikke finnes noe uangripelig grunnlag å basere bedømmelsene sine på. Vurdering av faglig kvalitet er som nevnt dels basert på implisitte standarder og faglig skjønn. Så sant evaluatoren ikke har en ubestridt autoritet på feltet, må han/hun regne med at de evaluerte kan finne en grunn til å forkaste evalueringsrapporter de ikke liker. En evalueringsrapport som ikke har legitimitet i de berørte fagmiljøene, vil bli et sårbart fagpolitisk redskap. Evalueringskomiteer har derfor dobbelt grunn til å bruke diplomatisk språk og uttale seg svært forsiktig om negative aspekter. De unngår den personlige belastningen det er å få sine vurderinger «slaktet» og skaffe seg fiender, samtidig som de bidrar til et godt samarbeidsklima for

behandlingen av rapporten. Som regel kan imidlertid de negative aspektene leses mellom linjene.[25]

3.3 Kritikk av makro- og mesoevalueringer

Det finnes en del litteratur som hevder at den type evalueringsaktivitet som har vokst frem i de seneste tiår, har en rekke symbolske funksjoner og generelt har liten innvirkning på den øvrige aktiviteten i fagmiljøene og forskningsrådene. En studie av effekter og bruk av 101 nordiske evalueringer av forskning konkluderte med at evalueringsrapporter først og fremst blir brukt til å legitimere forskningsbevilgninger og som ammunisjon i konkrete saker. Slik bruk av rapportene til å understøtte egne interesser kan også til en viss grad ha effekter på beslutningene som fattes. Oppfølging av anbefalingene i rapportene krever imidlertid en del forutsetninger som sjelden er til stede. Det kreves at evalueringen gir svar på et informasjons-behov, at den er metodologisk troverdig, at den blir effektivt formidlet til beslutningstakere, at den gir ammunisjon til støttespillere i aktuelle saker og at det finnes innflytelsesrike støttespillere, mener forfatterne (Luukkonen og Ståhle 1990, ref. også Brofoss 1997:72-74).

Det hevdes altså at på grunn av svakheter ved evalueringene, formidlingen av dem og mangel på sentrale aktører som ønsker å engasjere seg (evalueringen går kanskje på tvers av deres interesser), har evalueringene få eksplisitte instrumentelle funksjoner. De flytter sjelden på ressurser, men kan ha andre og mer symbolske eller politiske funksjoner, som f.eks. å dekke over svakheter og feil for å beskytte evalueringsobjektet («whitewash»), rituell handling som inngår som et standardkrav til organisasjonens virksomhet («posture») eller å utsette beslutninger («postponement») (Larsen 1985:217-218).[26]

En intervjustudie av 90 forskere som hadde vært gjenstand for ulike nordiske disiplin-evalueringer konkluderer med at evalueringene hadde minimalt med konkrete konsekvenser. De evaluerte mente imidlertid at evalueringene hadde egenverdi for forskningen, ga oppmuntring, bidro til selvrefleksjon etc., og flertallet mente at man bør utføre liknende evalueringer også i fremtiden (Luukkonen 1995).

Går vi over fra litteraturen på feltet og hører på hva aktørene på feltet hevder i ikke offisielle sammenhenger, er det ikke vanskelig å finne påstander om begrensninger ved fagevaluering som er adskillig krassere enn de vi finner på trykk. Det påstås både at oppdragsgiver får den konklusjonen han ønsker (fordi han velger ut evaluatorer etter hvilken konklusjon han ønsker) og at de som skal evalueres får vridd hodene på de som evaluerer dit de vil (at «site

[25] Det finnes også flere eksempler på rapporter hvor negative bedømninger fremstilles meget eksplisitt. Slike rapporter kan skape livlig debatt om evaluatorenes kompetanse og innfallsvinkel. «Evaluering av arbeidslivs- og arbeidsmiljøforskningen i Norge» (*NORAS evalueringsrapport 1/92)*, og den etterfølgende debatt i *Tidsskrift for samfunnsforskning* og *Forskningspolitikk* er et eksempel på dette.

[26] Fokus er her på strategisk og symbolsk bruk. Andre former er instrumentell bruk og opplysende bruk (Naustdalslid og Reitan 1994).

visits» er en effektiv kanal for de evaluerte til å få formidlet sine ønsker og ressursbehov til bevilgende myndigheter – med underskrift fra et eksternt ekspertpanel).

Hvis begge påstandene skal ha noe i seg, samtidig som de er åpenbart motstridende, må de nødvendigvis nyanseres. Det er mange parter involvert i en fagevaluering, og evaluatorene står under press fra minst to parter, samtidig som de prøver å gjøre det beste ut av det (Van der Meulen 1997). De vil ikke tråkke noen på tærne eller ødelegge egne fremtidige samarbeidsmuligheter, samtidig som de må gi oppdragsgiveren et anstendig svar på spørsmålene som er stilt i mandatet for evalueringen. I tillegg kommer at evaluatorene forholder seg til ulike sett av faginterne normer og tradisjoner for evaluering – de har ulike faglige ståsted og hvert sitt personlige og «tause» grunnlag for å bedømme faglig kvalitet. I den grad det er «bias» i evalueringene kan altså retningen på den variere.

3.4 «Teigdeling»

Uenighet både om hvem som er autoriteter innen et gitt område og om avgrensingen av et gitt fagfelt, kan problematisere valg av evaluatorer. Ved makro- og mesoevalueringer er et tilleggsproblem å finne fagfolk som har bred nok kompetanse til å evaluere hele institutt og fagområder. Jo bredere og mer heterogent område som skal dekkes, jo viktigere *og* vanskeligere blir det å plukke ut de «riktige» evaluatorene. Man ønsker at evaluatorene skal ha en viss distanse til evalueringsobjektet, at komiteen skal ha den «riktige» balansen mellom ulike faglige ståsted, og at evaluatorene har tilstrekkelig faglig bredde og autoritet. Også andre typer krav, som å ha begge kjønn representert, skal tilfredsstilles. Den faglige bredden hos den enkelte evaluator kan være et av de kravene som er vanskeligst å ivareta. Ofte ender man opp med en komité som i liten grad har overlappende kompetanse, og man får en klar teigdeling i komiteen – der hver evaluator får «monopol» på det området han kan best (ref. Hansen og Jørgensen 1995). Kanskje dekker ikke hver ekspert mer enn en lite brøkdel av det området som evalueringsobjektet omfatter. Skulle man da unngå komiteer med én ekspert på hver «teig», får komiteene svært mange medlemmer.

Å sikre at all forskningen som skal evalueres dekkes av like god ekspertise kan ses som et mer prekært problem enn å skaffe overlappende ekspertise. Enhver ekspertgruppe vil ha *begrenset* ekspertise, den vil ikke ha like god kunnskap om alle typer forsking innen et bredt definert fagområde (Luukkonen 1991). Ifølge Chubin og Hackett (1990:80) bidrar økende spesialisering sammen med økt vekt på tverrfaglighet til at gruppen av folk som er kompetente til å evaluere et gitt spesialområde, blir mindre og mindre. Hver evaluator har spesialkompetanse innen stadig mindre deler av det som defineres som et relevant evalueringsobjekt.

Stilltiende forhandlinger og kompromisser er nærliggende i heterogene grupper. At gruppene er heterogene bidrar til at det er vanskeligere å formidle grunnlaget for sine vurderinger til resten av gruppen. Teigdeling og stilltiende kompromisser er enklest i en slik situasjon. Tidsbegrensninger og ønske om et godt arbeidsklima i gruppene tilsier også stilltiende

forhandlinger og kompromisser. Åpen konfrontasjon tar både tid og skaper et dårlig arbeidsklima. Det er mye enklere om man har en stilltiende regel om å stole på hverandres vurderinger og ikke fremme synspunkter man antar vil være i strid med flertallet. En slik teigdeling og «konfrontasjonsvegring» kan også føre til en form for falsk konsensus – man tror at man er enige, men en mer åpen og grundig diskusjon ville avdekket uenighet (Janis 1982).

3.5 Grunnlaget for makro- og mesovurderinger

Mandatene til forskningsevalueringer på miljø-, program- eller instituttnivå utført av fagfeller omfatter gjerne både vurdering av utført forsknings kvalitet, enhetenes rammebetingelser og å gi anbefalinger for fremtidig forskning. Evalueringsutvalgene har et bredt spekter av *informasjonskilder* til rådighet: evaluatorenes forhåndskunnskap om forskningen som skal evalueres, gjennomgang av forskningspublikasjonene, besøk/intervjuer ved de berørte institusjonene («site visits») og bakgrunnsdokumenter fra miljøene/instituttene. Det er også vanlig at komiteen eller forskningsrådet sender en liste med spørsmål til institusjonene/-miljøene som skal besvares før komiteen kommer på besøk.

Fagfelle-evaluatorer er gjerne opptatt av å være konstruktive og forsiktige. De ønsker å hjelpe forskningsmiljøene som blir evaluert til å forbedre seg, og når de skriver evalueringsrapporten legger de vekt på å få fram de positive sidene og hvordan noe kan bli bedre. Som nevnt må man ofte lese mellom linjene for å finne det negative. Noen intervjuede evaluatorer sier eksplisitt at dette er for å verne de «svake», hindre at de fratas midler og unngå personlige «tragedier». En annen grunn er usikkerheten som er forbundet med evalueringene. I tillegg til at evaluatorene ofte er klar over at andre vil være uenige i deres vurderinger og at de derfor synes de bør uttale seg forsiktig, mener ofte evaluatorer som foretar makro- og meso-evalueringer at de ikke har grunnlag til å felle noen endelig «dom» over de enkelte forskerne eller forskergruppene. Til det har de ikke nok informasjon om den enkeltes kompetanse og arbeid. Evaluatorene ser dessuten både svake og sterke grupper/forskere i hvert miljø/institutt, og synes derfor også at det er vanskelig å felle noen generell dom på miljø-/instituttnivå (Langfeldt, under arbeid).

Som nevnt baserer meso- og makro-evalueringer seg ofte på mer generelle kriterier enn mikro-evalueringer. Uten grundig gjennomlesning av publikasjonene og detaljkjennskap til de ulike forskningsfeltene, blir selve kvalitetsvurderingen mer overfladisk, basert på for eksempel hvor viktig forskningen anses å være for fagfeltet som helhet, eller om metodene er alminnelig anerkjente. Det kan også benyttes mer indirekte kvalitetskriterier, som hvor forskningen er publisert, antall siteringer eller forskningsmiljøenes/forskernes generelle vitenskapelige omdømme. I noen sammenhenger kan kvantitative indikatorer være en vesentlig del av informasjonen som evalueringskomiteer benytter seg av (Hansen og Jørgensen 1995).

Det kan imidlertid være vanskelig å svare på hvorvidt evalueringskomiteer som forventes å utføre «ren» *fagfellevurdering*, hovedsakelig benytter direkte eller indirekte kriterier på forskningskvalitet. Dels kan evaluatorer mene at de baserer seg på direkte vurderinger selv om vurderingene er betydelig påvirket av hva evaluatorene vet om andres bedømmelser, siteringshyppighet eller generelt faglig omdømme, dels kan slike indirekte bedømmelser være en del av praksis som man ikke ønsker å kommunisere videre (for eksempel i evalueringsrapporten eller til utenforstående). Norske intervjudata tyder på at evalueringskomiteer benytter et bredt sett av kilder og kriterier for å vurdere gruppers og institutters utførte forskning og fremtidige potensial – fra direkte vurdering av publikasjoners soliditet og originalitet til vurdering av motivasjon, pågangsmot og av gruppekjemi eller instituttklima (på bakgrunn av «site visits») og hvor man publiserer, generelt/internasjonalt omdømme og hva ens nære kolleger på feltet mener. Kriteriene og kildene varierte både innen og mellom komiteer. Evaluatorene ga også uttrykk for ulike syn på bruk av indirekte bedømmelser. Noen vektla nytten av å høre hva kolleger (utenfor evalueringskomiteen) mener om forskere på felt der de selv ikke er spesialister eller fremlegger renommé som et viktig kriterium, mens andre mente det ville være galt å la andre influere på vurderingene og at man slik risikerer å forsterke (muligens grunnløse) fordommer. Dataene inkluderer også noen evaluatorer med bakgrunn i anvendt forskning eller «brukersiden». Disse var mer tilbøyelige til å bruke indirekte kriterier enn det evaluatorer med bakgrunn i grunnforskning var (Langfeldt, under arbeid).

4 Fagfellevurderingers begrensninger

4.1 Begrensninger ved fagfellevurdering på mikronivå

Mye av denne rapporten har dreid seg om muligheten for «bias» i fagfellevurderinger. Kritikk av fagfellevurdering går også på en rekke andre forhold. Eksempelvis omfatter kritikken mot prosjektvurderinger i forskningsråd, foruten muligheten for «bias», at prosessen er ugjennomsiktig, at det er innebygde interessekonflikter i systemet, at evaluatorene har mulighet for å misbruke prosjektinformasjonen de får, at blant annet unge forskere vil komme dårlig ut når tidligere meritter tillegges stor vekt, at systemet er svært kostnadskrevende for alle parter, at det er vanskelig å forutsi hvilke prosjekter som vil lykkes og hvilke mislykkes og at det er problemer knyttet til seleksjon og kontroll av evaluatorer (Wood 1997:17). Flere av disse punktene er imidlertid også relatert til mulighetene for partiskhet. Seleksjon og kontroll, ugjennomsiktighet, interessekonflikter og misbruk av informasjon er på ulike vis knyttet til habilitetsproblematikken i fagfellevurderings-prosesser.

En annen innvendig mot prosjektvurdering er at den har en tendens til å være konservativ og risikoadvers. Dette er en form for «cognitive bias». Med mange søknader og lite midler å fordele er det de «sikre» prosjektene og ikke de med for eksempel utradisjonelle tilnærminger, som har størst sjanser til å nå gjennom nåløyet når etablerte fagfeller er satt til å vurdere blant annet muligheten for å gjennomføre prosjektet. En survey blant forskere som søkte om midler fra NIH fant at 60 prosent mente at «Reviewers are reluctant to support unorthodox or high-risk research» (Chubin og Hackett 1990:66).

Uansett om det dreier seg om ulike former for «bias» eller prosesskritikk knyttet til habilitetsspørsmål så er det muligheter for skjevheter og partiskhet i *utfallet* av fagfellevurderinger som er bakgrunnen for kritikken. Vi har imidlertid sett at begrepsinnholdet når det snakkes om «skjevheter» eller «partiskhet» er uklart. Det er langt fra gitt hva upartiskhet og likebehandling ville innebære når vurderinger er forutsatt å bygge på faglig skjønn og kriteriene kan være vage og implisitte. Både «bias» på grunn av at evaluatorene har et bestemt faglig utgangspunkt og på grunn av personlige eller profesjonelle interesser eller fordommer kan være underforstått, ubevisst eller på andre måter fortiet. Når en vesentlig del av grunnlaget for evalueringer ikke drøftes eksplisitt eller (kan) redegjøres for, vil det være vanskelig å avgjøre om evalueringer - både egne og andres - er upartiske og ikke-diskriminerende. Det finnes ikke klare kriterier som kan avgjøre slike spørsmål.

Det kan også hevdes at fagfellevurdering i seg selv innebærer en bevisst og ønsket faglig «bias». Fagfellevurdering baserer seg på etablerte meninger om hva som er adekvat og verdifullt, og dette vil nødvendigvis innebære en innebygget faglig partiskhet og diskriminering av faglig virksomhet som bryter med de etablerte normer og teorier i systemet (ref. Kuhn 1970). Slike konservative effekter kan imidlertid også motarbeides, f.eks. ved

egne tiltak for ukonvensjonell forskning og ved å bruke evaluatorer som er spesielt åpne og tolerante overfor slik forskning.

Den mest åpenbare form for diskriminering er den som rammer på grunn av kjønn, manglende bekjentskaper eller manglende institusjonell status. Åpenbar betyr imidlertid ikke at den er enkel å avdekke. Som GAOs evaluering av forskningsråd i USA påpeker, kan en sammenheng mellom slike faktorer og forskningskvalitet ikke utelukkes (GAO 1994). Det er imidlertid mulig å motvirke slik diskriminering, blant annet ved å ikke gi evaluatorer av manus opplysninger om forfatteren, eller å gi utfallet for utsatte grupper særlig oppmerksomhet når bedømmelsen ikke kan foretas uten opplysninger om forskerens identitet. I fag der publiserings- og siteringsindikatorer er allment akseptert som gode mål på en forskers kompetanse, kan også slike mål settes opp mot utfallet av fagfellevurdering for å avdekke eventuell forskjellsbehandling (Wennerås og Wold 1997).

Oppsummert er det en rekke begrensinger ved fagfellevurdering. Vurderinger baserer seg på skjønn, ikke på klare kriterier. Fagfeller vurderer forskning ut fra ulike grunnlag og faglige ståsted, og hvem som foretar en vurdering kan derfor være avgjørende for utfallet. Ulike former for fagfellevurdering kan være utsatt for en rekke ulike former for «skjevheter» eller partiskhet, men det er vanskelig å avgjøre om vurderinger er upartiske eller ikke.

Samtidig må det understrekes at flere av begrensningene ved fagfellevurdering bunner i at det stilles ulike krav som kan gå på tvers av hverandre. Når validitet (riktig utfall) krever et bredt spekter av vurderinger, og muligens uenighet og diskusjon, kan man ikke samtidig bedømme prosessen ut fra snevre reliabilitetsmål (samme utfall uansett hvem som bedømmer) (Harnad 1985). Likeledes kan kostnadseffektivitet gå på bekostning av både validitet, reliabilitet, og habilitetshensyn (Chubin og Hackett 1990:46). Helt uten hensyn til kostnader ville man kunne innhente uttalelser fra et stort antall eksperter langt utenfor landets grenser, og slik bedre løse habilitetsproblematikken og samtidig sørge for en bredde i ekspertuttalelsene som både sikret validitet og reliabilitet – hvis et annet like stort og bredt utvalg av eksperter ble plukket ut, ville resultatet bli noenlunde det samme. Hvis dette skulle gjøres for alle mikronivå vurderinger ville kostnadene imidlertid stå i et sterkt misforhold til ressursene som skulle fordeles ved hjelp av vurderingene. Forskersamfunnet ville dessuten bli så nedtynget i å skrive uttalelser at det knapt ville være tid tilbake til å forske.

4.2 Begrensninger ved fagfellevurdering på makro- og mesonivå

I det store og hele har meso- og makro-evalueringer de samme begrensningene som prosjektvurderinger. Kriteriene er uklare, utfallet kan avhenge av hvem som bedømmer og vurderingene kan på ulike måter være partiske. Partiskhet kan imidlertid ta en spesiell form her. En rekke ulike faktorer kan bidra til at evaluatorene sympatiserer med de evaluerte og ønsker å beskytte eller hjelpe dem. Evalueringsrapportene er offentlige og lojalitetshensyn og kollegiale normer bidrar til at man ofte prøver å formulere seg i positive vendinger.

Samtidig har slike evalueringer skilt seg fra mikro-vuderinger ved at det ikke er et knapt gode som skal fordeles, og at det ikke kreves rangering av evalueringsobjektene. Problemet med at fagfolk vurderer forskning ulikt, blir mindre prekært når evalueringen ikke innebærer karaktergiving, rangering og ressursfordeling.

Når evalueringskomiteer sammensatt av eksperter med ulik bakgrunn, tar for seg hele program, institusjoner eller fagfelt, og i mindre grad den enkelte forsker eller prosjekt, heves evalueringen opp på et *fagpolitisk nivå* som man i utgangspunktet skulle tro ville bidra til debatt og synliggjøring av de ulike faglige ståsted, implisitte kriterier og enkelte av de tause forutsetningene for vurderingene. Slik er det imidlertid ikke. Evalueringsrapportene diskuterer i liten grad kriterier og grunnlag for vurderingene. Riktignok blir både metoder og generelle kriterier ofte gjort rede for, og forsvart, men selve grunnlaget for vurderingene blir sjelden problematisert. Makro- og meso evalueringer kan derfor virke «ugjennomsiktige» på samme måte som mikro-evalueringer. De evaluerte får imidlertid vite *hvem* som har foretatt bedømmelsene, og gitt at de også kjenner deres faglige bakgrunn vil de ha en formening om de har fått en «rettferdig» bedømmelse eller ikke.

De begrensningene som kommer i tillegg ved vurderinger på meso- og makronivå er særlig knyttet til evalueringsmaterialets omfang. Det er vanskeligere å finne evaluatorer med et bredt nok kompetansegrunnlag til å vurdere hele forskningsprogrammer, institutter eller fagfelt. De faglige vurderingene kan derfor bli generelle og overfladiske i forhold til hva som er mulig ved vurderinger på mikronivå. De kan også i høyere grad enn på mikronivå være basert på indirekte kriterier – eksempelvis generelt renommé og hvor man har publisert. «Teigdeling» i heterogene evalueringskomiteer bidrar til at grunnlaget for vurderinger ikke drøftes, og at den som er «faglig nærmest» får det siste ordet i bedømmelsen av et miljø.

4.3 Hvorfor bruke fagfellevurdering?

Vi har sett at det er klare begrensinger ved fagfellevurderinger uansett om evalueringsobjektet er enkeltforskere/enkeltmanus eller forskningsinstitutt, programmer eller fagfelt. Når fagfellevurdering likevel er det dominerende evalueringskonseptet i forskersamfunnet er det på grunn av en helt avgjørende styrke. Vi får ingen *fagkyndig* vurdering uten at fagfeller deltar i vurderingen. Fagkompetanse er påkrevet for bedømmelse av faglig kvalitet på forskning, både av forskningens soliditet og av faglig verdi og originalitet. Velger man bort fagfeller som evaluatorer velger man samtidig bort en vurdering av disse aspektene.

Også kvantitative indikatorer som publiseringsindekser og siteringsindekser er i siste instans basert på vurderinger foretatt av fagfeller, blant annet fagfellevurdering av manus. Flere av svakhetene ved fagfellevurdering kan derfor ikke unngås ved å basere seg på slike indikatorer. Hvis fagfellevurdering for eksempel bidrar til konservatisme, kumulative fordeler

eller ivaretagelse av «old boys»-nettverk, forsterker bruk av kvantitative publiseringsindikatorer tvert imot slike tendenser (Niiniluoto 1987:19).

Begrensningene ved fagfellevurdering kan minimeres når man er oppmerksom på dem. «Kognitiv partiskhet» kan begrenses for eksempel ved å sikre stor bredde i fagekspertisen som brukes eller hyppig utskifting av evaluatorer. Forskjellsbehandling kan unngås ved å holde de evaluertes identitet skult for evaluatorene. «Teigdeling» og «bias» i evalueringskomiteer bør kunne reduseres ved å oppnevne komitémedlemmer med forutsetninger for å krysskontrollere hverandre (Hansen og Jørgensen 1996). Det kan imidlertid være dyrt å ivareta reliabilitet, validitet og habilitetshensyn. Som nevnt er et stort og bredt utvalg av eksperter det ideelle ut fra slike hensyn. Å sikre seg mot tilfeldigheter, skjevheter og partiskhet i vurderingsutfall kan koste uforholdsmessig mye sammenlignet med de midlene som skal fordeles eller den tilleggsverdien som en mer sikker og legitim vurdering gir.

Referanser

Ajenstat, Jacques (1993): »Empirical test of a computer-based support system for the evaluation of research grant proposals» *Research Evaluation, 3:68-74.*

Aubert, K. E. (1989): «Ideologisk vitenskap» *Nytt Norsk Tidsskrift, 6:233-247.*

Bozeman, Barry (1993): "Peer review and evaluation of R&D impacts." I Barry Bozeman and Julia Melkers (red.) *Evaluating R&D impacts: methods and practice.* Boston: Kluwer Academic Publishers.

Brofoss, K. E. (1997). *Metaevalueringen. En gjennomgang av Norges forskningsråds evalueringspraksis.* Oslo, Norges forskningsråd.

Burnham, J. C. (1992): «How Journal Editors came to Develop and Critique Peer Review Procedures» I H. F. Mayland & R. E. Sojka (red.) *Research Ethics, Manuscript Review and Journal Quality.* Madison: ACS Miscellaneous Publication.

Ceci, Stephen J. og Douglas P. Peters (1982): «Peer-review practices of psychological journals: The fate of published articles, submitted again» *The Behavioral and Brain Sciences 5:187-255.*

Ceci, Stephen J. og Douglas P. Peters (1982b): "Peer Review: A study of reliability." *Change, 14:44-48.*

Christiansen, John og Lene Christiansen (1989): "Research on Research: Evaluation of Evaluations in the Nordic Countries." *COS Forskningsrapport 3/89.* Copenhagen.

Chubin, Daryl E. og Edward J. Hackett, (1990): *Peerless Science.* New York: State University of New York Press.

Cicchetti, Domenic V. (1991): "The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation" *The Behavioral and Brain Sciences 14:119-186.*

Cole, Stephen (1983): "The Hierarchy of the Sciences?" *American Journal of Sociology 89:111-139.*

Cole, Stephen, Jonathan R. Cole and Gary A. Simon (1981): "Chance and Consensus in Peer Review." *Science, 214: 881-886.*

Cole, Stephen, Leonard Rubin and Jonathan R. Cole (1978): *Peer Review in the National Science Fondation. Phase one of a study.* Washington D.C.: National Academy of Science.

Cole J. R., og S. Cole (1985): «Experts' «Consensus» and Decision-Making at the National Science Foundation» I K. S. Warren (red.) *Selectivity in Information Systems. Survival of the Fittest.* New York: Praeger.

Daniel, Hans-Dieter (1993): *Guardians of Science. Fairness and Reliability of Peer Review.* Weinheim: VCH.

Fisher M., S. B. Friedman og B. Strauss (1994): «The Effects of Blinding on Acceptance of Research Papers by Peer Review» *JAMA 272:143-146.*

Fürst, Elisabeth (1988): *Kvinner i Akademia - inntrengere i en mannskultur?* Oslo: NAVFs sekretariat for kvinneforskning.

GAO (1994): *Peer Review. Reforms Needed to Ensure Fairness in Federal Grant Selection.* United States General Accounting Office, GAO/PEMD-94-1.

Garfunkel, J. R., M. H. Ulshen, H. J. Hamrick og E. E. Lawson (1994): «Effect of Institutional Prestige on Reviewers' Recommendations and Editorial Decisions». *JAMA 272:137-138.*

Gilbert, J. R., E. S. Williams og G. D. Lundberg (1994): «Is There Gender Bias in JAMA's Peer Review process?» *JAMA 272:139-142.*

Gulbrandsen, M. og L. Langfeldt (1997): *Hva er forskningskvalitet? En intervjustudie blant norske forskere.* NIFU: Rapport 9/97.

Hansen, H. F. og B.H. Jørgensen (1995): *Styring af forskning: Kan forskningsindikatorer anvendes?* Frederiksberg: Samfundslitteratur.

Harnad, Steven (1985): «Rational Disagreement in Peer Review» *Science, Technology & Human Values, 10:55-62.*

Hull, D. L. (1988): *Science as a process.* Chicago/London: The University of Chicago Press

Janis, Irving L. (1982): *Groupthink. Psychological Studies of Policy Decisions and Fiascoes.* Boston: Houghton Mifflin Company.

Kuhn, T. S. (1970): *The Structure of Scientific Revolutions.* Chicago: The University of Chicago Press.

Laband, D. N. og M. J. Piette (1994): «A Citation Analysis of the Impact of Blinded Peer Review» *JAMA 272:147-149.*

Langfeldt, L. (1998): *Fagfellevurdering som forskningspolitisk virkemiddel. En studie av fordelingen av frie midler i Norges forskningsråd.* NIFU: Rapport 12/98.

Langfeldt, L. (under arbeid): «Peer Evaluation as Collective Decision-Making. A Study of Six Expert Committees» Oslo: NIFU.

Larsen, Bøje (1985): "Forskningsevaluering - Problemer og muligheder." In Egil Fivesdal (ed.): *Nærbilleder af forskning: Organisasjonssociologiske studier*. Copenhagen: Nyt fra Samfundsvitenskaberne.

Lindsey, Duncan (1988): «Assessing precision in the manuscript review process: A little better than a dice roll» *Scientometrics* 14:75-82.

Luukkonen, Tertu (1995): «The impacts of research field evaluations on research practice» *Research Policy* 24:349-365.

Luukkonen, Tertu (1991): «Citation indicators and peer review: their time-scales, criteria, and biases» *Research Evaluation* 1:21-31.

Luukkonen, Tertu og Ståhle, Bertel (1990): "Quality evaluations in the management of basic and applied research" *Research Policy* 19:357-368.

Mahoney, Michael J. (1977): "Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System" *Cognitive Therapy and Research, 1:161-175.*

Mazuzan, George T. (1992): "Good Science Gets Funded... The Historical Evolution of Grant Making at the National Science Foundation" *Knowledge, 14:63-90.*

Merton, Robert K. (1968): "The Mattew Effect in Science" *Science 159:56-63.*

Merton, Robert K. (1988): "The Mattew Effect in Science, II. Cumulative Advantage and the Symbolism of Intellectual Property" *ISIS 79:606-623.*

Montgomery, H. og S. Hemlin (1991): *Judging scientific quality. A cross-disciplinary investigation of professorial evaluation documents.* Göteborg Psychological Reports, Vol. 21, No. 4

Naustdalslid, J. og M. Reitan (1994): *Kunnskap og styring. Om bruk av forskning i politikk og forvaltning.* Oslo: Tano

Niiniluoto, Ilkka (1987): "Peer review: problems and prospects" I *Evaluation of Research.* Nordic Science Policy Council, FPR-publication No.5.

NORAS (1992): *Evaluering av arbeidslivs- og arbeidsmiljøforsknigen i Norge.* Oslo: NORAS evalueringsrapport 1/92.

Norges forskningsråd (1993): *Habilitet og tillit i Norges forskningsråd.* Rapport fra ekspertgruppe oppnevnt av Hovedstyret for Norges forskningsråd, september 1993.

Norges forskningsråd (1996): *Evalueringshåndbok for Norges forskningsråd.* Oslo: Norges forskningsråd.

NOU (1988): *Med viten og vilje.* Oslo: Statens forvaltningstjeneste, NOU 1988:28.

Nylenna, M., P. Riis og Y. Karlsson (1994): «Multiple Blinded Reviews of the Same Two Manuscripts: Effects of Referee Characteristics and Publication Language» *JAMA 272:149-151.*

Pfeffer, Jeffrey, Gerald R. Salancik, and Huseyin Leblebici (1976): «The Effects of Uncertainty on the Use of Social Influence in Organizational Decision Making» *Administrative Science Quarterly* 21:227-245.

Ravetz, Jerome R. (1971): *Scientific Knowledge and its Social Problems.* Oxford: Clarendon Press.

Speck, B. W. (1993): *Publication Peer Review. An Annotated Bibliography.* Westport/London: Greenwood Press.

Travis G. D. L. og Harry M. Collins (1991): "New Light on Old Boys: Cognitive and Institutional Particularism in the Peer Review System" *Science, Technology & Human Values, 16:322-341.*

van de Kaa, D. J. (1993): «Picking the winners by consensus: Grant-giving practice in the Netherlands» I F. Q. Wood og V. L. Meed (red.) *Research Grants Maagement and Funding.* Canberra: Bibliotech.

van der Meulen, B. J. R. (1997): «The use of S&T indicators in science policy: Dutch experiences and theoretical perspectives from policy analysis» *Scientometrics 38:87-101.*

Wennerås, C. og A. Wold (1997): «Nepotism and sexism in peer review» *Nature 387, 22 May:341-343.*

Wood, F. Q. (1997): *The Peer Review Process.* Australian Research Council. National Board of Employment, Education and Training. Commissioned Report No. 54.

Wood, F. Q. (1995): *Issues and Problems in the Public Funding of University Basic Research.* Thesis, University of New England.

Zuckerman, Harriet og Robert K. Merton (1971): "Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System" *Minerva 9:66-100.*