Nya avhandlingar: "Öppna rum - om ungdomarna, staden och det offentliga livet", Björn Anderssons doktorsavhandling vid institutionen för socialt arbete, Göteborgs universitet 2002

Persson, Anders

Published in:
Socialvetenskaplig tidskrift

2002

Från redaktionen

Det finns flera skäl till att redaktionen beslutat att uppdra åt professor Haluk Soydan att göra detta temanummer. Både internationellt och i Norden har utvärderingsforskningen kommit att bli ett uppmärksammat forskningsområde. Från praktikerhåll har det riktats krav på att utvärderingsforskningen bör komma till nytta för att öka kvalitén i olika verksamheter som sjukvård och socialtjänst.

Att skapa en översikt över forskningsfältet kan därför ses som en angelägen uppgift. Soydan, som är verksam vid Centrum för utvärdering av socialt arbete och Stockholms universitet, har därför samlat ett antal författare för att kunna spegla den internationella utvecklingen i utvärderingsforskningen. Förhoppningsvis kan numret också ge näring till debatten om kunskapsbaserat socialt arbete. Naturligtvis kan inte alla riktningar täckas i ett enskilt tidskriftsnummer och det finns perspektiv som inte belysts. Vi vill därför fortsätta att publicera bidrag inom ämnesområdet och tar gärna emot artiklar.

Socialvetenskaplig tidskrift är en svenskspråkig tidskrift som vänder sig till en bred publik. Några av artiklarna i detta nummer är på engelska. För att behålla den svenskspråkiga karaktären på tidskriften finns artiklarna översatta till svenska och utlagda på vår hemsida: www.forsa.nu. Klicka på SVT och artikelförteckning.

Som vanligt presenterar vi också några recensioner av nya doktorsavhandlingar och intressanta böcker.

Vi vill till slut rikta ett stort tack till Haluk Soydan för det förtjänstfulla arbete som han lagt ner på detta nummer.

Redaktionen

Socialt arbete och utvärdering

Ungefär samtidigt som jag planerade detta temanummer av Socialvetenskaplig tidskrift läste jag om dramatiska förhållanden, framkallade av ett hormonbaserat preparat, diethylstilboestrol (DES) som använts för att motverka missfall hos kvinnor. Under 1940--60-talen har uppskattningsvis tre till sex miljoner kvinnor ordinerats detta preparat. I vissa länder användes det fram till 1980-talets första år. I de fem första av åtta studier som genomfördes för att studera preparatets effektivitet användes en retrospektiv och matchad forskningsmetod där kvinnor som använt DES jämfördes med kvinnor som inte använt preparatet.

Samtliga studier med denna metod visade att DES var ett framgångsrikt preparat för förebyggandet av missfall. Chansen att fortsätta graviditeten och föda levande barn var dubbelt så hög bland behandlingsgruppen. Sextiofem procent av kvinnorna i behandlingsgruppen hade levande födda spädbarn medan andelen levande födda spädbarn i »kontrollgruppen« var trettiotvå procent. De tre övriga studier som genomfördes hade en prospektiv experimentell forskningsdesign. Kvinnor med missfallsrisk randomiserades till behandlings- respektive kontrollgrupper; randomiseringen baserades på kvinnornas läkares bedömning av osäkerheten i DES användningen.

Forskningsresultaten beträffande DES effektivitet var mycket annorlunda i dessa tre studier: det kunde inte hittas någon effekt vad gäller chansen att föda levande spädbarn. Senare har dessa studier fått stor betydelse när det började visa sig att döttrar till DES användare rapporterades utveckla en sällsynt variant av vaginal cancer. Såväl döttrar som söner till dessa kvinnor rapporterades dessutom utveckla andra allvarliga sjukdomar. Det kom också att visa sig att DES användarna själva hade ökad grad av bröstcancerfall. Det var uppenbart att DES inte bara var effektlöst mot missfall, utan hade också långsiktigt skadliga effekter (Oakley 2000, 308). Användningen av DES har avbrutits.

DES forskningen visar ett generellt problem: interventioner med goda syften kan ha icke önskade och skadliga effekter!

Det finns också exempel på sociala interventioner med skadliga effekter; en av de mest kända är kanske Cambridge-Somervillestudien om rådgivningsprogram för unga flickor och pojkar. Den upprepade uppföljningen av effekterna av rådgivning till ungdomar i riskzon visade att de som hade fått stöd av socialarbetare hade mer brottslighet än dem som inte hade exponerats för rådgivning (McCord 1978).

Studier av interventioner i socialt arbete är ett underutvecklat område (Macdonald 1998). Litteraturöversikter i tidskrifter för forskning i socialt arbete är mycket ofta narrativa, anger inte kriterier för inkludering och exkludering av litteratur, och ger samma vikt åt studier som genomförts med

olika forskningsdesigner eller som har olika vetenskaplig styrka.

Det urval av studier som presenteras i detta temanummer måste värderas mot den bakgrund som här antyds: att utvärdering av interventioner som är riktade mot människors hälsa och sociala välbefinnande kan ha livsviktig betydelse, och att utvärdering av interventioner i socialt arbete har just börjat sin långa vandring mot ett bättre tillstånd där åtminstone skadliga effekter av stora interventionsprogram gjorts kända.

I temanumret möts bidrag från två praktik- och kunskapstraditioner: nämligen utvärdering och socialt arbete. Typiskt har både traditionerna egenpraktik och gör anspråk på teoretiska (i bästa fall disciplinära) fält. Huvudansatsen i detta nummer är således att belysa frågor som samtidigt förenar insatser från både traditionerna.

Oavsett om man vill studera effekterna av interventioner i socialt arbete eller andra aspekter som till exempel interventioners förlopp och processer blir det nödvändigt att vända sig till utvärderingsforskningen för val av lämpliga utgångspunkter, teorier, och definitivt för val av metoder och metodologiskt stöd.

Utvärdering som praktik har sedan senare delen 1950-talet fått luft under vingarna. Huvudsakligen i USA har vissa händelser burit fram utvärderingsforskningen. Dessa inkluderar utvärdering av försvarssystem som lanserades i kapprustningen mot kommunismen och Sovjetunionen (1950-talet); lansering av nya lagar mot diskriminering och för positiv diskriminering av etniska minoriteter och handikappade samt det så kallade Great Society programmen för reducering av fattigdom och sociala problem (1960-talet); den federala statens strävan att säkerställa bättre användning av skattemedel som beviljats organisationer och institutioner (1970-talet); program för att höja USAs internationella kompetens inom olika verksamhetsområden (1980-talet). Och senast under 1990-talet såväl i USA som i andra delar av världen har utvärderingsforskningen kommit till nytta för att säkra och öka kvalitet, träffsäkerhet och rättvisa i service inom ett antal verksamhetsområden som hälsovård, socialtjänst och skola.

Reformer inom skolan har varit det första och under lång tid huvudsakliga studieobjektet för svensk utvärderingsverksamhet inom det sociala området (Franke-Wikberg och Lundgren 1980). Utöver utvärdering av svenska skolreformer har utvärdering som professionell verksamhet i Sverige varit begränsad, särskilt mätt i termer av det som sedan lång tid tillbaka pågick i USA. Statsvetaren Evert Vedungs bok Utvärdering i politik och förvaltning som först kom ut i början av 1990-talet var det första tunga bidrag som på ett allmänt plan gav ansikte till utvärderingsforskningen i Sverige (Vedung 1991 och på engelska Vedung 1997). Det första omfattande försöket att i Sverige relatera socialt arbete och modern utvärderingsforskning publicerades i Scandinavian Journal of Social Welfare år 1998.

Det bör i detta sammanhang nämnas att det finns en växande diskurs om vikten att basera praktiskt socialt arbete på empiriskt prövade kunskaper – ett förhållande som torde vara väl känt för många av denna tidskrifts läsare. Debatten förs till exempel inom ramen för den amerikanska

organisationen, Society for Social Work and Research (en motsvarighet till FORSA i Sverige och andra nordiska länder) och den växande Campbell Collaboration för utveckling och distribution av systematiska forskningssynteser bland annat i socialt arbete (http://www.campbellcollaboration.org). I Sverige har Socialstyrelsen nyligen utrett kärnfrågorna kring empiriskt baserat socialt arbete på uppdrag av regeringen (Socialstyrelsen 2000).

Urvalet av artiklar i föreliggande temanummer har gjorts mot bakgrund av den internationella utvecklingen i utvärderingsforskningen och debatten om kunskapsbaserat socialt arbete. På ett begränsat utrymme kan naturligtvis inte alla väsentliga teman täckas. De artiklar som nu presenteras på svenska har valts med ledning av tre kriterier: översikt, exemplifiering och utvärderingsforskningens spetsfrågor.

I ett temanummer av en tidskrift som annars bevakar ett brett empiriskt och teoretiskt fält som socialt arbete är det viktigt att ge läsaren en översiktlig bild av utvärdering som professionell verksamhet.

På ett sätt kan man hävda att flera artiklar tillsammans ger en översiktlig bild av fältet. Mer specifikt, artiklarna om utvärdering som disciplinär verksamhet och om utvärderingsmodeller fyller detta syfte. Framför allt finns en djungel av utvärderingsmodeller som alla betonar någon aspekt eller några aspekter av hur man bör utvärdera insatser och program. En övergripande sammanfattning av utvärderingsmodeller är ett utmärkt sätt att bilda sig en uppfattning av det som fältet erbjuder.

Flera artiklar i detta temanummer presenterar eller ger exempel på konkreta och empiriska utvärderingsstudier. Huvudavsikten med en presentation av empiriska studier har varit att försöka fånga in hur utvärderingsforskare gjort när de säger att de utvärderat en insats eller ett program. Exempel på empiriska studier utgör, enligt min mening, ett pedagogiskt sätt att utveckla sin egen utvärderingspraktik. Genom att studera andras studier kan man lära sig undvika misstag och förbättra sitt eget arbete.

I temanumret ingår även ett urval av artiklar som var och en presenterar det som jag uppfattar som utvärderingsforskningens spetsfrågor, internationellt sett. Det är viktigt att komma ihåg att detta bara är ett urval. Spetsfrågor som internationellt debatteras och forskas om är huvudsakligen av metodologisk karaktär snarare än paradigmatisk. De paradigmatiska frågorna återkommer oftast i (begränsade) debatten om utvärdering som disciplinär verksamhet eller i modelltänkandet vilket är mycket utbrett i utvärderingsforskningen. Spetsfrågorna återfinns huvudsakligen i metod- och metodologisk utveckling. Några exempel är frågor som berör datainsamling, effektmätning, validitet, itemvaliditet, systematiska forskningssynteser och forskningsanvändning.

Temanumret börjar med en artikel av Michael Scriven (USA) som uppfattar utvärderingsforskningen som en egen disciplin. Närmare bestämt är utvärderingsforskning en »transdisciplin« som statistik och matematik. Transdiscipliner är analytiska verktygsdiscipliner med egen plattform. Scriven har utvecklat dessa tankar under en längre tid men började först publicera dessa under 1990-talet.

Den föreliggande artikeln har tidigare varit publicerad i Scandinavian Journal of Social Welfare år 1998. Det är nu efter några år intressant att konstatera att Scrivens ansats fortfarande står oemotsagd, men också utan att någon på allvar följt spåren för vidareutveckling av ansatsen.

Evert Vedung (Sverige) presenterar en översiktlig bild av utvärderingsmodeller. Med en historisk blick typologiserar han utvärderingsmodeller efter grundläggande värdekriterier. Hans exempel kommer från olika policyområden inklusive naturvetenskapligt inriktade områden som skogsvård och miljöproblem. Även om exemplen inte är lånade från det sociala arbetets fält visar Vedungs exempel att det i princip är samma sorts designproblem och mätproblem som utvärderingsforskarna måste lösa inom olika interventions- och policyområden.

Ian Shaw (Storbritannien) har under senare år varit en framgångsrik förespråkare för användning av kvalitativa metoder i utvärdering av praktiskt socialt arbete. I sin artikel fortsätter han i samma bana för att argumentera för fördelarna med kvalitativa metoder när man utvärderar socialt arbete. Hans anspråk på kvalitativa metoders förmåga att hantera kausalitet är särskilt järvt. Han presenterar också ett konkret fall av utvärdering av socialt arbete i Storbritannien.

Med ökat intresse för studier av interventioners effekter har frågor kring utfall och utfallsmått åter blivit aktuella. Utfall är avsedda och icke-avsedda förändringar hos enheter till följd av påverkan från sociala interventioner. I denna artikel beskriver Haluk Soydan och Bo Vinnerljung (Sverige) några av de problem som de mött i utvärderingsverksamhet.

Av olika skäl – till exempel för att fånga in empiriska variationer eller att säkra ett större empiriskt material – har man särskilt i USA utvärderat interventioner eller program på flera orter (»sites«) parallellt. Edward Mullen (USA) som har många års erfarenhet av flerortsstudier presenterar en rik flora av utvärderingsstudier som genomförts i USA. Han utvecklar en intressant typologi som systematiserar olika former och typer av flerortsutvärderingar.

Under ett par decennier har avancerade metaanalyser genomförts som underlag för upplyst beslutsfattande i professionell praktik och policyutveckling. Metanalys är ett samlingsbegrepp för statistiska metoder för genomförande av systematiska forskningssynteser. Mark Lipsey (USA), som spelar en ledande roll i denna utveckling presenterar grunderna för metanalys. Han sammanfattar också metodologiska erfarenheter från ett stort antal metaanalyser. Dessa ökar kraftigt vår insikt i effektstudiers styrka och svagheter.

Anthony Petrosinos (USA) bidrag utgör ett kompletterande inslag till metanalyser. Petrosino följer upp tidigare spår och argumenterar för en integration av metaanalytiska ansatser med programteoriutvärdering. Medan den förra ansatsen avser systematisering av resultat från flera utvärderingar avser det senare studium av underliggande antaganden och mellanliggande variabler i ett interventionsprogram vilka anses bidra till förståelsen av varför programmet borde »fungera«.

Individer utgör den traditionella allokeringsenheten i randomiserade studier. Det kan finnas flera skäl, inte minst etiska, att

inte använda individer som allokeringsenhet. I denna artikel för Robert Boruch (USA), Ellen Foley (USA) och Jeremy Grimshaw (Canada) fram möjligheten att använda andra enheter i radomiserade studier. Sådana enheter kan vara geopolitiska »ställen« som grannskap, polisdistrikt, skolor såväl som administrativa enheter och grupper av individer. Denna typ av randomisering kallas också grupprandomisering eller klusterrandomisering. Författarna är mycket generösa med att beskriva flera utvärderingsstudier där klusterrandomisering använts.

Kari Jess och Siv Nyström (Sverige) sammanfattar erfarenheter av och resultat från en svensk kvasi-experimentel utvärderingsstudie, Krami-studien. I artikeln rapporterar de dels en klientstudie, dels en samhällsekonomisk studie av de studerade rehabiliteringsprogrammen. KrAmistudien är intressant inte bara på grund av sin för svenska förhållande sällsynta design, men också för de positiva effekter som Krami-programmet tycks ha på sina klienter.

Slutligen publicerar vi i temanumret personliga reflektioner av Bengt-Åke Armelius (Sverige) kring den historiska utvecklingen om hur psykoterapin successivt baserats på empiriskt prövade kunskaper. Med referens till debatten om möjligheterna att införa systematiska utvärderingar och empiriskt dokumenterade kunskaper som grund för åtgärder och beslut inom socialtjänsten känner han igen mycket av de argument och de känslor som uttrycks i den allmänna debatten från den tid då man inom den kliniska psykologin började med psykoterapiforskning.

Avslutningsvis några ord om språken i artiklarna. Ett flertal av artiklarna har skrivits av engelskspråkiga författare för att sedan översättas till svenska av utomstående översättare i Sverige. Redaktionsmedlemmar, redaktionssekreteraren och gästredaktören har satsat åtskilliga timmar för att granska såväl fackterminologin som svenskan i dessa artiklar. Trots detta är vi inte nöjda med översättningarna. Vi har valt att publicera fyra artiklar i originalspråket, engelska. Svenska översättningarna av dessa artiklar finns på tidskriftens hemsida, www.forsa.nu.

Stockholm, september 2002

Haluk Soydan
Gästredaktör
Centrum för utvärdering av socialt arbete, och
Stockholms universitet

Litteratur

Franke-Wikberg, Sigbrit, Lundgren, Ulf (1980). Att utvärdera undervisningen. Del I. Stockholm: Wahlström & Wistrand.

Oakley, Ann (2000). Experiments in Knowing. Gender and Method in the Social Sciences. New York: The New Press.

Macdonald, Geraldine (1998). »Promoting evidence-based practice in child protection« Clinical Child Psychology and Psychiatry, 3 (1), 71-85.

McCord, J. (1978). »A thirty-year follow-up of treatment effects« American Psychologist, March, 284-289.
Scandinavian Journal of Social Welfare: Special Issue – Evaluation Research and Social Work (1998). (7) 2.
Socialstyrelsen (2000). Nationellt stöd för kunskapsutveckling inom socialtjänsten. SoS-rapport 2000:12.
Vedung, Evert (1991). Utvärdering i politik och förvaltning. Lund: Studentlitteratur.
Vedung, Evert (1997). Public Policy and Program Evaluation. New Brunswick: Transaction Publishers.

TACK: Föreliggande temanummer har krävt avancerat administrativt arbete över nationsgränser och ovärderligt stöd till gästredaktören. Jag tackar forskningsassistenten Johan Glad för hans noggranna och effektiva arbete.

Den nya utvärderingsvetenskapen

michael scriven

Är utvärderingsforskningen en egen disciplin eller närmare bestämt en »transdisciplin« som statistik och matematik? Kan man tala om att utvärderingen är en modern konst grundad på vetenskap snarare än en renodlad tillämpad vetenskap? I artikeln reflekterar författaren över den nya utvärderingsvetenskapen och dess roll.

Inledning

En tidigare version av den här artikeln kallades »Utvärdering – en modern konst«, men utvärdering är ju lika mycket vetenskap som konst, och det här temanumret är ett bra tillfälle att betona utvärderingens vetenskapliga sida. Liksom inom socialt arbete är utövarna av utvärdering mycket väl medvetna om i vilken utsträckning tillämpningen kräver en konst grundad på vetenskap snarare än att vara en renodlad tillämpad vetenskap i stil med att bygga broar eller beräkna vilket år Hale-Bopp-kometen kommer att återvända. Det finns tillfällen vid utövandet av dessa konster när man behöver försäkra sig om att man har fast vetenskaplig mark under fötterna; till exempel i inledningsskedet av en ny vetenskap som utvärdering eller i en fullt utvecklad äldre vetenskap som socialt arbete, som då och då behöver ta sig en funderare på och se över sina landvinningar och sin framtid. Därför tänker jag ta det här tillfället i akt att betrakta beskaffenheten av och grunderna för den relativt nya vetenskapen utvärdering. Jag vill börja med att säga att det inte är något nytt med utvärdering – inom och utanför vetenskapsgrenarna – det enda som är nytt är att utvärderingen frigörs för att bilda en egen vetenskap, något som börjar med ett erkännande av dess legitimitet.

Michael Scriven is a Professor of Psychology, Claremont Graduate University, California, where he works at the School of Behavioral and Organisational Science.

Vetenskapens epistemologiska trygghet

Nuförtiden har det faktiskt blivit inne att angripa legitimiteten hos vetenskapen själv och då i synnerhet föreställningarna om sanning och objektivitet, som är av central

betydelse. Borde det kanske göra oss försiktiga med att försöka få med utvärderingen i vetenskapsgrenarnas skara? Vi får börja med att ta oss an dessa angrepp och den epistemologiska skepsis som de grundar sig på, inte bara för att om utvärdering skall bli betraktad som en vetenskap, så skulle den få samma behandling, utan av ett mera speciellt skäl. Skeptikernas särskilda angreppspunkter – begreppens objektivitet och partiskhet – är i själva verket avgörande inslag inom utvärderingsområdet. Om angreppen är vederhäftiga skulle dessa centrala begrepp inte överleva och utvärderingen skulle gå under tillsammans med dem.

Man träffar på denna skepsis vid diskussioner om teorier i socialt arbete på samma sätt som vid diskussioner om utvärderingsteori. Angreppen görs ofta av dekonstruktionister som säger att de har upptäckt att objektivitet är en myt eller att vetenskap inte är ett sökande efter sanning utan ett maktverktyg som överklassen använder sig av. Det är faktiskt ganska roligt för någon som har tillbringat över tjugofem år av sitt liv som yrkesfilosof att se så många människor med fina akademiska meriter i andra ämnen gå i samma fälla som varenda filosofistuderande gör under sitt första år. (Å andra sidan finns det ganska många yrkesfilosofer som har gått i samma fälla under hela sitt yrkesliv.) Men i vilket fall som helst kan man avfärda skeptikernas angrepp på vetenskap och objektivitet eftersom sist och slutligen det sunda förnuftet inte låter sig förlöjligas. Det kan visserligen vara modernt att driva med vetenskapen, men det finns ingen gränslinje mellan vetenskap och sunt förnuft. Eller som Einstein uttryckte det i Physics and Reality: »Hela vetenskapen är ingenting annat än en förbättring av vanligt tänkande.« Vi har inte för avsikt att ge upp våra välgrundade meningar om möblerna i det här rummet, förekomsten av idoga och lata människor, nyttan av värktabletter och Internets objektiva realitet.

Naturligtvis skiftar det som räknas som vetenskap och sunt förnuft en smula efterhand som vi gör nya upptäckter och upptäcker misstag. Nuförtiden är det förnuftigt att inte dela vattenflaska med någon som har en kraftig förkylning, men förr i tiden skulle någon som vägrade att göra detta ha ansetts irrationell eller vara behäftad med sociala fördomar. Så vetenskapen förändras och sunt förnuft förändras – men nästan alla förändringarna sker stegvis och är noga genomtänkta. Vår fantasi fångas av sporadiska radikala begreppsmässiga förändringar, paradigmskiftena – men Galileo och Einstein gjorde bara mycket små detaljförändringar i data de tolkade. Vi behöver inte vara rädda för att vår tro på att bakterier överför sjukdomar kommer att visa sig vara kapitalistisk hjärntvätt. Även om det i princip är möjligt och inträffar emellanåt – och dessutom har synnerligen stort nyhetsvärde – är sådana händelser så pass sällsynta att vi inte behöver betrakta vår tilltro som otillbörlig. Det finns objektiva sanningar och det finns partiskhet, och de kan båda säkert identifieras och särskiljas utan det naiva antagandet att de alltid skulle skilja sig på samma punkt och att det vi är mest övertygade om aldrig skulle vara fel. Den vurm som förespråkar det motsatta – den aktuella vågen av skepsis – kommer att gå över. Man skulle lika gärna kunna sluta upp att använda räkning när man beräknar sin inkomstskatt eftersom

matematikens grunder fortfarande ifrågasätts. Den typen av diskussioner hör hemma runt filosofernas kaffebord – den är ovidkommande i den riktiga världen.

För att ta ett exempel som ligger närmare området samhällsvetenskap och utvärdering, så vet alla mycket väl att man kan förhålla sig ganska objektiv till vissa saker, som till exempel förekomsten av en långvarig diskriminering av kvinnor i moderna västerländska samhällen. Att säga att vi kan förhålla oss objektiva till det är inte det samma som att säga att vi bör vara oengagerade eller oberörda av det. Det säger bara att vi kan vara säkra på det, vi kan veta att det är sant och inte en projicering av våra känslor. Det är helt enkelt ett faktum, på samma sätt som det är ett faktum att Hitler orsakade miljontals oskyldiga människors död. Att vi känner till det visar inte att vi inte kan förhålla oss objektiva till det. Utvärderingar som den här kan vara objektiva – det var bara den positivistiska neurosen som hävdade att det var omöjligt.

En del människor kan inte förhålla sig objektiva till vissa av de här frågorna eftersom de blir för djupt berörda av dem. Det innebär att tillförlitligheten i slutsatserna de har dragit om dem blir lidande och att känslorna i viss mån tar överhanden över deras förnuft. Vi vet att det är mycket svårt för alla att förhålla sig objektiva till vissa saker, som till exempel de goda egenskaperna hos våra egna barn eller karaktären hos den äkta hälft som just har begärt skilsmässa. Vi är ganska bra på att särskilja dessa fall från mera generella fall. Det är helt enkelt tråkigt att hålla på att påminna oss själva om att vi ibland har fel när vi tror att vi är objektiva och helt enkelt befängda om vi därav drar slutsatsen att det inte finns någon objektivitet. Till och med den till synes mest felfria bil kan gå sönder utan minsta föregående varning. Av det följer inte att alla bilar är opålitliga eller att vi aldrig skulle ha rätt om vi påstår detta om en viss bil.

Det vi måste få från de radikala epistemologerna är specifika exempel på partiskhet som vi hittills skulle ha förbisett. Om de är övertygade om att så kallad objektivitet är rena skämtet borde de kunna hitta några exempel som hittills har undgått oss. I själva verket finns det säkert sådana, och det är troligt att de har varit till skada för någon. Vi borde vara beredda och villiga att börja arbeta med att kontrollera bevisen och sedan identifiera och gottgöra offren och inte ägna oss åt filosofiskt gnäll om den allestädes närvarande omöjligheten att det skulle finnas säker kunskap eller objektivitet. Är den säker så är den säker. Termen som den används inom vetenskapen och det sunda förnuftet innebär inte den absoluta omöjligheten för misstag, bara den extremt låga sannolikheten för misstag, en sannolikhet som är så låg att man i praktiken inte behöver ta hänsyn till den.

Mer än så ger vi alltså inte för det allmänna angreppet på vetenskap och objektivitet. Det är helt enkelt ett filosofiskt misstag och bör inte tas på större allvar än ett angrepp på matematikens grunder – men vi bör alltid vara villiga att lyssna om det dyker upp något specifikt, något av praktisk vikt.

Utvärderingens historiska osäkerhet

Filosofiska fördomar som exemplet med

radikal skepsis som diskuterades ovan har ofta haft större inverkan på vetenskapen än man insett vid tillfället. Utvärderingens historia är ett sådant fall. Tabubeläggningen mot den infördes av logiska positivister – som ansåg att utvärderingsmässiga påståenden inte ens kunde tas som förslag, än mindre kunde vara empiriskt sanna eller falska – och har dröjt sig kvar i fyrtio år efter det att positivismen förlorat alla anhängare av betydelse. Positivism är i dag inget annat än en fågelskrämma som radikala epistemologer använder som måltavla, men positivismens åsikt om utvärdering genomsyrar fortfarande de flesta vetenskaperna. Ta till exempel de naturvetenskapliga läroplansreformernas historia efter Sputnik. Samtliga inledde med att förteckna naturvetenskapens viktigaste logiska begrepp och förfaranden, som de buntade ihop till ett slags föreställning om vetenskaplig metod. Detta bildar den grund på vilken det riktiga, faktiska innehållet sedan har byggts. De vanliga begreppen består av bland annat observation, beskrivning, mätning, klassificering, generalisering, förklaring, hypotesprövning och så vidare. Dessa förteckningar var aldrig särskilt imponerande och förefaller bara bli sämre om man skall döma av en som just införlivats i den nya amerikanska läroplanen för de naturvetenskapliga ämnena. Men de har en sak gemensam: de utesluter det allra viktigaste begreppet i den vetenskapliga repertoaren, det enda som är oumbärligt inom alla vetenskaper. De utesluter utvärdering.

Om man öppnar en vetenskaplig tidskrift inom ett slumpmässigt valt område och stannar vid en slumpmässigt vald artikel är det troligt att man finner en presentation som börjar med en genomgång av tidigare genomförda arbeten. Denna genomgång är ytterst selektiv och urvalet görs på grundval av kvaliteten på eller betydelsen av tidigare forskning. Detta kommer förmodligen att efterföljas av en diskussion om planen för den forskning som skall redovisas: skäl kommer att anges till varför man ansåg att den valda planen efter förhållandena var optimal. Det är en utvärderingsmässig slutsats grundad på resonemang, det vill säga just det som sades vara en självmotsägelse. Sedan framläggs resultaten av undersökningen, och rönens kvalitet och betydelse kommer att diskuteras – ytterligare två utvärderingsinsatser. Utvärdering är i själva verket den vetenskapliga undersökningens probersten. Det finns vetenskaper som ägnar sig mycket litet åt klassificering och andra som inte gör mycket annat. Det finns vetenskaper som redovisar enstaka händelser, andra som bara går in för generella principer. Men det finns inga vetenskaper som underlåter att utvärdera utformningen eller resultatet av sitt arbete eller andra forskares arbeten. I synnerhet finns det inga vetenskaper som inte kan skilja mellan bra och dåligt vetenskapligt arbete och vid gränsfall skilja det bästa arbetet från pseudovetenskap, till exempel astronomi från astrologi – och den skillnaden är helt och hållet grundad på utvärdering.

Se på resten av vetenskapsmännens vetenskapliga liv, just de vetenskapsmän som fortfarande inte vill förläna utvärderingen legitimitet som vetenskaplig metod. De undervisar, och de utvärderar varje elev de ger undervisning. Om dessa utvärderingar ifrågasätts försvarar de dem

med hjälp av sakförhållanden och resonemang. Det vill säga, de anser att dessa utvärderingar är objektiva, vilket är något helt annat än att anse att de är ytterst exakta. De granskar arbeten som lämnas in till tidskrifter och ägnar sig därvidlag åt ren utvärdering. De bedömer ansökningar om forskningsmedel – det handlar om utvärdering. De tar sökande till en tjänst på sin institution eller forskningsgrupp under övervägande, vilket är en process helt och hållet grundad på utvärdering. Den egentliga vetenskapliga processen består inte bara till stor och avgörande del av utvärdering, utan varje vetenskaplig disciplins sociala verksamhet hålls samman av ett nät av utvärderingar. I förbigående kan nämnas att alla vetenskapsmän inte bara ägnar sig åt och försvarar utvärdering i sitt privatliv utan också tar råd från sådana som är experter på det, till exempel konsumenttidskrifter eller utredningar om mediciner som görs av statliga organ.

Kan man då inte säga att uteslutningen av utvärdering, tabubeläggningen av utvärdering, som kännetecknar vetenskapens historia under det senaste århundradet i sanning är ett märkvärdigt fenomen? Hur kunde det hända? Denna uteslutning tror jag har mycket djupa rötter. Även om den uppenbarligen i hög grad stöddes och främjades av positivisterna dröjde den kvar så länge efter den offentliga avrättningen av den filosofin att det måste finnas fler förklaringar till den. Jag tvivlar på att doktrinen om värderingsfrihet kunde ta ett sådant grepp och hålla kvar det utan att åstadkomma en genklang i vetenskapsmännens själar. Vad kan den genklangen vara för något?

Enligt min åsikt är det ingen händelse att den religiösa traditionen, i likhet med den vetenskapliga, har samma bisarra, nästan schizofrena inställning till utvärdering. »Dömen icke, så skola I icke bliva dömda« säger Bibeln och ändå är dömandet själva poängen med varenda liknelse och grundtanken hos varenda profet. Är det inte konstigt att den förbjudna frukten i Första mosebok satt på kunskapens träd på gott och ont? Bibeln säger att Gud förbjöd Adam och Eva att äta frukten eftersom den skulle göra dem lika gudarna: gudomen förbehöll sig själv rätten att utvärdera. Och ändå tvekade inte Gud att straffa dem som var olydiga mot honom genom att äta frukten från det trädet trots att de inte kunde veta att det var syndigt att göra det. Under årtusendena som har gått sedan dess har prästerna inom många religioner fortsatt att förbehålla sig rätten till den särskilda kunskapen om vad som är gott och ont. Så den religiöse lekmannen är förbjuden att utvärdera eller att döma, på samma sätt som den vanlige vetenskapsmannen förbjöds att behandla utvärdering som en respektabel kognitiv process. Trots det ägnar sig alla lekmän ständigt åt det, precis som alla vetenskapsmän.

Jag tror att utvärderingen framkallar den här schizofrena responsen eftersom den söker sig ända ned till ångestens och rädslans primitiva källor. I den reviderade standardnomenklaturen för psykiatrisk klassificering finns det numera en åkomma som kallas för utvärderingsångest. Jag tycker att det här bekräftar den ångest jag talar om. Vi är med rätta nervösa inför en sluttentamen. Vi är neurotiskt nervösa när vi hävdar att detta visar att tentamina är olämpliga. När jag hör utbildningsteoretiker hävda att lärare aldrig någonsin borde betyg-

sätta studenternas arbeten, och jag funderar över detta totalt förryckta förslag, vet jag att jag måste söka djupa orsaker till det och till det allmänna fenomen som det är ett exempel på – de ofta förekommande angreppen på något som helt tydligt är värdefull utvärdering. Jag minns det raseri med vilket många fakultetsmedlemmar mötte de första förslagen om att studenterna skulle tillåtas eller kanske till och med uppmuntras att utvärdera dem. Det var ungefär samma slags reaktion som skulle ha mött en hädelse. »Du skall inga andra gudar hava jämte mig« är det man närmast kommer att tänka på – de nya utvärderarna inkräktar på de gamla reviren. Vetenskaparen, en utvärderare in i ryggmärgen inom sitt eget område, blir förbittrad när makten övergår till andra och uppfinner en absurd doktrin för att rättfärdiga tillbakavisandet. Doktrinen är inte absurd bara för att den är rena självmordet, vilket ett ögonblicks allvarlig begrundan av deras eget arbete snart uppenbarar, utan den är absurd på grund av att var och en som följer ens den enklaste form av vetenskaplig metod skulle se att den är falsk.

Om vi bara kan se klart på det är det här en scen från ett stort drama i tänkandets historia, och dramat är den långsamma processen att göra vetenskapen till en självrefererande verksamhet. I det här exemplet ser man hur långt vetenskapen har varit från att studera sin egen process innan de lagt fram dogmer om hur deras egen process är beskaffad. Kort sagt visar fallet hur tydligt som helst att vetenskapen är långt ifrån att vara en vetenskaplig verksamhet. Nu talar jag i och för sig i egen sak när det gäller det här angreppet. Jag hävdar att försöket att tabubelägga utvärderingen var otillbörligt och att vetenskapsgrenen hölls tillbaka under större delen av århundradet på grund av denna blinda fördomsfullhet.

Ibland förefaller det mig som om böjningen av det relevanta verbet, åtminstone i vetenskapsmännens ögon, skulle låta ungefär så här: »Jag utvärderar, du uttrycker en åsikt, och de är helt enkelt känslomässiga.« De ville inte ge utvärderingen legitimitet eftersom de inte ville släppa vargen lös. Så länge den är fastbunden kan vi använda den. Om den släpps loss kan den bita oss. Med tanke på att ett tillbakavisande av utvärderingens legitimitet vore att skära hjärtat ur hela vetenskapen och allt praktiskt liv var det en desperat åtgärd. Med tanke på att vi inte själva har för avsikt att ge upp den, bara förmena den sin legitimitet och därmed andras användning av den, är detta verkligen en ovärdig inställning. Det är en inställning som intellektuella borde skämmas över att ha. Jag kan inte finna någon annan förklaring till en sådan inställning än en rädsla för att själv hamna i händerna på andra, en rädsla för att en granskning skulle avslöja ens egna tillkortakommanden.

I det mänskliga tänkandets historia är tabubeläggningen av utvärderingen onekligen ett av de mest fascinerande, mystiska och på många sätt tragiska teman som finns. Det är tragiskt, inte bara patetiskt, eftersom en av konsekvenserna av tabut, som går raka spåret tillbaka till Max Webers tidiga version av doktrinen om värderingsfrihet, var att samhällsvetenskaperna slutade befatta sig med mänskliga behov. De var stolta över att det inte förekom någon utvärdering i deras arbete och underlät därmed att inrikta sitt arbete mot de ställen där en vetenskaplig inställning

till det sociala sönderfallet skulle ha kunnat leda till stora förbättringar. De letade inte efter behov, de lade inte märke till desperata behov, och i vilket fall som helst ansåg de inte att de hade med det att skaffa. Om ni tvivlar på det här, läs recensionerna av de två första böckerna om fattigdom i USA, som kom ut långt efter det att vi hade passerat halvsekelskiftet. De är väldokumenterade och välskrivna – och mottogs med stort förakt med hänvisning till att det inte kunde finnas någon fattigdom att tala om i USA. Den metodologiska myten hade förblindat vetenskapens iakttagande ögon. När man läser ordväxlingarna mellan företrädarna för socialt arbete och samhällsvetenskap på Internet eller i tidsskrifter kan man än i dag känna spänningen mellan vetenskaparmodellen och hjälparmodellen, en spänning som rätt och slätt består av rök från den positivistiska soptippen.

Tragedin låg i att samhällsvetenskapens fokusering på mänskliga behov, som det skedde inom det medicinska området, i stor utsträckning inom socialt arbete och annorstädes, och som har börjat dyka upp under de senaste årtiondena, kunde ha räddat otaliga liv under de långa mörka tiderna under vårt eget nittonhundratal. Så det här är inte blott och bart ett intellektuellt gruff. Det stod mycket på spel i den här kampen och det är en kamp där de vetenskapliga krafterna skämde ut sig. Men det fanns förstås en och annan framstående person som gick mot strömmen – Gunnar Myrdal, till exempel. Men de kunde inte hejda strömmen i USA och andra engelsktalande länder.

Ni tycker kanske att ni har hört allt det här förut, många gånger, och att poängen bara är kommentaren om de hycklande vetenskapsmännen som utövade utvärdering men förmenade den dess legitimitet. Faktum är att det bara är en replik man kan glida över. Visst hycklade de, men det finns en mycket allvarligare fråga att ställa: Hade de rätt? När man upptäcker att en predikant är äktenskapsbrytare, lögnare och tjuv så klandrar man hans hyckleri, men man drar inte slutsatsen att han hade fel. Det är en helt annan sak. Och likadant är det i det här fallet. Vi har avslöjat hyckleriet, men vad skall vi göra med de argument som användes för att stödja ståndpunkten? För att uttrycka sig rakt på sak: Skall man utgå från att de hade fel bara för att de var hycklare?

Det är uppenbart att vi inte bara kan utgå från det. Vi måste ha starka argument för det felaktiga i ståndpunkten. Ett av dessa argument finns i ett motsägelsebevis som är underförstått i det jag har sagt: om utvärdering är otillåten eller ovidkommande så är det mesta av vetenskapen och det mesta praktiska beslutsfattandet också otillåtet. Om man tror, vilket förefaller rimligt och som jag har hävdat ovan, att detta är en absurd inställning, så måste vi avfärda doktrinen om värderingsfrihet. Men vi måste också titta närmare på de uttryckliga argument som har framförts för att tabubelägga utvärderingen. Jag skall strax åtminstone antyda varför de var missriktade. Först när de kan vederläggas kan vi känna oss fria från tabut och utvärdering kan försöka uppnå vetenskaplig status.

Sätt det jag har sagt i motsats till den åsikt som nästan är en del av den intellektuella kulturen i dag, nämligen åsikten att den värderingsfria inställningen avslöjades som humbug av krafterna inom den politiska liberalismen och radikalismen.

De och alla andra som blivit övertalade av dem anser att den radikala vänstern var riddaren i skinande rustning som dräpte den värderingsfria draken. Det de gjorde var i själva verket fullständigt ovidkommande vad doktrinen om värderingsfrihet beträffar. Det de trodde sig göra var att visa att samhällsvetenskaperna var ett exempel på kejsarens nya kläder, eftersom just de samhällsvetenskapare som avfärdade utvärderingens berättigande inom vetenskapen i själva verket tog ställning i en rad frågor, ställningar som hade hämtats från deras yttre politiska och personliga värderingar. Den typen av argument visar precis samma totala okunnighet om frågan som ledande vetenskapsmän själva uppvisade. Doktrinen om värderingsfrihet hävdade aldrig någonsin att vetenskapsmän inte hade både politiska och personliga värderingar eller att vetenskapsmännen inte skulle använda dem för att bestämma vilken vetenskap de skulle ägna sig åt eller om de skulle ägna sig åt vetenskap. Doktrinen om värderingsfrihet var något helt annat. Det var en doktrin om att utvärderingsmässiga påståenden inte hade någon vetenskaplig status, det vill säga inte kunde bekräftas med hänvisning till logik eller vetenskapliga metoder. Det hävdades inte att de inte hade någon inverkan på vetenskapsmännen eller ens att de inte skulle ha någon inverkan på vetenskapsmännens beslut om huruvida de skulle ägna sig åt vetenskap eller vilken vetenskap de skulle ägna sig åt eller hur de skulle använda sig av den. Radikalerna dräpte i sanning en drake – den var en mytisk fiende.

Bland de argument som framfördes ovan om hyckleriet i den vetenskapliga inställningen finns det däremot ingen hänvisning alls till de värderingar som driver fram sådana beslut. Det var alltid klart och tydligt och förnekades aldrig för ett ögonblick av någon företrädare av värderingsfri vetenskap att det är yttre värderingar som driver fram dessa personliga beslut av vetenskapsmännen. Det som framförs i argumentet ovan är att hyckleriet härrörde från den faktiska användningen av utvärdering i själva hjärtat – i teorin och praktiken – i den logiska kärnan – hos varje vetenskap. Slår man ihop denna genomgripande och uppenbara roll med avfärdandet av dess berättigande får man hyckleri som grundats på självbedrägeri. Det som den radikala vänstern visade var att en doktrin som aldrig någon var anhängare till var oförenlig med en praxis som alla anslöt sig till. Det är inte någon särskilt upphetsande slutsats.

Det enda seriösa sättet att vederlägga doktrinen om värderingsfrihet är att ta itu med en undersökning av vetenskapens struktur, inte dess sociala omgivning, och visa att vetenskap i allt väsentligt är utvärderande. Och det var det jag gjorde, litet löst, när jag genomförde tankeexperimentet att slå upp en vetenskaplig tidsskrift och noggrannare betrakta de kognitiva processer som innehållet innefattar. Det ger oss de motsägelsebevis vi behöver. Det ger oss också slutsatsen om hyckleri, om man inser att vilka vetenskapsmän som helst hade kunnat göra det vi diskuterade och upptäcka att det de gjorde var oförenligt med det som vetenskapsmän säger om vetenskapens beskaffenhet. Om man fastställer slutsatsen om hyckleri på detta sätt är den välgrundad snarare än ett angrepp på en fågelskrämma.

Men det innebär att vi sitter där med klara argument för att det är omöjligt att

vetenskapligt rättfärdiga värderande slutsatser, vilket naturligtvis tyder på att det är omöjligt att ha en utvärderingsvetenskap.

Argumenten för att utvärdering är en omöjlighet

Det finns två sådana som är nära besläktade, och de är mycket enkla. Det första var argumentet att vetenskapen bara ägnar sig åt, eller bara borde ägna sig åt, att beskriva världen som den är, det vill säga bara ägna sig åt beskrivningar och inte tala om hur den borde vara, det vill säga rekommendationer. Den andra är en version av Humes argument, att man utifrån deskriptiva premisser – vetenskapliga data eller observationer – inte kan dra giltiga utvärderingsmässiga slutsatser. Det första argumentet är uppenbart felaktigt om man bara bryr sig om att titta närmare på bevisen. Vetenskapen ägnar sig faktiskt åt hur världen ser ut, men det kräver att man också ägnar sig åt hur den kan beskrivas på bästa sätt, med de bästa förklaringarna, de bästa sätten att undersöka den och med de bästa verktygen för sådana undersökningar. Med andra ord måste vetenskapen vara utvärderande för att kunna vara deskriptiv.

När man väl fått klart för sig att vetenskapens logik kräver användning av utvärderingens logik, legitimerar det erkännandet av det sätt på vilket tillämpad vetenskap alltid och vederbörligen utfärdar rekommendationer om hur saker bör göras, till exempel hur temperaturen i en järnsmältugn skall kontrolleras, hur en bro bör byggas, hur en stor damm skall konstrueras för att komma upp i den bestämda kapaciteten, etc. Så vetenskapen har inte bara med instrumentell utvärdering att göra utan även med tillämpad utvärdering. Faktum är ju att när konsumenttidskrifter rekommenderar den bästa datorn för vissa ändamål ägnar de sig åt produktutvärdering, ett slags tillämpad vetenskap. Och dessa tillämpade matematiska och byggnadstekniska exempel skiljer sig inte från exempel på samhällsbyggande, till exempel tillämpad psykologi eller tillämpad sociologi, utom när det gäller parametrarnas komplexitet. Det är med andra ord ingen principiell skillnad, inget brott mot logiska begränsningar.

Det andra argumentet, att man inte kan dra utvärderingsmässiga slutsatser av faktiska premisser, verkade ganska förnuftigt på den tiden när Hume framförde det, när slutsatser på det hela taget betydde deduktiva slutsatser. Men nuförtiden förstår man härledningens beskaffenhet bättre. Man förstår att slutledningen är en särskild typ av härledning, en som gör sina slutsatser logiskt nödvändiga utifrån de givna premisserna, men man förstår också att det finns två andra typer av härledning som bara gör sina slutsatser troliga: statistisk härledning, den mest erkända, och beviskraftig härledning, den typ som styr de flesta juridiska härledningar och sådana som grundar sig på sunt förnuft. (Ett specialfall av beviskraftig härledning är luddig härledning, som är mycket mer berömd än sitt upphov.) Beviskraftig härledning är härledning av slutsatser som vid första påseendet är troliga, med de förutsättningar som givits. Detta innebär inte att man kan räkna ut en siffermässig sannolikhet för slutsatserna, vilket är möjligt för statistisk härledning, utan att de flesta bevisen talar för dem, som man brukar säga, eller att de troligtvis

är sanna, under i övrigt lika förhållanden, som man ofta säger. Det som ofta talar för den beviskraftiga härledningen är att den förefaller ge de bästa förklaringarna till premisserna och att dessa premisser förefaller vara av ett sådant slag som det på det hela taget alltid finns förklaringar till (med andra ord är determinism tillämplig).

Beviskraftig härledning är just den typen som leder oss från premisser om behov och sätt att tillfredsställa dessa behov och tillgängliga resurser, till slutsatser om vad som bör göras, det vill säga slutsatserna om programutvärderingar. Det är också den typen som leder oss från kunskap om den täckning som en högskolekurs har plus information om bakgrundkunskaperna hos studenterna som läser kursen, tiden som är tillgänglig för kursen och de angivna nivåerna, till slutsatser om betyg för deras specialarbeten, det vill säga slutsatser grundade på prestationsutvärderingar. Och samma sak gäller personalutvärderingar. Här har vi med andra ord utvärderingens grundläggande logik, underförstådd i den praktiska utvärderingsprocessen inom de många delområden som utvecklade sina egna procedurer på grund av det praktiska behovet av att göra det, ogenerade av vetenskapens underlåtenhet att erkänna deras berättigande. Och det humeska argumentet, i likhet med radikalernas angrepp på doktrinen om värderingsfrihet, får därmed betraktas enbart som ett tjusigt angrepp på en ståndpunkt som ingen intar.

Bergen av invändningar mot det berättigade i en utvärderingsvetenskap kan därför bestigas, och det är dags för oss att inrätta vårt basläger. Vad är egentligen utvärdering? Är det numera en disciplin snarare än en rad underförstådda tillämpningar, och har den disciplinen bidragit med något som främjar vetenskapens framsteg?

Utvärderingens beskaffenhet och logiska grunder

Som vi har sett är utvärdering inte en svårtillgänglig process begränsad till vetenskap eller etik. Småbarn lär sig mycket tidigt att i sin omgivning skilja mellan det de gillar och det de ogillar. Det rör sig här inte om utvärdering utan om att lägga grunden till utvärdering. Efterhand som de blir äldre lär de sig att förknippa orsakerna till att resultaten blivit behagliga med själva välbehaget, och har därmed börjat identifiera tillstånd eller saker som inte omedelbart står till buds men som eventuellt, och ofta indirekt, är värdefulla för dem. Fortsatt utbildning och mognad utsträcker omfattningen av de överväganden som innefattas i utvärdering till andra värden än välbehag och ger en färdighet i utvärdering vad avser endast dessa andra värden. På så sätt lär vi oss att utvärdera olika slags saker i enlighet med normen som gäller för just den saken, inklusive konståkning, motorcyklar och hyresfastigheter.

Och så är vi där till slut och har blivit förtroliga med det naturliga och enkla utövandet av vägning och bedömning av styrkan och stabiliteten i de många överväganden som måste göras i samband med fastställandet av förtjänster och värden, det vill säga själva utvärderingsprocessen. Det är faktiskt en av de allra vanligaste och viktigaste kognitiva processer som vi är i stånd att utföra. Men det är en som framkallar blandade känslor i oss eftersom

vi växte upp under sådana förhållanden att våra handlingar och vi själva ofta utvärderades, inte sällan till vår nackdel, och ofta med bestraffningar förknippade med dessa utvärderingar. Så vi slits mellan tillbakavisande och godtagande, mellan ogillande och erkännande av nyttan av den. Till och med inom de vetenskapsgrenar där motsättningarna ofta inte är lösta och resultaten motsägelsefulla.

Fundera på hur många gånger någon som undervisar i en introduktionskurs i samhällsvetenskap har sagt: »Det är inte lämpligt att vetenskapsmän gör värderingar« utan att de tänker på att de just yttrat en självmotsägelse. Det här är inget område där vetenskapsmän är immuna mot logiska misstag av enklaste slag.

För att betrakta den mer positiva sidan av vår uppgift, låt oss se om vi kan ställa upp utvärderingsvetenskapens grundläggande grundsatser. Vi måste börja med att skilja mellan två slags vetenskap, som vi kan kalla grunddiscipliner och transdiscipliner. Gränsen mellan dem är inte särskilt tydlig, men den generella skillnaden är att grunddiscipliner i första hand är inriktade på empiriska undersökningar av fysiska, sociala, beteendemässiga eller psykologiska skeenden, medan transdiscipliner är: 1) analytiska snarare än empiriska, 2) verktygsdiscipliner som underlättar de undersökningar och utredningar som genomförs av grunddisciplinerna snarare än att utföra arbetet själv, 3) också självständiga studieområden, vilket ger dem rätt att kallas egna discipliner. De främsta transdisciplinerna är nog statistik, sannolikhetsteori, mätvetenskap, logik och utvärdering. Varje disciplin har ett eget studiefält, i vissa fall en hel akademisk institution, men varje fält är formellt och inte empiriskt. Varje transdisciplin är också ett viktigt verktyg inom ett halvdussin grundfält. Och var och en får mycket av sin utvecklingskraft från försöken att lösa problem som uppstår i dessa andra discipliner. Experimentell design är ett annat exempel, liksom beslutsteori, och möjligen etik.

Det finns en viktig skillnad i transdisciplinernas historiska utveckling. Logik och utvärdering utvecklades till en sofistikerad – men för det mesta underförstådd – nivå samtidigt som grunddisciplinerna mognade. Logiken, som började underförstått, kom fram ganska tidigt som en egen disciplin för två tusen år sedan. Statistik och sannolikhetsteori kom senare, som biprodukter av vissa vetenskapliga och matematiska verksamheter, men utvecklade sedan snabbt sina egna teorem och esoteriska begrepp liksom otaliga värdefulla tjänster åt grunddisciplinerna, beträffande allt från kvantmekanik till demografi. Utvärdering och mätvetenskap är fortfarande i hög grad underförstådda i vanligt vetenskapsutövande och har först på senare tid börjat utvecklas självständigt, och då för att åter förse grunddisciplinerna med nyttiga verktyg. Mätvetenskapen gjorde förstås det här tidigare än utvärderingen. Experimentell design var litet före dessa två grupper. Den delade sig i och med Fishers och andras arbeten, och har bevisat sitt värde, men det är klart att mycket av experimentell design inom de fysiska vetenskaperna har utvecklats på egen hand och inte haft någon större glädje av den självständiga disciplinens arbete. Beslutsteori är också ett mellanliggande fall. Vad den här typen av skillnader

beträffar, vill jag hävda att vi nu kan peka på ett flertal ytterst viktiga bidrag från utvärderingens sida till vanlig vetenskapsutövning, varav jag nämner några nedan. Detta är en synnerligen viktig milstolpe eftersom det är mycket svårt att avfärda ett ämne som har blivit eller bidragit med ett viktig verktyg till ens egen och andra grunddiscipliner.

Om man har i åtanke det sätt på vilket S. S. Stevens inrättade det mätvetenskapliga fältet genom att först identifiera kärnbegreppen och sedan utveckla någon teori om dem, kommer man att se likheter med det sätt på vilket jag inrättat utvärderingen. I det följande gör jag en allmän översikt av området.

1. Utvärdering är en process där enheters värde, förtjänst eller betydelse bestäms; utvärderingar blir resultatet av denna process. Utvärdering kan vara extern eller intern eller en blandning härav, och den kan vara kvantitativ eller kvalitativ eller en blandning härav. Den är starkt men inte alltid skarpt avgränsad från förklaring.
2. Det finns bara fyra grundläggande utvärderingsgrunder, varav den ena – en kognat – måste uppträda i alla utvärderingsmässiga slutsatser (eller ges av de slutsatserna utifrån sammanhanget). Utvärderingsgrunderna är gradering, rangordning, poängsättning och fördelning. Var och en av dessa måste understödjas av ett lämpligt och separat utredningsförfarande. Det finns bara två samband – poängsättning medför rangordning, och fördelning medför, men är inte begränsad till, en blandning av gradering och rangordning. Rekommendationer är en del av den tillämpliga utvärderingsvokabulären och består av kontextberoende slutsatser baserade på någon kombination av de grundläggande utvärderingsgrunderna.
3. Utvärdering är en självständig disciplin som bestäms av 1 och 2, och den utvecklar sina egna modeller, teorier och förfaranden. Dessa täcker frågor som att anföra logiken i de grundläggande utvärderingsgrunderna, kartlägga sambandet mellan utvärdering och förklaring, planera, beskriva, klassificera, göra förutsägelser och ge rekommendationer. Vidare att skapa och fastställa metoder för interna synteser (att integrera delutvärderingar i en övergripande utvärdering) och externa synteser (att integrera utvärderingar som gjorts av flera utvärderare i en övergripande utvärdering – som också är ett slags metaanalys). Utvärdering använder också, som verktyg, många metoder och tekniker från andra discipliner.
4. Utvärdering är också en transdisciplin, en disciplin som förser andra discipliner med verktyg. Vissa av dessa verktyg har utvecklats som en viktig del av disciplinerna efterhand som de kommit fram ur sina förvetenskapliga ursprung. Således är intradisciplinär utvärdering – utvärdering av hypoteser, data, resultat, utredningsmässiga förfaranden (experimentell design, till exempel),

instrument, tidigare arbeten – en viktig del av alla andra discipliner, mycket viktigare än de tilläggskomponenter som härletts ur de flesta transdisciplinerna. Det är i själva verket det nyckelverktyg som skiljer dessa discipliner från irrationell eftergivenhet och ren spekulation. Arbetet inom den självständiga disciplinen utvärdering kan nu påvisa många sätt på vilket intradisciplinär utvärdering kraftigt kan förbättras (till exempel utvärdering av förslag, kandidater, studerande, vetenskapliga instrument, programvaror och maskinvaror och genom att förbättra den interna syntesen).

5. Utvärdering har massor av namngivna praktiska områden, där kompetent utvärdering har pågått i många år, ibland i århundraden eller årtusenden. Bland dessa finns produkt-, program-, prestations-, personal-, förslags- och policyutvärderingar, liksom institutionell utvärdering, läroplansutvärdering, litteraturkritik, kvalitetssäkring inom industrin, etiska bedömningar, juridiska granskningar, metautvärdering (utvärdering av utvärdering), liksom de intradisciplinära utvärderingarna som nämndes ovan. Den grundläggande logiken i utvärdering är precis samma inom vart och ett av dessa områden, och – vilket tål att understrykas – många av de professionella metoder som utvecklats inom ett av dem fungerar inom många av de andra. I dagens läge finns det gott om utrymme för förbättringar av utvärderingen inom alla dessa områden. Medan en del av det helt enkelt kan åstadkommas genom överflyttning av metoder eller begrepp som redan slagit rot inom andra områden (till exempel skillnaden mellan primära (kriteriegrundande) indikatorer och – vanligen otillåtna – sekundära (växelverkande) indikatorer) återstår mycket att utveckla från de grundläggande satserna (som framgår av punkt 3 ovan) och genom att tillämpa utvärderingens grundläggande logik, som är – liksom inom juridiken och det mesta av vårt praktiska liv – beviskraftig logik (det vill säga härledning av slutsatser som gjorts vid första anblicken), inte slutledning eller statistisk härledning.
6. Utvärdering är en nyckelprocess inom alla meningsfulla aktiviteter i vardagslivet, inklusive konstnärliga, fritidsmässiga och reflekterande. Inom många av dessa områden, liksom inom många av de tillämpade fälten, är de nuvarande nivåerna på utvärderingskvaliteten mycket begränsade och priset man får betala för undermåliga utvärderingar är mycket högt. Det beror delvis på dålig täckning i skolan inom relevanta ämnesområden (till exempel försäkring, investering, dubbla yrkesutbildningar för dagens arbetsmarknad, strategier för beslutsfattande), men också delvis på dålig täckning i vilken som helst av läroplanerna för K-8 (från förskolan till och med åttonde studieåret) avseende de grundläggande principerna för eller utövan-

det av utvärdering (till exempel skillnaden mellan god utvärdering och ren åsikt eller beskrivning, skillnaden mellan engagemang och bias, metoder för biaskontroll, värdet av och metoder för kritisk utvärdering och självutvärdering, färdighet i att utvärdera produktutvärderingar). Med tanke på detta och situationen som beskrevs under punkt 5 ovan, är det kanske dags för professionen att överväga en rejäl insats för att öka mängden utvärdering som finns i vanliga läroplaner, inte minst de som gäller naturvetenskap, samhällsvetenskap och yrkesförberedelser.

7. Färdighet i utvärdering är mycket värdefullt inom många angränsande aktiviteter som inte är rent utvärderande, till exempel planering, målförtydligande, diagnosticering, idékläckning, ledning, förordande, förklarande, problemsökning, undervisning och yrkesutbildning. Även om det är lämpligt och viktigt för utvärderare att företa och delta i sådana aktiviteter bör det stå klart att det i sig inte är någon ersättning för att skaffa sig eller tillämpa utvärderingsfärdigheter (eftersom det inte rör sig om alternativa sätt att gripa sig an utvärdering), och det är troligt att de ligger utvärderaren i fatet vid utvärdering av samma utvärderingsobjekt vid ett senare tillfälle.
8. De flesta utvärderingstillämpningar påverkar människor i grunden eftersom de står i samband med kvaliteten på deras arbete eller deras värde. Alla de angränsade aktiviteter som finns uppräknade under punkt 7 påverkar människor. Det är därför viktigt att utvärdering används med ansvar och viktigt att det står klart att många av dess tillämpningar och angränsade aktiviteter kräver mellanpersonliga färdigheter som ligger långt från de färdigheter som krävs för att utveckla logiken i disciplinens utredningar.

Dessa överväganden antyder att betydande förbättringar inom yrkesmässig utvärdering har möjlighet att kunna komma till avsevärd nytta för det mesta av vad människor tar sig för. Jag tror att vårt arbete inom utvärdering, på samma sätt som inom socialt arbete, bör låta sig inspireras av den möjligheten. Vi borde ständigt vara på alerten, inte bara inför möjligheten att utvidga resultaten från ett utvärderingsområde till ett annat, utan också inför chansen att nå ut från den akademiska världens vanliga områden till sådana som alla medborgare har intresse i. Utvärdering är inte längre bara en samling olika aktiviteter, må vara indelade i discipliner. I likhet med den medicinska vetenskapen är den nu inte bara en akademisk disciplin utan även något mycket mer än en akademisk disciplin.

Noter

1. Vissa avsnitt av den här artikeln lades fram för Australian Evaluation Societys årsmöte 1996 på Nya Zeeland i augusti förra året. Kommentarer och kritik är mycket välkomna: P. O. Box 69, Point Reyes, CA 94956, scriven@aol.com, fax 415 663-1913. En version av artikeln har tidigare publicerats i Scandinavian Journal of Social Welfare, volym 7, nummer 2, 1998.

Utvärderingsmodeller

evert vedung

Vilka frågor kan man egentligen besvara genom en utvärdering och hur skall man gå tillväga för att utvärdera? I artikeln presenteras en översiktlig bild av olika utvärderingsmodeller. Med en historisk överblick typologiseras utvärderingsmodellerna efter grundläggande värdekriterier.

En ambitiös utvärdering försöker besvara två frågor. Motsvarar resultatet det som man kan förvänta eller önska sig utifrån något värdekriterium? Är detta resultat på något sätt en produkt av interventionen?

När utvärdering började praktiseras i Sverige inom skolpolitik på 1950-talet och kom upp som en nyhet i USA omkring 1965 besvarades dessa båda frågor på ett mycket karakteristiskt sätt. Frågan om kriterieuppfyllelse uppfattades som en fråga om interventionen uppfyllde sina egna mål. Motsvarar de resultat som uppnås de i förväg uppställda målen med interventionen? Utvärdering innebar att besvara denna måluppfyllelsefråga. Problemet med interventionens effekter löstes på ett liknande karakteristiskt sätt. Lösningen bestod i att använda sig av bästa möjliga naturvetenskapliga metoder. Den uppläggning som favoriserades var det randomiserade eller matchade experimentet. På slumpmässiga grunder borde akademiska forskare skapa två likvärdiga grupper, en experimentgrupp och en kontrollgrupp. Forskarna borde sedan utsätta experimentgruppen men inte kontrollgruppen för interventionen och skrupulöst noga mäta utvecklingen i båda grupperna före, under och efter interventionen. Om det uppstod några skillnader mellan grupperna borde de kunna tillskrivas interventionen eftersom allt annat utom interventionen genom randomiseringen var lika.

Båda dessa ganska stränga synsätt finns fortfarande kvar i utvärderingsdiskursen. Det som hänt är att de trängts tillbaka. Alla medger att offentliga insatser kan utvärderas också på många andra sätt än som uppnående av på förhand uppsatta mål. Man kan utvärdera mot ekonomiska mål som produktivitet och utfallseffektivitet, mot intressenters mål och mot brukarmål för att nu nämna några. I utvärdering av högre utbildning och forskning har det uppstått en specialitet som utvärderar mot de professionellas (forskarnas) egna kvalitetsmål i självvärderingar samt mot forskarkollegers kvalitetsuppfattningar. Det har också skett en förskjutning från att se utvärdering som forskning mot att se utvärdering som

Evert Vedung är professor i statsvetenskap och arbetar på Institutet för bostads- och urbanforskning och Statsvetenskapliga Inst. vid Uppsala universitet.

demokratisk i bemärkelsen participativ och deliberativ, som en angelägenhet för vanliga berörda.

Den översikt som här skall ges sker utifrån ett smalt perspektiv. Utvärderingsmodeller skall ordnas på basis av de grundläggande värdekriterier som används för att bedöma verksamheten.[1] En taxonomi över utvärderingsmodeller grupperade efter grundläggande värdekriterier presenteras i figur 1.

Figur 1.
Utvärderingsmodeller.

Substansmodeller
- Måluppfyllelsemodellen
- Bieffektsmodellen
- Brukarmodeller
- Intressentmodeller
- Professionella modeller: kollegiebedömning
- Professionella modeller: självvärdering

Ekonomiska modeller
- Produktivitetsmodellen
- Effektivitetsmodeller:kostnads-effektivitetsanalys
- Effektivitetsmodeller: kostnads-intäktsanalys

Substansmodeller är helt och hållet inriktade på interventionens, verksamhetens, sakinriktade aspekter. Det är resultaten eller verksamheten i sak som står i centrum. I vilken riktning och hur mycket har energipolitiken påverkat energiflödena i samhället? Vilka effekter på miljön har det miljöpolitiska programmet fått? Även ekonomiska modeller är inriktade på saken men inte enbart. De undersöker också vad det kostar att producera saken. Skillnaden mellan ekonomiska modeller och substansmodeller är att de förra även beaktar kostnader.

Substansmodellerna utgör en mångfacetterad grupp. Det urval som tas upp är måluppfyllelsemodellen, bieffektsmodellen, brukarorienterade modeller, intressentmodeller och professionella modeller. Viktigast bland ekonomiska tillnärmelser är produktivitets- och effektivitetsmodellerna.

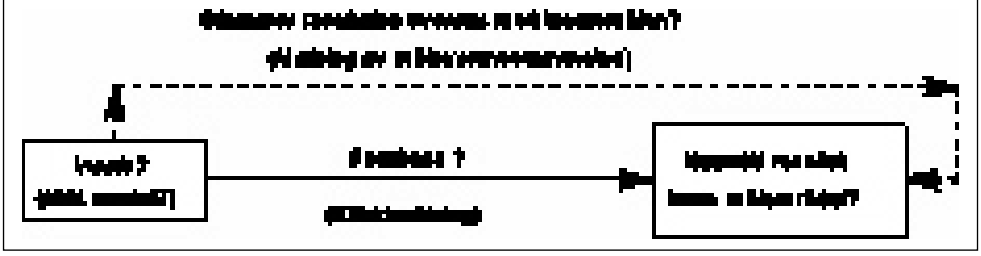
Måluppfyllelsemodellen

Vid måluppfyllelseutvärdering undersöks om de resultat som insatsen faktiskt producerat inom målområdet motsvarar insatsmålen. Måluppfyllelsemodellens grund-frågor är två. Stämmer resultaten inom målområdet överens med de beslutade insatsmålen? Beror detta i så fall på insatsen? Det första momentet kallas här mätning av målöverensstämmelse, det andra effektmätning.

Måluppfyllelsemodellen är föredömligt tydlig och enkel. Först måste en utvärderare klara ut hur målen för styrningen ser ut samt undersöka om målen har uppnåtts inom målområdet. Sedan kan hon ställa frågan om insatsen i så fall bidragit till måluppfyllelsen. Måluppfyllelseutvärderingens enkla anatomi illustreras i figur 2.

1 Framställningen utgör en kraftigt förkortning och skärpning av kapitel 4 i min Utvärdering i politik och förvaltning (Lund: Studentlitteratur, 1998, 2:a upplagan, www.studentlitteratur.se) och kap 4 i Public Policy and Program Evaluation (New Brunswick, N.J.: Transaction Publishers, paperback 2000, www.eurospan.co.uk.

Figur 2.
Måluppfyllelsemodellen för utvärdering



Ett gripbart exempel må anföras. Det institutionaliserade målet i 1978 års svenska energisparplan för befintlig bebyggelse var att »nettoenergiförbrukningen« i 1978 års byggnadsbestånd år 1988 skulle vara cirka 35 TWh lägre per år än det var år 1978. Detta innebar en reduktion på 30 procent. För att uppnå detta förbluffande klart uttryckta mål anslog riksdagen betydande belopp för isolering av fastigheter samt till rådgivning för energisparande. En måluppfyllelseutvärdering av energisparplanen skulle innebära att vi undersökte om energianvändningen i 1978 års byggnadsbestånd verkligen var 35 TWh lägre per år vid 1988 års utgång samt om det uppmätta resultatet på något sätt hängde samman med den av riksdagen antagna planen.

Måluppfyllelseutvärdering avser nästan alltid kollektiva aktörers mål. Det betyder att utgångspunkten är mål som institutionaliserats genom att uttryckligen åberopas i beslutssituationen. Modellen kan avse offentliga interventioner på alla tänkbara politiska och administrativa nivåer från den subkommunala över EU-program till den globala nivån.

Måluppfyllelsemodellen kan också tillämpas på interventioner av olika aggregeringsgrad på samma hierarkiska nivå. Evaluanden kan vara den samlade beslutade rikspolitiken i en politisk sektor, ett stort programbeslut bland flera i en sektor eller en liten konkret myndighetsåtgärd i en sektor.

Måluppfyllelsemodellen kan tillämpas på alla led i implementeringskedjan. Vilka är målen för inflödet i verksamheten (eller för förvaltningen, slutprestationerna, brukarnas mottagande, brukarnas åtgärder, samhällsutfallet, utvärderingen och återkopplingen)? Hur ser det faktiska läget ut beträffande inflödet (förvaltningen, slutprestationerna, brukarnas mottagande, brukarnas åtgärder, samhällsutfallet, utvärderingen och återkopplingen)?

I en inflödesutvärdering jämförs önskat inflöde i form av pengar, kompetens och materiel med faktiskt. Hur stora intäkter förväntade sig exempelvis myndigheten kamma in genom att sälja tjänster och hur mycket flöt faktiskt in? En förvaltningsutvärdering ställer önskad omvandling mot faktisk. Vad planerade exempelvis myndighetens ledning att göra för att sprida sina intentioner ner till den operativa fältpersonalen och vad åstadkoms i praktiken? En slutprestationsutvärdering ställer frågan

om de faktiska slutprestationerna i form av tjänster eller andra styrmedel gentemot adressaterna motsvarar vad som krävs. I en brukarutvärdering ställs värdekriterier för brukarnas mottagande av styrningen, t.ex. vilken grupp av individer man ville nå fram till, mot det faktiska mottagandet, t.ex. vilka individer som faktiskt nåddes av insatserna. Åtgärdsutvärderingar jämför önskade brukaråtgärder med faktiska. I en utvärdering av samhällsutfallet tar utvärderaren reda på det faktiska utfallet och värderar det i ljuset av något bedömningskriterium. Uppgiften för en metautvärdering respektive en återkopplingsutvärdering slutligen är att undersöka om de önskemål som förelegat eller föreligger med utvärderingen respektive återkopplingen har realiserats.

Av pedagogiska skäl skall jag i fortsättningen utgå från fallet att mål avser samhällsutfall.

Måluppfyllelsemodellens styrka

Måluppfyllelsemodellen har överösts och överöses med störtfloder av kritik. Nära nog alla som förespråkar någon annan modell gör det i polemik mot måluppfyllelse. Här skall två skäl för måluppfyllelsemodellen nämnas: demokrati- och objektivitetsargumentet.

Det första skälet, demokratiargumentet, hämtar sin näring ur föreställningen om den parlamentariska styrkedjan och därmed den representativa demokratins perspektiv. Den förda politikens egna officiella mål är inte vilka hugskott som helst. De har i konstitutionell ordning fastställts i beslutande församlingar av folkets representanter. Målen är institutionaliserade och offentligt fastlagda. Detta gör att statens, landstingets eller kommunens mål för ett politikområde har en annan status och legitimitet än exempelvis intressentgruppers mål och förväntningar, riktade mot samma politikområde. Legislaturers mål får också en särställning gentemot andra intressenters genom att de kommit till under ansvar. I princip måste ju politiker ta hänsyn till inte bara vad de vill utan även vad som är resursmässigt möjligt att genomföra. Särintressen kan alltid driva krav och sätta upp mål utan att som ansvariga politiker behöva tänka på helheten och den finansiella situationen.

Detta gäller också andra organ i offentliga sektorn såsom regeringar och myndigheter, vilka handlar på delegation från beslutande församlingar. Om en myndighet beslutar inrätta ett program i syfte att uppnå vissa mål, så får dessa mål sin legitimitet av att regeringen delegerat denna beslutanderätt till myndigheten och att regeringen i sin tur fått denna rätt delegerad till sig av riksdagen.

Måluppfyllelsemodellen har alltså sina poänger ur den representativa demokratins synvinkel och därmed indirekt i ett medborgarperspektiv.

Det andra skälet, objektivitetsargumentet, baseras på föreställningen att måluppfyllelsemodellen erbjuder en objektiv lösning på måttstocksproblemet vid utvärdering. Eftersom interventionsmålen är uttryckligen angivna eller åtminstone tydligt underförstådda i texter och förarbeten, kan de i princip fastställas empiriskt.

Genom att ta dessa som värdekriterier - deskriptiv värdering - behöver utvärderaren inte ställa upp egna värdenormer vid bedömningen - preskriptiv värdering. Utvärdering med deskriptiv värderingsmetod blir en bedömning av andra ordningen, en bedömning med utgångspunkt i andras mål och värden. Då värdegrundsfrågan påstås kunna lösas objektivt, kan hela utvärderingen bedrivas objektivt. Eftersom målen är exogent givna, består utvärderingsuppgiften i att undersöka om anvisade medel bidragit till faktisk måluppfyllelse. Detta studium av medel blir ett rent empiriskt spörsmål, som kan bedrivas objektivt med vetenskapliga metoder.

Måluppfyllelseutvärdering har således starka sidor, främst ur den representativa demokratins synvinkel. Men det är också uppenbart att den lider av besvärliga brister, mot vilka läsarens uppmärksamhet nu skall riktas.

Måluppfyllelsemodellens svagheter

De viktigaste invändningar, som framförs mot måluppfyllelsemodellen som substansmodell är oklarhetsargumentet, bieffektsargumentet, argumentet om strategiska motiv och toppen-nerargumentet.

Det första skälet, oklarhetsargumentet, går ut på att insatsmål på grund av sin dimmighet är otjänliga som värderingsgrunder. Att politiska och administrativa mål är diffusa är en truism, som i den förvaltningspolitiska diskursen upprepas till leda. Här skall en suddighetsform, oprioriterade målkataloger, och de oklarheter detta skapar uppmärksammas. Problemet med målkataloger gör sig gällande dels inom ett och samma program, dels mellan program.

I samband med varje större reform redovisas regelmässigt en hel uppsättning mål. Visserligen kan måhända ett av dem pekas ut som huvudmålet men samtidigt framhålls att det skall vägas mot alla de andra, kanske potentiellt motstridande målen utan att det exakt sägs hur vägningen skall gå till. Portalparagrafen i 1979 års svenska skogsvårdslag (SFS 1979:429) innehöll en sådan oförlöst målkonflikt mellan virkesproduktion å ena sidan och naturvårdens och andra allmänna intressen å den andra. Texten löd:

Skogsmarkmeddessskogskallgenomlämpligt utnyttjande av markens virkesproducerande förmåga skötas så att den varaktigt ger en hög och värdefull virkesavkastning. Vid skötseln skall hänsyn tas till naturvårdens och andra allmänna intressen.

Mer sades inte i skogsvårdslagen. I skogsstyrelsens förordning (SKSFS 1986:6) däremot utvecklades naturvårdens och andra allmänna intressen till åtta mål (jag har markerat dem med siffror):

Skogsbruket skall bedrivas med hänsyn till skogens betydelse för (1) växter och (2) djur, för (3) vattenbalans och (4) lokalklimat samt för (5) friluftsliv och (6) rekreation. Hänsyn skall tas till (7) värdefull kulturmiljö och (8) landskapsbild.

Exakt hur avvägningen mellan virkesproduktion och de åtta andra hänsynen skulle ske sades inte. Oklarheten var säkerligen

befogad. Det är svårt att i centrala beslut på kontor i Stockholm och Jönköping specificera hur prioriteringen skall gå till vid varje enskild skogsavverkning i Ekshärad i Värmland eller Vemdalen i Härjedalen. Denna klokhet skapar emellertid problem för måluppfyllelsemodellen. Något klart förväntat utfall kan nämligen inte utpekas, vilket innebär att programmålen inte ger någon säker vägledning i utvärderingsarbetet.

Oklarhetsargumentet är en bärig invändning mot måluppfyllelsemodellen. Om politiker eller administratörer inte klarat ut vad de vill uppnå, så hamnar utvärdering enligt måluppfyllelsemodellen i ett intellektuellt moras. En lösning är naturligtvis att ta målen ett för ett och undersöka hur de uppfyllts.

Samtidigt bör understrykas att offentliga mål inte nödvändigtvis behöver vara oklara. Ibland finner politiker för gott att sätta klara, t.o.m. kvantifierbara mål, vilket vårt ovan anförda exempel 1978 års energisparplan visar. I dessa fall träffas naturligtvis inte måluppfyllelsemodellen av oklarhetsargumentet.

Ännu starkare är den andra invändningen, bieffektsargumentet. Medveten offentlig styrning får i allmänhet konsekvenser, som inte förutsågs av reformbärarna i beslutsögonblicket. »Det är svårt att sia, särskilt om framtiden«, som Niels Bohr skämtsamt uttryckte det. Om utvärderingen som i måluppfyllelsemodellen koncentreras på de reformbärande krafternas ursprungliga mål med reformen, kommer den per definition inte att spåra upp oförutsedda bieffekter. Detta kan ge en skev bild av vad reformen åstadkommit. En aktivitet som ger många spinoffeffekter måste väl rimligen vara bättre än en verksamhet, som inte utlöser någon extra energi alls.

Förekomsten av oavsiktliga bieffekter i avsiktligt handlande utgör för övrigt det kanske starkaste skälet för att överhuvudtaget göra utvärderingar. Att lägga upp dem på ett sätt som omöjliggör spårandet av sådana bieffekter vore då helt förkastligt.

Det tredje skälet mot måluppfyllelsemodellen, argumentet om strategiska motiv, påstår att denna utvärderingsform är oförmögen att hantera aktörernas strategiska överväganden. Utvärderingar som utgår från officiella sakmål ser bara toppen på ett isberg; sakmålen avslöjar blott en obetydlig del av de överväganden, som kan ligga bakom interventionen. De syften som redovisas offentligt kan vara av symbolisk karaktär och inte avsedda att förverkligas, medan de reella, outtalade motiven pekar åt andra håll. Kanske var huvudmotiven bakom reformen att hålla samman partiet, vinna väljare eller bereda marken för en koalitionsregering. När dessa mål uppnåtts, är man ointresserad av att genomföra reformen i sak.

För statsvetare tillhör denna analys av drivkrafterna bakom offentliga ingripanden det intellektuella allmängodset. Som kritik av måluppfyllelsemodellen är argumentet om strategiska motiv emellertid i ett avseende inte träffande. Rent logiskt måste det gå att bruka måluppfyllelsemodellen och ändå ta hänsyn till reformbärarnas strategiska syften. De kan nämligen användas teoretiskt, som förklaringar till varför det avsedda sakresultatet blev som det blev. Det faktum att initialt deklarerade sakmål ej blivit uppfyllda kan bero på att dolda strategiska övertygelser varit de verkliga drivfjädrarna bakom reformen. Att ge

sådana förklaringar är en huvudpoäng i utvärderingsforskningen och med en viss utvidgning skulle de kunna rymmas inom ramen för måluppfyllelsemodellen.

Det fjärde och sista skälet, toppen-ner-argumentet, säger att måluppfyllelse-modellen utgår ifrån en konventionell syn på förhållandet mellan huvudman och exekutiv enligt vilken exekutiven (förvaltningen) även i praktiken lydigt verkställer huvudmannens (politikens) beslut.

Toppen-ner-argumentet är emellertid felaktigt. Måluppfyllelseutvärdering tar inte för givet att förvaltning och andra genomförare faktiskt också är beslutsfattarnas tjänare. Det finns ingenting i måluppfyllelsemodellen som säger att insatsen gett upphov till det avsedda utfallet. Tvärtom är det just detta kausalsamband, som skall undersökas i måluppfyllelsemodellens effektanalys. Måluppfyllelsemodellen består ju av en uppsättning frågor som ställs vid utvärdering, inte en samling svar på dessa frågor. Kontentan av studien kan bli att insatsen inte haft någon inverkan på utfallet. Invändningen att modellen tar för givet att program de facto är målsökande robotar drabbar således inte måluppfyllelse-modellen ty det är just detta som modellen siktar till att granska.

Måluppfyllelseutvärderingens styrka är att den tar officiella mål på allvar. Att detta är en god egenskap hänger ihop med den representativa demokratins teori. I ett styrperspektiv behöver både medborgare och politiker målutvärdering för att kontrollera om deras exekutiver sköter sig.

Men kraftiga invändningar kan också riktas mot måluppfyllelsemodellen. Av dessa är bieffekts- och oklarhetsargumenten särskilt allvarliga. Det är nu hög tid att presentera en annan målorienterad modell, som tar hänsyn till den tunga bieffekts-kritiken - bieffektsmodellen.

Bieffektsmodellen

Margaret Thatcher lär ha sagt: »There is only one rule in politics: the unexpected will happen.« Problemet oförutsedda bieffekter skapar problem för måluppfyllelsemodellen. Hur kan icke föregripna bikonsekvenser upptäckas och värderas i en utvärdering, som tar sin utgångspunkt i och enbart är inriktad på i förväg angivna mål? För att lösa detta problem måste måluppfyllelsemodellen utvidgas till att omfatta sökande efter följder vid sidan av målområdet. Detta kallas här bieffektsmodellen .

Bieffektsmodellen liknar måluppfyllelsemodellen i så måtto att på förhand uppställda mål bibehålls som viktigaste värdekriterium. Det nya med bieffektsutvärdering är att letande inom målområdet efter resultat kompletteras med sökande utanför målområdet efter sidokonsekvenser. Att bieffektsmodellen ändå, liksom måluppfyllelsemodellen, är baserad på mål framgår av uttrycket »sidoeffekt«. Sidoeffekter kan endast finnas i relation till huvudeffekter, vilka pekas ut av avsikterna i insatsen. Men mål kompletteras som värdekriterium av kriterier för bieffekter. Den underliggande idén är att offentliga ingrepp kan åstadkomma annat än avsedda resultat. De kan ge upphov till glada överraskningar, men också skapa nya problem. Lösningar på problem blir problem, som måste lösas.

Ett exempel på en lösning som blivit ett

problem är värmepumparnas roll i svensk energiförsörjning. Med början på 1970-talet utbetalades statliga pengar till installation av värmepumpar för återvinning av spillenergi som en lösning på oljeproblemet och kärnkraftsproblemet. Värmepumparnas snabba utbredning i landet var en konsekvens av det statliga stödet. Efter en tid upptäcktes att köldmediet CFC i värmepumparna kunde läcka ut och med lång tidsfördröjning skada jordens skyddande stratosfäriska ozonlager. En liten del av skadorna på ozonskiktet kan ses som en oavsiktlig bieffekt av det svenska statliga stödet till värmepumparna. Från 1980-talets mitt blev därför värmepumparna ett miljöproblem, som måste lösas politiskt.

Att offentliga interventioner kan ge upphov till bieffekter är en självklarhet. I interventionsbärarnas perspektiv kan bieffekt definieras som en i vart fall partiell faktisk konsekvens av insatsen, som inte kan räknas till de eftersträvade huvudeffekterna. Huvudeffekter kan definieras som de faktiska effekter som i vart fall delvis produceras av insatsen och som reformbärarna medvetet önskar uppnå med insatsen i fråga. Huvudeffekt är således i just detta fall knuten till insatsens sakinriktade mål samt till vad insatsens införare trodde sig kapabla att uppnå. Vidare är den per definition både förutsedd och positivt värderad av interventionens tillskyndare.[2]

2 Vissa personer använder termer som interna och externa effekter, varvid de förra hänför sig direkt till insatsen, exempelvis eliminering av mygg som en intern effekt av ett myggutrotningsprogram medan tillkomst av ett nytt rekreationsområde blir en extern effekt. Se Vedung 1998:60.

Offentliga interventioner kan också få perversa effekter. Perversa är sådana följder av offentliga ingrepp, som blir rakt motsatta de som avsågs. Eftersom dessa självmotverkande konsekvenser uppträder inom målområdet för ingreppet, så kan de inte betecknas som bieffekter. De är heller inga huvudeffekter, eftersom de inte är medvetet önskade av interventionens tillskyndare. Perversa effekter av avsiktligt handlande bör behandlas som en separat kategori.

Perversa effekter är också något annat än nolleffekter. Nolleffekter betyder att insatsen inte får några konsekvenser alls inom målområdet. Vid perversa effekter uppträder konsekvenser av interventionen, men rakt motsatta de som avsågs.

Eftersom perversa effekter och nolleffekter uppträder inom målområdet, har måluppfyllelsemodellen inga problem med att spåra dem. Men detta betyder också att modellen inte kan upptäcka och fastställa sidoeffekter, eftersom dessa per definition faller utanför det förspecificerade målområde, som måluppfyllelsemodellen täcker.

Bieffekter kan vara förutsedda och beaktade i den kalkyl, som föregick programbeslutet. De kan vara både positiva och negativa.

Efter oljekrisen 1973 inrättades i Sverige ett nationellt stödprogram för energihushållning i bostadshus. Genom utvidgning förvandlades det år 1978 till den ovan omtalade energisparplanen för den befintliga bebyggelsen. Genom statliga bidrag och lån för tilläggsisolering av fastigheter ville regeringen Fälldin I spara bort kärnkraften. Det avsedda utfallet var naturligtvis att spara energi. År 1988 skulle

energianvändningen vara 35 TWh lägre per år i 1978 års bebyggelse. Samtidigt förutsågs att stödet skulle kunna få en ogynnsam inverkan på förmögenhetsfördelningen i befolkningen, ty välsituerade människor som villaägare och ägare av stora hyreshus kunde förväntas utnyttja medlen i högre grad än ekonomiskt svaga grupper. Det rörde sig här om en bieffekt, som reformens tillskyndare inte ville åstadkomma, men som de var beredda att ta för att befrämja det överordnade målet, energieffektivering. »Man får ta det onda med det goda«, var en fras som förekom vid denna tid.

Ytterligare en förutsedd negativ bieffekt var att kulturhistoriskt värdefull bebyggelse skulle kunna förstöras genom att putsade fasader och träpaneler kläddes in med plåt i samband med tilläggsisolering.

Samtidigt förutsågs positiva sidokonsekvenser. Stödprogrammet skulle höja den ekonomiska temperaturen och öka sysselsättningen. Genom tätning runt fönster, i väggar och på vindar skulle draget minska, vilket skulle leda till ett behagligare inomhusklimat. Väggisolering och treglasfönster skulle hålla buller ute.

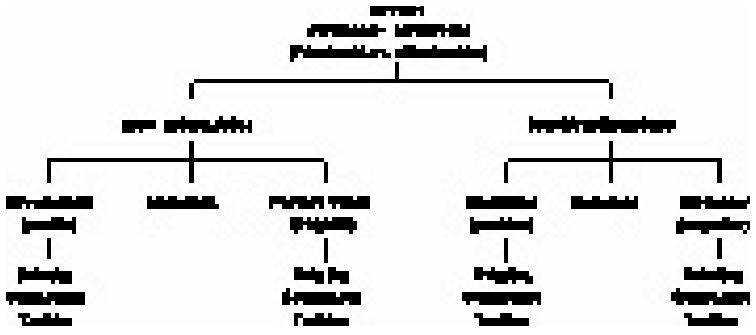
En studie som undersökte programmets effekter på energianvändning men även på förmögenhetsfördelning och sysselsättning och som dessutom uppmärksammade stadsmiljöaspekter och komfortfrågor skulle vara en bieffektsutvärdering.

Hittills har jag uppehållit mig enbart vid förutsedda effekter. Men många bieffekter är oförutsedda. Liksom sina förutsedda motsvarigheter kan de vara både positiva och negativa. Som exempel på en negativ oförutsedd sidokonsekvens kan vi anföra att tilläggsisoleringen innebar att husen blev mycket tätare vilket ledde till ökad radioaktivitet framför allt genom radon (denna oförutsedda konsekvens är särskilt intressant eftersom den första fälldinregeringen med sitt energisparprogram ville spara bort kärnkraften, som ansågs farlig genom den radioaktiva strålningen).

Resonemanget om huvudeffekter, perversa effekter, nolleffekter och bieffekter sammanfattas i klassifikationsträdet i figur 3, som visar vilka konsekvenser som kan studeras i utvärderingar.

Perversa effekter kan inträffa långt fram i styrkedjan, t.ex. i fjärde, femte eller längre

Figur 3.
Huvudeffekter, nolleffekter, perversa effekter och bieffekter



bort liggande led. Dessutom uppträder de många gånger efter mycket lång tid.

Rekommendationen att utvärderarna skall beakta perversa effekter är truistisk. Om insatsen får rakt motsatta effekter måste det vara något fel på den. Men varför är det så viktigt att beakta bieffekter? Naturligtvis därför att de är centrala element i varje helhetsbedömning av en offentlig verksamhet. Om det visar sig att sidoverkningar, som på förhand varit kända och positivt värderade, inte kommit till stånd, trots att styrningen varit i kraft under avsedd tid, så bör detta få följder för reformprogrammet. Om sysselsättningseffekten av ekonomiskt bidrag till isolering av hus är mycket mindre än man på förhand kalkylerat med, så finns det mindre anledning att behålla styrmedlet än förut, även om förväntningar om energibesparingar blivit infriade.

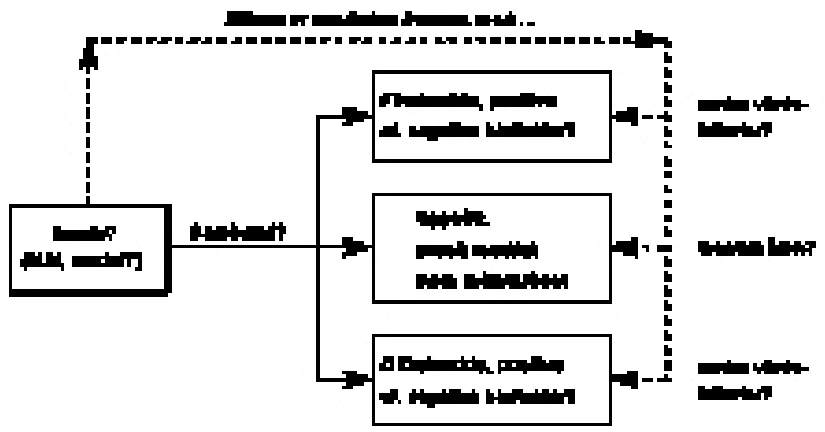
Därmed skulle vi ha avlägsnat oss långt från insatsens ursprungliga intentioner. Betoningen ligger på samtliga effekter, avsedda och oavsedda, negativa och positiva. Men fortfarande gör utvärderaren en distinktion mellan förutsedda och oförutsedda effekter, vilket förutsätter att på förhand uppställda mål används som organiserande princip. Även den bieffektsbeaktande utvärderingen är delvis målbaserad.

Om verksamhetens samtliga effekter inom och utanför målområdet undersöktes och bedömdes, skulle utvärderingens struktur på resultatsidan bli som i figur 4.

Jag rekommenderar starkt bieffektsutvärdering framför måluppfyllelseutvärdering. Ett av de kraftfullaste argumenten för att överhuvudtaget göra utvärderingar är, som redan framhållits, insikten att resultaten av offentliga ingrepp är oförutsebara, eftersom oavsiktliga händelser alltid inträffar. Det måste vara en viktig uppgift vid utvärdering att klarlägga dessa bieffekter.

En utmaning för bieffektsutvärderaren är vilka värdekriterier som skall användas

Figur 4.
Bieffektsmodellen med preciserade frågor om sidokonsekvenser



vid bedömningen. Idealt skulle hon vilja balansera värdet av den avsedda huvudeffekten mot värdet av de lyckliga och skadliga bieffekterna. Negativa sidoeffekter skulle då kunna bli en motvilligt accepterad kostnad för att uppnå den övergripande, värdefulla effekten. Starka positiva bieffekter skulle också öka godtagbarheten av program med svag måluppfyllelse. För att kunna genomföra denna kalkyl krävs kriterier för huvudeffekten, varje typ av bieffekt samt för avvägningen dem emellan.

Den deskriptiva värderingsteori som anvisas av måluppfyllelsemodellen - att insatsens egna mål skall utgöra måttstocken - är uppenbart otillräcklig av följande skäl. Om vissa bieffekter inte är förutsedda, så är naturligtvis varken värdekriterier för bedömningen av dessa var och en för sig eller för sammanvägningen av dem till ett allomfattande värde på hela insatsen inte angivna. Därför är officiella mål otillräckliga för att värdera oförutsedda bieffekter.

En möjlig lösning på kriterieproblemet ligger i en mer kreativ tillnärmelse till värderandet. Utvärderaren kan kartlägga om huvudeffekten och förutsedda bieffekter verkligen inträffat och värdera dem i ljuset av angivna mål. Vid sidan härav kan hon också lokalisera oförutsedda bieffekter men överlåta till olika användare av utvärderingen att själva göra den sammanvägande värdebedömningen. Eftersom modellen litar till andras värdekriterier, vilar den fortfarande på en deskriptiv värderingsteori. Men beaktandet av oförutsedda bieffekter tvingar utvärderingens nyttjare att beakta andra värden än dem, som från begynnelsen var inkorporerade i interventionen.

Brukarmodeller

Brukarorienterade utvärderingsmodeller inriktar sig på mötet mellan verksamheten och dess klienter. Huvudfrågan är om den offentliga verksamheten i leveransögonblicket motsvarar den kvalitet, som mottagaren kräver eller behöver. Brukarmodeller använder antingen adressaternas förväntningar (önskemål) eller deras behov som normativt bedömningskriterium. Att utvärdera mot behov och att utvärdera mot förväntningar är emellertid principiellt skilda saker, eftersom de förra ofta anses vara något annat än utsagda önskemål. För enkelhetens skull utgår jag ifrån att brukarmodellen tar sin utgångspunkt i slutmottagarnas önskemål och förväntningar.

En parallell till konsumenters inköp på en marknad anses ibland kasta ljus över brukarutvärdering. När konsumenten köper i butiken, är det inte producentens syfte utan hennes egen bedömning av varans värde som avgör. Brukarorienterad utvärdering bygger på föreställningen att förvaltningen producerar för kunder på en marknad.

Trots att svenska språket rymmer uttryck som »fängelsekund«, kan parallellen med kunders inköp på marknadsplatsen inte drivas för långt. I motsats till kunden betalar brukaren oftast inte personligen för den service hon får, i vart fall inte i full utsträckning, utan det får de skattskyldiga göra. I tänkandet kring brukare ingår därför också en deltagardemokratisk aspekt, som innebär att brukarna skall kunna göra sina stämmor hörda gentemot serviceproducenten och därmed påverka eller ta ansvar för tjänstens innehåll.

Brukarmodeller innehåller ett diskursivt, resonerande, diskuterande och influerande moment, som bör få komma till uttryck vid brukarorienterad utvärdering. Brukarmodeller hänger ihop med ett brukardemokratiskt synsätt. Om måluppfyllelsemodellen bygger på den representativa demokratins teori innehåller brukarmodellerna inslag av participativ demokratisk teori.

Det finns många varianter av brukarorienterad utvärdering. Brukarmodeller kan använda adressaternas förväntningar (önskemål) eller deras behov som normativt bedömningskriterium. Att utvärdera mot behov och att utvärdera mot förväntningar är principiellt skilda saker, eftersom de förra ofta anses vara något annat än utsagda önskemål. Brukarutvärdering kan utföras av brukarna själva som självvärdering men också beställas och utföras uppifrån. Låt mig resonera utifrån det senare fallet.

Vid brukarevaluering är utvärderarens första uppgift att lokalisera klienterna. Ibland finns det flera klientgrupper; vid utvärdering av akademisk utbildning kan vi dels tänka oss studenterna, dels studenternas framtida arbetsgivare som klienter. Eftersom samtliga troligen inte kan kontaktas, måste ett urval göras. Sedan pekar den klientdrivna modellen i olika riktningar. En möjlighet är att undersöka om den faktiska klientgruppen avviker från den avsedda, en annan att granska brukarnas bedömningar av insatsen. Brukarmodellerna säger ingenting om vilka verksamhetskomponenter som bör studeras. Det mesta är tillåtet. Klienterna kan bli ombedda att fälla omdömen om utfallet, slutprestationerna, tjänstens tillgänglighet, tjänstens kvalitet, tjänstens administration eller tjänstens kostnader. Ett annat karakteristiskt drag är att effektfrågan, som är så central i måluppfyllelseutvärdering och bieffektsutvärdering, ibland inte reses i brukarorienterad utvärdering. Utvärderarna nöjer sig med att be klienterna värdera någon aspekt på den service de får.

Ett annat notabelt inslag i brukarmodellen är värdepluralism. Olika brukare bedömer servicen utifrån skilda värdegrunder. De kan vara djupt oeniga vid bedömningarna. Brukarmodellen tillåter en konfliktsyn på offentliga program och deras mottagande.

Det har tagit oerhört lång tid innan brukarna upptäcktes i offentlig förvaltning. Idéerna om den politikerstyrda förvaltningen och den parlamentariska styrkedjan - förvaltningen som politikernas exekutiv - har dominerat tänkandet. Numera används emellertid brukarmodeller i en rad utvärderingssammanhang, särskilt när det gäller offentlig serviceproduktion inom barnomsorg, sjukvård, åldringsvård och biblioteksväsen eller program där adressaternas aktiva deltagande är nödvändigt för hela verksamheten. Vid universitet får de studerande lämna synpunkter på kurslitteraturen och undervisningen. De bedömer lärarnas prestationer utifrån exempelvis systematisering av lärostoffet, inställning till debatt och diskussion, förmåga att stimulera studenternas intresse, motivation och självständiga tänkande, omsorg om studenterna och entusiasm.

Här följer några exempel på marknadsundersökningar som görs eller har gjorts inom den statliga sektorn: Integrationsverket undersöker hur invandrare ser på den offentliga introduktionsverksamheten; Riksskatteverket vill veta hur de skattskyldiga ser på skatteväsendet, bemötandet,

personalens kompetens och utförandet av olika tjänster; Arbetsmarknadsstyrelsen frågar hur kunderna upplever arbetsförmedlingarnas service; Byggnadsstyrelsen vill ha sina kunders (hyresgästers) syn på hur verket sköter lokalförsörjningen; socialdepartementet spörjer ett stort antal föräldrar i hela landet hur de ser på barnomsorgen; Naturvårdsverket undersöker vilken trovärdighet miljömyndigheterna har hos allmänheten; och Patentverket tar reda på hur uppfinnarna upplever kötiderna.

För att ytterligare konkretisera kan vi ta biblioteksundersökningar som exempel. Besökare kan tillfrågas om bibliotekets tillgänglighet, bokbestånd, utrymmen och serviceanda: a) är öppettiderna godtagbara eller har biblioteket stängt när folk är lediga från arbetet och skulle kunna låna? b) är böckerna uppställda på ett begripligt sätt? c) finns den rätta blandningen av gammalt och nytt, barnböcker och vuxenlitteratur? d) är reservationstiderna för långa? e) tar det för lång tid innan återlämnade böcker utplaceras i hyllorna? f) är lokalerna tillräckligt rymliga? g) är ljusförhållandena tillfredsställande? h) blir låntagarna trevligt bemötta av personalen?

Brukarmodeller bör ses som ett komplement till de tidigare presenterade modellerna, eftersom de reser andra problem. Kravet att förvaltningen skall vara brukarorienterad även om brukarna inte betalar för servicen är sunt i vissa avseenden och vissa fall. Det förefaller väldigt rimligt att ta hänsyn till boklånarna när det gäller öppettider och personligt bemötande, för att ta två exempel. Generellt bör brukaråsikter få spela en större roll vid bedömningen av offentlig service än vid utvärdering av myndighetsutövning eller myndighetsinformation.

Men brukarkrav kan inte i alla lägen tillåtas väga tyngre än principen att förvaltningen skall vara följsam mot de folkvalda och därmed ytterst medborgarna. Detta beror på att brukarmodeller har begränsningar. Enligt teorin om rationella val vill brukarna ha någonting för ingenting. De vill ha bättre service därför att de själva kommer i åtnjutande av den samtidigt som de i stor utsträckning slipper betala eftersom kostnaderna sprids ut över en stor mängd skattebetalare. Brukarna har fiskala illusioner. De har begränsad information om vad tjänsterna kostar, vilket gör att de underskattar kostnaderna samtidigt som de kräver mer tjänster. Brukarna kan således inte förväntas anlägga ett totalt kostnads-intäktsperspektiv, vilket det politiska systemet tvingas göra. Detta gör att även ekonomiska modeller har en legitim roll att spela i offentlig politik, ett spörsmål till vilket vi skall återkomma.

Intressentmodeller

Intressentmodellernas värdekriterier är farhågor och undringar hos interventionens berörda. Detta är något helt annat än att ta sin utgångspunkt i insatsens mål. Det är också något annat än att fråga brukarna. Skillnaden är att intressentmodeller i princip beaktar samtliga berörda under det att måluppfyllelsemodellen och brukarmodellerna tar hänsyn endast till vardera en grupp, nämligen beslutsfattare respektive brukare.

En uppställning av tänkbara intressenter i ett offentligt program presenteras i figur 5.

Figur 5.
Intressenter i en offentlig insats och dess utvärdering

INSATSINTRESSENTER:	
Medborgarna:	Folket som väljer representanter till beslutande församlingar på olika nivåer i det politiska systemet
Högsta beslutsfattare:	Makthavare direkt ansvariga för att insatsen införts, fortsättes, trappas av eller avslutas
Politiska opponenter:	Insatsens politiska motståndare
Myndighetsledning:	Personer ansvariga för ledning och samordning av insatsens förvaltning
Handläggare centralt:	Personer som på central nivå direkt arbetar med insatsens förvaltning
Reg. o lokala myndigheter:	Personer som regionalt och lokalt är delaktiga i insatsens förvaltning
Gräsrotsbyråkrater:	Operativ fältpersonal
Enskilda mellanhänder:	Enskilda som på uppdrag av offentliga instanser eller på annat sätt deltar i insatsens förvaltning
Slutmottagare:	Insatsens målgrupp (adressater, brukare, klienter)
Andra slutmottagare:	Aktörer, kollektiva eller individuella, som (oavsiktligt) berörs av insatsen utan att vara avsedda slutmottagare

De tio grupper som listas i figur 5 representerar de viktigaste parter, som berörs av en insats i offentliga sektorn. Listan är kompilerad med tanke på en nationell reform, beslutad av högsta legislaturen. Självfallet skulle en förteckning över intressenter i kommunala verksamheter se annorlunda ut. Intressentmodellen är för övrigt öppen och ger inget klart svar på frågan vilka de berörda är.

Intressentutvärdering, som på engelska kallas »stakeholder approach«, kan ske på olika sätt. Intressenterna kan själva utgöra utvärderingslaget och själva exekvera utvärderingen. Vi kallar detta den svenska SOU-modellen. Utvärderingen kan också utföras av särskilt utsedda utvärderare, som kan vara fristående forskare, konsulter eller anställda vid myndigheten. Poängen är att alla intressenter skall beaktas men att själva utvärderingen utförs av särskilda utvärderare. Detta är den nordamerikanska intressentmodellen. I fortsättningen skall jag resonera utifrån den nordamerikanska modellen.

Alla tio intressentgrupper i figur 5 är tänkbara bidragsgivare i en utvärdering. Även om utvärderaren i praktiken måste göra ett urval förutsätter den fortsatta diskussionen att samtliga intressenters förväntningar beaktas i intressentmodellerna. Intressentansatsen tolkas som en holistiskt, inriktad på hela spektrum av berörda.

Utvärderaren börjar med att ta reda på vilka som har intressen i insatsens tillkomst, genomförande och konsekvenser. Hon spårar upp insatsens utarbetare, finansiärer och genomförare, ringar in målgruppen och de personer och institutioner, som drabbas av kostnader och sidokonsekvenser. Hon pekar ut människor, som utan att veta om det har intresse i interventionen.

Sedan de berörda blivit identifierade, kan den nordamerikanska intressentmodellen fungera på olika sätt. I Effective Evaluation föreslår Guba och Lincoln (se Vedung 1998:77) att alla berördas »farhågor« (concerns) och »frågor« (issues) får bilda utgångspunkten. En »farhåga« kan vara något som en part befarar kan medföra nackdelar eller något som den gärna vill få närmare belagt. »Virtually any claim, doubt, fear, anticipated difficulty, and the like expressed by anyone with a legitimate basis for making such a representation could be entertained as a concern.« En »fråga« å andra sidan är »any statement, or focus that allows for the presentation of different points of view; any proposition about which reasonable persons may disagree; or any point of contention«.

För att komma fram till en lämplig uppläggning av undersökningen utesluter intressentmodellens förespråkare inget tillvägagångssätt. De hyser emellertid förkärlek för kvalitativa metoder. Nyckelorden är direkt observation, samtal, dialog och växelverkande sökprocess. Utvärderaren måste observera och samtala med klienter, programadministratörer och andra intressenter. Vad hon ser eller får till svar inverkar på vad hon gör därnäst. Efter en tid börjar hon kanske upptäcka både uppgivna och verkliga syften med verksamheten och de farhågor och frågor, som olika intressenter har rörande både styrningen och utvärderingen. Efterhand blir utvärderaren mer involverad och kan själv börja argumentera för vilka farhågor och frågor, som skall tas med i undersökningen. Den slutliga inriktningen bestäms av utvärderaren. För varje farhåga och varje fråga specificerar utvärderaren vilken typ av information som behövs och hur den skall samlas in.

Det är typiskt för intressentmodellen att utvärderaren tänkes leta ganska länge efter problemen eller uppläggningen. Måluppfyllelse t.ex. pekar på ett helt annat sätt än intressentmodellen ut vad utvärderingen skall handla om. Den nordamerikanska intressentmodellen är påfallande öppen. Tanken är att utvärderaren måste vara lyhörd för de berördas farhågor och undringar och låta dessa styra nästa steg i sökprocessen. Hon skall genom växelverkande kommunikation med de berörda successivt komma fram till vilka farhågor och frågor, som skall tas på allvar och utredas. Utvärderingens uppläggning bestäms gradvis. Intressentutvärdering är lyhördhetsutvärdering.

För den slutliga datafångsten kan alla strategier komma i fråga. Men för det mesta föredras observationsmetoder och interrogativa metoder framför dokumentmetoder. I många fall sätter intressentutvärderarna mest värde på direkt observation genom platsbesök och personliga iakttagelser. Djupintervjuer är en annan favoriserad metod för materialinsamling. Poängen är att programmet studeras i sitt naturliga sammanhang utan kontrollgrupp. Det skall vara fallstudier. Resultatrapportering, slutligen, kan variera från berörd till berörd. Nyckelordet tycks vara »porträttering«, som kan vara skriftlig men lika gärna muntlig eller visuell. Det sammanhängande helhetsintryck som förmedlas av ett porträtt är viktigt. Utvärderarens egna bedömningar spelar en mindre roll eftersom både val av värdegrunder och värderandet skall skötas av intressenterna. Normalt måste det bli

fråga om flera värdegrunder och många olika bedömningar. Intressentmodellen blir därmed i praktiken utpräglat pluralistisk. Till sist avgörs om resultaten skall rapporteras i en monografi, vilket inte alls är nödvändigt.

Intressentmodellen har många förtjänster. Demokratiska argument, kunskapsargument, användningsskäl och målhanteringsargument talar till dess fördel.

De demokratiska argumenten för intressentmodellen utgår från participativa och deliberativa synsätt. Demokrati innebär inte bara att medborgarna i val röstar fram eliter, som skall fatta beslut på medborgarnas vägnar. Medborgarna bör också få delta i offentligt beslutsfattande mellan valen. Dessutom är deliberation och debatt viktiga värden i en demokrati ty härigenom kan människor få klarhet i sina preferenser. I högre grad än måluppfyllelsemodellen och bieffektsmodellen tillgodoser intressentmodellerna dessa participativa och deliberativa värden

Enligt kunskapsargumentet vore det dåraktigt av utvärderaren att avskärma sig från de insikter i reformen, som dess berörda besitter. Intressenterna hyser dessutom ofta farhågor om bieffekter, implementeringsbarriärer och samspel med andra program, som kan ge ytterst värdefulla uppslag för fortsatt undersökning. Det är därför lätt att instämma i rekommendationen att nästan varje utvärdering bör inledas med en aktörsbestämning och intervjurundor.

Detta kunskapsskäl är emellertid inget huvudargument för intressentmodellen. Ett sådant är däremot påståendet att den ökar resultatens användbarhet. I den amerikanska debatten kring utvärdering diskuteras intressentmodellen helt och hållet i detta avnämarperspektiv. Resultat som tas fram i måluppfyllelseutvärderingar och bieffektsgranskningar kommer sällan till användning. Rapporterna staplas på varandra i beslutsfattarnas bokhyllor, olästa och bortglömda. Även om viljan är god så nyttjar parlamentariker, generaldirektörer, byråchefer och lokala politiker inte det framtagna informationsunderlaget till att fatta mer genomtänkta beslut. Utvärdering lider, så kan kritiken formuleras, av ett elakartat skrivbordslådesyndrom.

Detta till synes grovt irrationella handlande har förbryllat utvärderingsforskarna. Varför beter sig de styrande så oklokt? Huvudanledningen sägs vara att utredarna arbetar i suverän isolering, utan samfärdsel med avnämarna. Därför tenderar de att penetrera problem, som intressenterna inte känner något behov av att få utredda. Intressentmodellen ökar sannolikheten för att frågor tas upp som berörda grupper finner relevanta. Därmed kommer resultaten lättare till verklig nytta.

Intressentmodellen kan också ses som en strategi att hantera notoriskt oklara och osammanvägbara mål. Intressentmodellen erbjuder en praktisk lösning på bekymret med många mål samtidigt. Intressentmodellen hör ihop med en intressegruppsdemokratisk syn på offentlig verksamhet.

Målen för offentliga verksamheter är nästan undantagslöst diffusa. Framför allt kännetecknas de av att många olika mål skall balanseras mot varandra utan att någon på förhand kan avgöra hur detta skall gå till. Avvägningen kan därför inte göras på vetenskaplig väg utan måste ses som en politisk uppgift. Därför gäller det att låta

de många intressenternas många mål på många områden bilda utgångspunkter för utvärderingen. Att intressenterna då visar sig djupt oense måste anses normalt och naturligt i en fri stat. Intressentmodellerna är till sin natur politiska och pluralistiska.

Det finns också uppenbara nackdelar med intressentmodellerna som substansmodeller. De är kostsamma och omständliga eftersom varje intressent måste kontaktas och utfrågas.

Intressentmodellerna är oskarpa (Karlsson 1995). De har inget entydigt svar på frågan vilka som egentligen är berörda av ett offentligt program. De skulle behöva kompletteras med en politisk teori om urval av berörda. Intressentmodellerna likställer vidare alla bemälda, hur de nu än valts ut. Men i ett konstitutionellt politiskt system måste väl ändå de folkvalda politikerna - de överordnade beslutsfattarna - vara viktigare än experter på sakområdet eller oavsedda slutmottagare? Intressentmodellen saknar regler för hur dylika prioriteringar bör gå till. Den är ett uttryck för kålsuperi. Risken finns att de bäst organiserade och mest engagerade intressenterna konsulteras medan vaga och mer svårdefinierbara intressenter lämnas utanför. Även på denna punkt skulle intressentmodellerna behöva baseras på en mer problematiserad politisk filosofi. Intressentmodellerna och deras intressegruppsdemokratiska grundval kan kommaikonfliktmedmåluppfyllelsemodellens synsätt, som hänger ihop med föreställningen om representativ demokrati.

Samtidigt skall konstateras att intressentmodellen ger utvärderaren möjlighet att gå förbi de stora korporativa, byråkratiska och partipolitiska etablissemangen på nationell nivå och koncentrera uppmärksamheten på nätverken i de lokala sammanhangen. Detta är en synpunkt som särskilt framhållits av den amerikanska utvärderingsforskaren Robert Stake.

Intressentmodellen reser frågetecken kring vetenskaplig objektivitet i resultatredovisningen. Hur skall utvärderaren handla om hennes ståndpunkt går stick i stäv mot mäktiga intressenters? Fakta kan bli en fråga om förhandlingsstyrka snarare än rationell diskussion och empiriska belägg. Det finns också risk att utvärderingen sker mot utvärderarens egna mål på ett icke redovisat sätt.

Intressentmodellen är alltså suddig i konturerna och relativt kontroversiell. Trots detta talar demokratiska argument, användningsargument och mångmålsargument till dess fördel. Ett annat starkt skäl för den är kunskapsargumentet: att den kan användas som sökstrategi för att komma fram till en lämplig uppläggning i utvärderingens inledningsskede, när det gäller att i grova drag få grepp om reformens innebörd, genomförande och utfall.

Professionella modeller - kollegebedömning

Särskilt i utvärdering av högre utbildning och forskning har det utbildats ett ganska unikt tillvägagångssätt - kollegebedömning - som skiljer sig kraftigt från särskilt mål- och bieffektsmodellerna men även de ekonomiska modellerna som vi senare skall ta upp. Kollegebedömning innebär att medlemmar av en profession utses att utvärdera professionsmedlemmarnas

arbete och verksamhet utifrån yrkeskårens egna kvalitetskriterier. Advokater utvärderar advokater, kirurger kirurger, ingenjörer ingenjörer och professorer professorer. Utvärderingen utförs av kolleger, som ju per definition är professionella likar. Även professonell utvärdering kan utföras på flera sätt. Självvärdering är en form. Den professionelle utvärderar då sina egna insatser eller professionella på en institution utvärderar tillsammans institionens insatser. En annan form är kollegial utvärdering genom utomstående likar. Idealt sett bör likarna i detta fall vara lite bättre inom sitt område än de kolleger som skall utvärderas. De utvärderades förtroende för utvärderarna är viktigt. Ofta kombineras den interna, självvärderande typen av professionell utvärdering med den externa.

Den mest bekanta kollegemodellen är s.k. peer review, som används vid utvärdering av vetenskap. Tanken är att renommerade forskare på ett område får i uppgift att utvärdera hur ett forskningsprojekt, ett forskningsprogram, en universitetsinstitution eller en hel disciplin i ett land står sig kvalitetsmässigt och relevansmässigt. Vetenskapsmännen och deras insatser utvärderas av kolleger (för exempel, se Allardt m.fl. 1987 sociologi, Danielsen m.fl. 1988 historia, Öhman & Öhngren 1991 psykologi och Engwall 1992 ekonomi).

Termen »peer review« har på senare tid vanligen syftat på en procedur för att välja ut bidrag till vetenskapliga tidskrifter. Till en tidskriftsredaktion sänds anonymt artiklar som redaktionen vidarebefordrar till någon kollega för peer review i syfte att utröna om artiklarna håller sådan kvalitet att de förtjänar att publiceras. Peer review används också för att ge råd till forskningsfonder om vilka forskningsprojekt som skall ges stöd. Ansökningarna skickas ut till en grupp kolleger för granskning. Kollegial bedömning innebär i detta fall att forskningen bedöms i förväg, ex ante. Det är forskningsplanen som värderas, inte de uppnådda resultaten. Peer review har vidare använts för att utreda och bedöma brott mot god forskningsetik och gott forskaruppträdande. Här avser jag med kollegial bedömning en procedur för att i efterhand utreda och bedöma genomförande, prestationer och utfall i offentlig verksamhet.

Kollegebedömning pekar inte i detalj ut de kriterier och standarder som bör vara vägledande vid granskningen eller sakområden som bör granskas. Frågan om värdekriterier och standarder tillåts variera från profession till profession.

Vad har egentligen kollegebedömning med utvärdering i offentliga sektorn att göra, kanske någon undrar. Svaret ligger förborgat i principen om den professionsstyrda förvaltningen. På vissa områden i den offentliga sektorn är målsättningar och kvalitetsnormer så invecklade och föränderliga att politikerna funnit för gott att överlämna deras närmare utformning till vetenskapligt utbildade professioner. Professionens egna medlemmar får genom kritisk debatt och diskussion utveckla bedömningsgrunder och kvalitetskriterier. Exempel på professionella som denna ställning är arkitekter, domare, professorer, lärare, läkare, veterinärer och ingenjörer. Detta gör det naturligt att också anförtro utvärderandet åt professionerna.

Utvärdering av sektorsforskning och

universitetsvetenskap kan ses som bedömning av statens forskningspolitik. Forskning ombesörjs inte av en webersk byråkrati, ej heller av korporationer eller politiker. Den sköts av professionella vetenskapsidkare. Detta får stundom till följd att även forskningsutvärdering sker i kollegiala former. Vid forskningsutvärdering arbetar experterna oftast på någons uppdrag, t.ex. ett forskningsråds. Forskningsrådet skriver direktiv, som i allmänna termer innehåller de problem och frågor man vill ha belysta. De utvärderingar som startades under 1987 av Byggforskningsrådets vetenskapliga nämnd uppmanades beakta

a relevansen av problemval och analysuppläggning
b analysmetodernas tillämplighet
c hållbarhet hos genomförda resonemang och konklusioner
d arbetets inplacering i förhållande till diskursen på området
e möjligheterna att praktiskt tillämpa forskningsresultaten
f värdet av den genom projektet eventuellt uppbyggda forskningsmiljön
g överensstämmelse mellan ursprungliga intentioner och slutligt resultat.

Därefter kontaktas lämpliga sakkunniga, vilka bör ha högre kompetens inom området än kollegerna som skall utvärderas. Experterna måste vara oberoende av de utvärderade; de bör inte ha samarbetat med varandra inom det aktuella sakområdet. En viktig skillnad går mellan kollegebedömningar där de utvärderade får föreslå och på förhand godkänna urvalet av kolleger och sådana där detta inte sker.

Sedan expertgruppen - kollegiet - utsetts organiseras det fortsatta arbetet. Oftast ombes de utvärderade sända in relevant forskningsmaterial, så att de sakkunniga kan läsa in sig. Detta tyder på att någon form av självvärdering ligger till grund för den kommande kollegebedömningen. Därefter besöks varje forskare och forskargrupp för presentationer och samtal. Kollegegruppen diskuterar kanske vilka kvalitetsgrunder som skall appliceras på forskningsresultaten eller forskningsprocesserna. Efter en tid skrivs en preliminär rapport. Den cirkuleras till berörda forskare som kan beredas tillfälle att yttra sig. Därefter komponeras den värderande slutprodukten. Men även om utvärderingskollegiet interagerar med de forskare som utvärderas så är det kollegiet som är ansvarigt för slutrapporten och därmed bedömningen. Hela proceduren brukar vid Byggforskningsrådetsvetenskapliganämnd ta ca 18 månader.

Kollegebedömningens styrka ligger i val och tillämpning av värdekriterier. För detta bör de professionella själva vara mest lämpade. Det är svårt för politiker att sätta kvalitetsmål på forskning. Dessutom torde måttstockarna liksom forskningen utvecklas över tid i takt med kunskapsmassans tillväxt. Utvärderingar bör emellertid inte begränsa sig till problemet om forskningen motsvarar krav på kvalitet i resultat eller i process utan även komma med förklaringar av varför exempelvis resultat inte motsvarar eller inte uppfyller krav på kvalitet. Otvivelaktigt torde kolleger t.ex. från ett annat land vara skickade att ställa upp lämpliga kvalitetskriterier och utföra bedömningar med ledning av dessa. Mer

tveksamt är om de också är mer kompetenta än andra att leta upp förklaringar. I den mån framgång och misslyckande beror på verksamhetens organisering borde kollegerna kanske kompletteras med experter på forskningsorganisation.

Man skulle på lite tydligare sätt kunna kombinera kollegebedömning med självutvärdering. De utvärderade skulle då först ombes bedöma värdet av sin egen produktion. På grundval av dessa bedömningar men naturligtvis även från egna utgångspunkter skulle sedan de sakkunniga göra sin betygsättning. Innan de slutliga omdömena slås fast skulle de utvärderade beredas tillfälle att yttra sig över de sakkunnigas resultat.

Ekonomiska modeller - produktivitet

Gemensamt för alla substansmodeller är att de inte tar in kostnader i kalkylen. Samtidigt som de undersöker kriterieuppfyllelse, oavsett om det är insatsmål eller brukares och intressenters önskemål och farhågor, så beaktar de inte uppoffringar för att uppnå värdenormerna. Hänsynstagande till kostnader å andra sidan ingår alltid i de ekonomiska utvärderingsmodellerna. Den rena kostnadsmodellen som enbart beaktar kostnader skall inte beröras här. Sökarljuset skall riktas mot produktivitet och effektivitet, två ekonomiska modeller som beaktar såväl substans som kostnader.

Produktivitet definieras som prestationer dividerade med kostnader. Den ofta citerade svenska programbudgetutredningen formulerade denna tanke så här:

Med produktivitet avses alltså förhållandet mellan prestationer och uppoffringar inom en myndighet eller en del därav, dvs. produktiviteten utgör kvoten mellan ett mått på mängden utförda prestationer och ett mått på använda produktionsresurser (output dividerad med input).

Produktivitet kan exakt uttryckas med den matematiska formeln i figur 6.

Figur 6.
Begreppet produktivitet

$$\text{Produktivitet} = \frac{\text{prestationer (output)}}{\text{insatser (input)}}$$

Kvotformeln i figur 6 kan självklart operationaliseras på olika sätt. Låt mig ta exempel från en finsk biblioteksstudie (Vedung 1998:81). Vid analysen av produktivitetsutvecklingen i finska kommunbibliotek användes följande mått:

$$\frac{\text{antalet boklån}}{\text{boklånens kostn. i mark}} = \text{kostnadsprod.}$$

Ett alternativ till kostnadsproduktivitet kunde ha varit arbetsproduktivitet, vilket kan illustreras med uttrycket:

$$\frac{\text{antalet boklån}}{\text{antal arbetade timmar}} = \text{arbetsprod.}$$

Skillnaden är att kostnaderna i det förra fallet anges i pengar, i det senare fallet i antal arbetstimmar, dvs. i fysiska termer. Det bör understrykas, att kostnader kan

kalkyleras på båda sätten vid produktivitetsanalys. Andra valmöjligheter som kan övervägas är antal böcker i biblioteket: kostnader, antal invånare i kommunen: kostnader och antal låntagare:kostnader. Tidsenheten kan vara budgetår, kalenderår, eller eventuellt också kalendermånad.

Semantiskt och retoriskt intressant med produktivitet är att den också betecknas som »inre effektivitet«. Detta återspeglar förhållandet att produktivitet är ett internt mått. Produktivitet fångar ett förhållande inom organisationen, men beaktar inte verksamhetens effekter på omgivningen eller omgivningens bedömning av effekterna.

För att säga något klokt om en myndighets produktivitet behövs referensfall. Ett biblioteks produktivitet måste jämföras med något för att ge ett intryck av hur god produktiviteten verkligen är. Detta kan tekniskt uttryckas så att utvärderarna måste ha standarder på produktivitetskriteriet för att ange vad hög och låg produktivitet verkligen innebär. Åtskilliga standarder brukas: jämförelse med tidigare prestationer, med liknande institutioner i landet, liknande institutioner i andra länder, brukarförväntningar och intressentmål.

Låt oss återvända till biblioteken för att illustrera problemen med att hitta giltiga indikatorer på slutprestationer. Är verkligen antalet utlånade böcker ett relevant och heltäckande prestationsmått? Relevant är det nog. Att låna ut böcker är rimligen bibliotekets viktigaste uppgift. Men heltäckande är måttet inte. En finsk utredning konkluderar att endast 30-50 procent av bibliotekens klienter är boklånare. De övriga besöker biblioteket för att läsa dagstidningar och tidskrifter men detta avsätter inga spår i utlåningsstatistiken. De frekventerar referensavdelningen för att slå i ordböcker och encyklopedier eller musikavdelningen för att lyssna. Biblioteken svarar också för informationsservice. I den finska undersökningen diskuterades självfallet problemen innan forskaren beslöt sig för att stanna för boklån. Att det är en begränsning går dock inte att komma ifrån.

Den finska studien skulle naturligtvis ha kunnat använda alla dessa delmått och försökt aggregera ihop dem till ett mått på den totala produktiviteten. Då skulle undersökarna emellertid ha tvingats ta ställning till besvärliga viktningsproblem.

En annan svårighet vid produktivitetsutvärdering är att kvaliteter ofta förbises. Böcker är av varierande kvalitet. Hur kan detta mätas i en produktivitetsutvärdering?

Produktivitet som mått på offentliga sektorns varor och tjänster kan också kritiseras från principiella utgångspunkter. Viktigast är kritiken att det mäter ett internt förhållande inom organisationen och inte det vi helst vill veta, nämligen de resultat den offentliga interventionen gett upphov till hos slutmottagarna eller ute i samhället, vad detta resultat är värt samt om förtjänsterna uppväger uppoffringarna. I biblioteksfallet är det inte boklånen i och för sig som är intressanta; folk kan ju låna böcker, lägga dem på hög i hemmet och efter påstötningar lämna dem tillbaka olästa. Viktigare är läsandet av de lånade böckerna. Avgörande är emellertid läsarnas upplevelser av läsandet. Upplevelserna kan delas in i bildning och förströelse. Det som

den kostnadsmedvetne, finkulturellt inriktade utvärderaren kanske vill åt är

$$\frac{\text{värdet av bildning och förströelse genom lånade böcker}}{\text{kostnader för att förvärva denna bildning}}$$

Därmed är hon inne i ett effektivitetsresonemang.

Förespråkarna för produktivitetsmätningar kan inte komma ifrån att produktivitet mäter slutprestationer och att detta inte är någon idealisk mätpunkt för att bedöma den offentliga sektorns verksamheter. Den offentliga institutionen kan ju göra fel saker, dvs. slutprestationerna kanske inte leder till det önskade utfallet.

Ekonomiska modeller - effektivitet

De andra ekonomiska modeller som används vid utvärdering av offentlig verksamhet är effektivitetsmodeller . Effektivitet (»efficiency«) mäts i utfallsledet och kan registreras på två sätt: i kostnads-intäktsanalys (K-I-effektivitet, »cost-benefit efficiency«) och i kostnads-effektivitetsanalys eller synonymt kostnads-nyttoanalys (» cost-utility«).

Om effektivitet mäts i en kostnads-intäktsanalys ställs insatsens intäkt i kronor i förhållande till kostnaden i kronor för intäktens förvärvande. K-I-effektivitet kan anges som kvoten av det monetariserade värdet av de resultat som verksamheten producerat och de monetariserade kostnaderna. Om effektivitet däremot likställs med vad som mäts i kostnads-effektivitetsanalyser, ställs insatsens fysiska avkastning i förhållande till kostnaden för avkastningens förvärvande. Vid kostnads-effektivitetsanalys beaktas med andra ord de monetariserade kostnaderna precis som vid kostnads-intäktsanalys medan däremot värdet av effekterna anges endast i fysiska termer. De två effektivitetsmåtten kan uttryckas med formlerna i figur 7.

Det vi i formlerna menar med »insatsens effekter« är sådant, som just insatsen gett upphov till. Det vi är ute efter är de effekter, som just insatsen åstadkommit.

Som alla andra utvärderingsmodeller ger produktivitets- och effektivitetsmät-

Figur 7.
Effektivitet mätt genom kostnads-intäktsanalys och kostnads-effektivitetsanalys

$$\text{Effektivitet (kostnads-intäktsanalys)} = \frac{\text{värdet av insatsens effekter (i kr, euro)}}{\text{insatsens kostnader (i kr, euro)}}$$

$$\text{Effektivitet (kostnads-effektivitets-analys, kostnads-nyttoanalys)} = \frac{\text{insatsens effekter i fysiska termer}}{\text{insatsens kostnader (i kr, euro)}}$$

ningar endast delperspektiv. De bortser från andra krav, som normalt ställs på offentlig verksamhet, exempelvis rättssäkerhet, representativitet, deltagandevärden och offentlighetsvärden. Hur avvägningen mellan dessa värden och produktivitet/effektivitet skall ske, kan ingen utvärdering med vetenskapliga anspråk ge besked om. Avvägningen kan endast ske med hjälp av offentlig debatt, opinionsbildning, kompromisser, majoritetsbeslut och maktspråk, dvs. politik.

Slutord om val av utvärderingsmodell

Den genomgång som här har gjorts har skett utifrån ett smalt perspektiv på utvärdering. Vi har valt att organisera framställning efter värdekriterier som används vid utvärdering. Därmed har vi i stort sett förbigått en mängd andra problem, som t.ex. om utvärdering skall ske som akademisk forskning av utbildade vetenskapsidkare eller om den skall genomföras av berörda professionella, intressenter eller brukare. Vi har i stort sett bara snuddat vid om utvärdering skall ske som experiment eller om den skall ske i form av fallstudier. Vi har inte heller systematiskt behandlat om utvärdering skall vara demokratisk i bemärkelsen participativ och deliberativ och ske i dialog med de berörda.

Genomgången har visat att den totala uppslutning, som i 1960-talets nordamerikanska och europeiska utvärderingsvåg rådde bakom måluppfyllelsemodellen och därmed den representativa demokratins princip, har brutits upp och ersatts av en situation där många modeller konkurrerar. Till en del beror detta på att läran om offentlig politik utvecklats mot större sofistikation, t.ex. vad gäller bieffekter av offentlig planering, vilket lett till att måluppfyllelsemodellenövergåttibieffektsmodellen. Vidare har under 1990-talet den offentliga sektorns kostnader kommit i förgrunden, vilket resulterat i en renässans för ekonomiska utvärderingsmodeller. Andra intressenter än folkets valda representanter har med ökad styrka börjat resa krav på deltagande i och inflytande över förvaltningsprocesser, vilket lett till att brukarmodeller och intressentmodeller kommit till användning. Den ökade användningen av professionella modeller är lite speciellt eftersom krav på detta kommer uppifrån och inte från professionerna själva. Allt i allt har utvärdering utvecklats från uniformitet till pluralism.

Det finns en stark tendens i nutida utvärderingslitteratur att helt avfärda måluppfyllelsemodellen. Från en utgångspunkt är denna totala nedvärdering oberättigad. Ur den representativa demokratins synvinkel är nämligen måluppfyllelsemodellen helt central, eftersom den grundas på idén om den parlamentariska styrkedjan. Mål satta på hög politisk nivå (eller på förvaltningsnivå men på delegation från den politiska nivån) är inte vilka mål som helst. De har framkommit genom konstitutionella procedurer, som gjort att de institutionaliserats som representativa församlingars mål. Medborgare och valda ombud har legitim anledning ta reda på huruvida dylika mål har uppnåtts på fältet. I annat fall kan det representativa demokratiska styrelseskicket inte fungera.

Det representativdemokratiska argumentet till försvar för måluppfyllelsemodellen kan emellertid inte göra något åt det faktum, att modellen råkar in i svårigheter särskilt med målkataloger. Dess största analysmässiga nackdel är dock att den ej förmår beakta sidoeffekter. Av denna anledning föredrar jag bieffektsutvärdering framför måluppfyllelseutvärdering. Bieffektsutvärdering kan ses som måluppfyllelseutvärdering utbyggd med sidoblickar på sidokonsekvenser. Bieffektsmodellen hämtar också sin styrka ur den representativa demokratins princip.

Ekonomiska modeller har från slutet av 1980-talet vunnit terräng, bl.a. därför att de offentliga finanserna försämrats. Även dessa modeller hänger samman med den parlamentariska styrkedjan och den representativa demokratisynen. Medborgarna och deras valda representanter behöver inte bara kunskap om sakinriktade resultat av offentliga åtgärder utan även om kostnader för att uppnå dessa resultat. Häri ligger också de ekonomiska modellernas styrka. En annan styrka är att de förmår reducera en verksamhets värde till enkla, lätt begripliga tal. Vi bör emellertid hålla i minnet att ekonomiska modeller i likhet med andra modeller ger endast ett partiellt perspektiv. Faran med ekonomiska modeller är att beslutsfattare fascineras av deras matematiska precision och felaktigt tror att de ger allomfattande, slutgiltiga svar.

Om måluppfyllelsemodellen, bieffektsmodellen och ekonomiska modeller hämtar sin grundläggande legitimitet ur en representativ demokratisyn synes intressentmodeller och brukarmodeller bygga på en deliberativ och participativ demokratisyn. Medborgarna skall inte delta i offentliga sektorns angelägenheter enbart via de allmänna valen. De bör delta även i sin kapacitet av slutmottagare eller annan intressent. Intressentmodeller och brukarmodeller skall kännetecknas av öppen debatt genom överbevisning och saklighet. De tänkes upprätta arenor för överläggningar och informerade samtal där folkvettet är direkt involverat. På det sättet kan de komplettera det representativa demokratiska systemet. Brukar- och intressentmodeller kan komma i konflikt med de modeller som drivs av föreställningar om den representativa demokratin. I denna framställning har de senare modellerna getts ett principiellt försteg. Med hänvisning till den parlamentariska styrkedjans primat för den offentliga sektorn kan brukar- och intressentmodeller endast komplettera men inte ersätta ansatser, som tar sin utgångspunkt i programmål och offentliga kostnader.

Styrkan i kollegebedömning och självvärdering är att de kan fånga in och bedöma kvaliteter. Detta är oundgängligt inom områden som kan kännetecknas av ytterst specialiserad kunskap. Paradigmfallet är naturligtvis akademisk forskning. Även dessa professionella modeller måste avvägas mot representantmodellerna.

Faran med alla utvärderingsmodeller är att de används alltför okritiskt och att beslutsfattare felaktigt tror att en enda modell kan lämna slutliga svar på centrala frågor. Därför bör det hållas i minnet att varje modell endast ger perspektiv och partiella svar. Av denna anledning rekommenderas att flera modeller får komma till användning. Pluralism är den hållning, som för närvarande ter sig mest rimlig.

Litteratur

Allardt, Erik & Sverre Lysgaard & Aage Bøttger Sørensen. (1987), Sociologin i Sverige: Vetenskap miljö och organisation, Uppsala: Swedish Science Press.

Danielsen, Rolf. (1988) Historia i belysning: Sex perspektiv på svensk historisks forskning, Uppsala: Swedish Science Press.

Engwall, Lars. Red. (1992) Economics in Sweden: An Evaluation of Swedish Research in Economics. London: Routledge.

Karlsson, Ove. (1995) Att utvärdera – mot vad? Om kriterieproblemet vid intressentutvärdering, Stockholm: HLS Förlag.

Vedung, Evert. (1997) Public Policy and Program Evaluation, New Brunswick, N-J.: Transaction Publishers.

Vedung, Evert. (1998) Utvärdering i politik och förvaltning, Lund: Studentlitteratur.

Öhman, Arne & Bo Öhngren. Red. (1991) Two Faces of Swedish Psychology: 1. Frontiers in Perception and Cognition: An Evaluation of Swedish Research in Cognitive Psychology, Uppsala: Swedish Science Press.

Summary
Evaluation models

The present classification, based on value criteria used in evaluation, has demonstrated that the total agreement which once existed about the appropriateness of the goal-attainment model has been replaced by a situation where several models compete. Evaluation has evolved from uniformity to pluralism. Yet, every model provides partial perspectives only. Combinations of several models are recommended.

There is a strong tendency in contemporary evaluation literature to recommend stakeholder and client-oriented evaluation and debase particularly the goal-attainment model. In many areas this is entirely justified. On one account, I take exception to this tendency. From a representative democracy point of view the goal-attainment model and particularly the side-effects model are important, since they are based on the conception of the parliamentary chain of influence. Policy goals, set by parliaments and governments, and program and project goals set by lower-level administrator on delegation from parliaments and governments are not just any goals. Established through a constitutionally determined procedure, they are institutionalized as the collective goals of the state. Principals of various kinds have legitimate reasons to ascertain whether goals set by their agents have in fact materialized in the field. Otherwise, they cannot function as principals in the representative system of government.

The major drawback with the goal-achievement model is its lack of focus on side-effects. For this reason, I prefer side-effects evaluation to goal-achievement evaluation.

Client-centered evaluation has a role to play, particularly as far as government services are concerned. The strength of peer review and other professional models lies in their capacity to capture and judge qualities. Their paradigm area of application is academic research and higher education. The stakeholder model provides the broadest view possible of government interventions, promises to take all involved into consideration, and may give legitimacy to programs.

Economic models will stay with us forever in public policymaking. It must be kept in mind, however, that like other designs they provide partial perspectives only. The danger with economic models is that decision-makers are fascinated by their mathematical precision and wrongly believe that they provide comprehensive, final answers. While the goal-attainment model, the side-effects model and economic models get their fundamental justification from the theory of representative democracy, stakeholder models and client-oriented models relate to deliberative and participative democracy. One of their rationales is to establish deliberative arenas to supplement representative democracy with some citizen participation and deliberation.

Directions in Qualitative Evaluation

ian shaw

It is regarded by many as not far short of bad taste to advance passionate claims based on the superiority of this or that methodology. The argument of most mainstream evaluation theorists is for a 'horses for courses' approach that aims to identify the strengths of different methods and discourage evaluators from over-claiming the relevance and application of any one approach to evaluation. I use this article to develop a few outline arguments in support of turning on their heads some conventional arguments about methodological choices for evaluation. I touch on four areas where qualitative methodology enables evaluators to re-cast central aspects of evaluation practice, viz causal understanding, methodological choice, the evaluation of professional practice, and the uses of evaluation.

Introduction

It is regarded by many as not far short of bad taste to advance passionate claims based on the superiority of this or that methodology. The argument of most mainstream evaluation theorists is for a 'horses for courses' approach that aims to identify the strengths of different methods and discourage evaluators from over-claiming the relevance and application of any one approach to evaluation. For example, if the evaluator is operating with an evaluation-as-accountability perspective – eg measuring results or efficiency – then the randomised, controlled, clinical trial (RCT) provides the 'gold standard, the Rolls Royce of evaluation approaches' (Chelimsky, 1997: 101). Qualitative strategies, for example case study designs, should not be used to tackle questions and problems which are the province of quantitatively-oriented methods. However, if the perspective is

Ian Shaw is Reader in Social Work at Cardiff University, School of Social Sciences, Cardiff (Wales). He is the author of Qualitative Evaluation (1999 Sage) and co-editor of the journal Qualitative Social Work (Sage).

one of evaluation for greater knowledge and understanding of a given policy or programme, then qualitative approaches may be the method of choice.

I have some sympathy for this position. Even the best arguments for pulling the coach of evaluation methods behind the horses of some favoured paradigm are in the end self-defeating, and I do not intend to use scarce space rehearsing the reasons for this conclusion[1]. In that sense I agree with the sociologist David Silverman when he says 'there are no principled grounds to be either qualitative or quantitative in approach. It all depends on what you are trying to do' (Silverman, 1997b:14). Demonizing positions with which we disagree is a fool's errand. It entraps us in sentimentality and superstitious practices (Shaw, 1999). I like the story of Philip of Macedon who apparently employed a man with a stick, atop of which was a pig's bladder. The sole function this man fulfilled was to exercise the freedom to wake Philip at any time of the night, and beat him about the head with the bladder, as a reminder that he was mortal. We forget our methodological mortality at our peril. There are, as Lather remarks, 'no innocent positions' (Lather, 1991: 85), and we need to 'protect our work from our own passions' (Lather, 1986: 77).

And yet the horses-for-courses approach of the evaluation mainstream imposes a premature closure on questions that are far from closed, and thus reinforces the status quo, some issues remaining forever out of court for methodological challenge from other quarters. In practice this operates as a line of defence for traditional evaluation methodologies. Evaluating outcomes, understanding causal processes, providing confident generalizations, and so on, remain the territory of just some evaluators, and 'merely serve to sanctify one perspective at the expense of another' (Chelimsky, 1997: 108). Boundaries thus become fixed at some arbitrary point in time. Why should we agree to fix a division of labour that is the simple happenstance of the century's turn? This is likely to perpetuate unhelpful hierarchies. If Chelimsky is right when she concludes that RCTs are the Rolls Royce then maybe case studies are a Trabant or at best a rather quaintly English Morris Minor.

I want to use this space to develop a few picture-board outline arguments in support of turning on their heads some conventional arguments about methodological choices for evaluation. I will touch on four themes, viz causal understanding, methodological choice, the evaluation of professional practice, and the uses of evaluation.

My general position is that evaluation is characterized by a cluster of evaluative purposes. These enable plausible and productive responses to questions such as methodological choice, evaluation theorising, evaluation ethics, and advocacy evaluation. Evaluation is best understood as entailing the conduct of evaluative research rather than a discrete set of evaluation axioms or methodology separate from the wider research enterprise. In this regard, evaluation theory and methodology owe almost

1 have attempted this in some detail in Shaw (1999) Qualitative Evaluation London: Sage. Eg Chapter 3.

as much to work undertaken by writers and researchers who would not regard themselves as evaluation theorists, as it does to confessedly evaluation theorists.

The boundary fence between research and evaluation is not the only one that needs dismantling. A relevant and rigorous evaluation requires the development of inter-professional evaluation theorising and strategies, such that education, health, criminal justice, law, human services and so on, are mutually attentive.

Evaluation is more – much more – than programme evaluation. Qualitative evaluation promises distinct but coherent perspectives on policy, programme and practice evaluation. Qualitative evaluation offers credible partial solutions to problems of causal analysis and outcome evaluation. It also enables us to avoid an unduly instrumental and rational approach to the uses of evaluation[2].

I don't want to be misunderstood as to the scope of the claims I am trying to support. It would be silly and self-defeating, for example, to leave the impression that qualitative evaluation should be the order of the day for all evaluative purposes. Rather, I hope to unsettle unquestioning faith in the evaluative benefits of some forms of mainly quantitative evaluation – those in particular that place undue confidence in the possibilities of controlling and precisely measuring independent and dependent variables.

I am not convinced, either, that all forms and traditions of qualitative methodology lend themselves equally or even directly to evaluative purpose. This would simply replace one variety of uniformitarianism with another, through a tendency to treat qualitative methodology in an unduly homogenous way. As a corrective to this, I believe there is a need to develop the case for a dialectical mix of methods within qualitative research. This will need to proceed through the development of a set of critical features of knowledge for different qualitative methodologies. A helpful starting point for this is the paper by McKeganey and colleagues, in which they discuss the benefits and limitations of interviewing and observation methods as part of a study of professional decision-making when people may be offered a place in a home for the elderly (McKeganey et al., 1988). This initial analysis needs to be extended to the full range of qualitative strategies, and tied to the critical features of their associated knowledge claims (c.f. Greene and Caracelli, 1997: 12-13). I do not attempt this task here, but it is with such considerations in mind that I have deliberately avoided giving overall definitions of what constitute qualitative or quantitative studies, but have restricted myself to a number of illustrative examples, mainly from the fields of interpretive sociology, ethnography and case studies.

2 This paper has been written primarily to an evaluation audience. I have corresponding misgivings about the ways in which a majority of mainstream qualitative research academics are silent on the evaluation relevance of their field. For a recent example of the dog that does not bark, see Atkinson, Delamont, Coffey, Lofland and Lofland (2000).

Understanding causes

The conventional division of labour is that qualitative inquiry is useful for generating hypotheses/questions, and describing processes, while quantitative and more statistical designs are needed to analyze outcomes and verify hypotheses. Miles and Huberman summarize this view – from which they vigorously dissent - as:

> qualitative studies are only good for exploratory forays, for developing hypotheses - and … strong explanations, including causal attributions, can be derived only through quantitative studies.
> (Miles and Huberman, 1994: 147)

This is too limiting. Neither quantitative nor qualitative evaluation can solve questions of cause and effect in a straightforward way. Led by recent work on realist evaluation, changes have taken place in thinking regarding the nature of cause, and the corresponding models of causal hypotheses which flow from that thinking. The central idea is that there are underlying causal mechanisms which cannot be understood by surface workings and measurement. Hence, 'events themselves are not the ultimate focus of scientific analysis...Reality consists not only of what we can see but also of the underlying causal entities that are not always discernible' (House, 1991: 4). The underlying reality produces actual events, of which we have empirical experiences and sense impressions.

This is often described as a generative concept of causality.

> When we explain an outcome generatively we are not coming up with variables or correlates that associate with one another; rather we are trying to explain how the association itself comes about. The generative mechanisms thus actually constitute the outcome. (Pawson and Tilley, 1997: 408) [italics in original]

The conventional concept of causation as regularities and associations is dismissed in favour of causal entities which have 'tendencies interacting with other tendencies in such a way that an observable event may or may not be produced' (House, 1991: 5). House quotes Manicas and Secord who say that, 'For the standard view of science, the world is a determined concatenation of contingent events; for the realist, it is a contingent concatenation of real structures. And this difference is monumental'. Hence, instead of merely documenting the sequence and association of events, the realist seeks to explain events.

While this view of cause does not necessarily require a qualitative methodology, it does clearly lend itself to such methods.

> Qualitative studies are not designed to provide definitive answers to causal questions...(but) it can still be an appropriately qualified pursuit. (Lofland and Lofland, 1995: 136, 138)

Miles and Huberman are even less reserved. They describe the conventional view as 'mistaken' (Miles and Huberman, 1994: 147), and insist that qualitative evaluation research is well equipped to,

1. Identify causal mechanisms.
2. Deal with complex local networks.
3. Sort out the temporal dimension of events.
4. Cycle back and forth between different levels of variables and processes.
5. Provide a way of testing and deepening single case explanations through analytic induction.

Causal accounts will be local and 'now-oriented' (Lofland and Lofland, 1995: 141). Miles and Huberman develop analytic methods which address causal attribution in both single and multiple case explanations. For example, they advocate the use of field research to map the 'local causal networks' which informants carry in their heads and to make connections with the evaluator's own emerging causal map of the setting. Such maps start from 'causal fragments' which lead on to linked building of logical chains of evidence. Such causal networks

are not probabilistic, but specific and determinate, grounded in understanding of events over time in the concrete local context - and tied to a good conceptualisation of each variable. (Miles and Huberman, 1994: 159)[3]

Much of this reasoning was anticipated by Cronbach's arguments regarding causal models. Rejecting the idea of causation as events that can be predicted with a high degree of probability, Cronbach developed twin arguments. First, he argued that causes are contingent on local interactions of clusters of events. More than one cluster may be sufficient, but no one cluster is necessary. Second, he accepted that there are usually missing events or conditions that affect the outcome of a given programme, but about which we know little. He was the first evaluation theorist to produce a plausible explanation of contextual factors in evaluation. Hence he concludes that 'after the experimenter with his artificial constraint leaves the scene, the operating programme is sure to be adapted to local conditions' (Cronbach et al., 1980: 217). Furthermore, 'a programme evaluation is so dependent on its context that replication is only a figure of speech' (p. 222).

Qualitative evaluation cannot resolve the problems of causal conclusions any more than quantitative evaluation, but it can assess causality 'as it actually plays out in a particular setting' (Miles and Huberman, 1994: 10).

Lofland and Lofland make the important observation that causal answers are by and large based on passivist conceptions of human nature. Qualitative inquiry has often steered away from causal accounts, not because the methodology is weak in

3 An interesting connection can be drawn between Miles and Huberman's assessment and current work, mainly in American evaluation writing, on logic models. A qualitative methodology stance on logic models suggests the potential value of inductive 'logic models'. I see the existing USA approach as a way of combining a consensual stakeholder input with outcome analysis. The drawback of this approach is that it under-emphasizes the likely persistence of variant informal logic models. Inductive, informal logic models offer a use of 'logic' in a sense not far from ideas of frames of meaning as used in Anthony Giddens and others.

that area but because of a commitment to an activist conception of human nature. The Loflands argue that an activist conception will lead to a focus on questions that address both structures and strategies. This will involve 'deciphering and depicting exactly what sort of situation the participants are facing' (Lofland and Lofland, 1995: 146), and understanding the 'incessantly fabricated' strategies people construct to deal with the situation.

Take for example, Silverman's work on HIV counselling. He is right to conclude that 'it is usually unnecessary to allow our research topics to be defined in terms of…the »causes« of »bad« counselling or the »consequences« of »bad« counselling' (Silverman, 1997a: 34), insofar as such topics reflect the conceptions of social problems as recognized by professional or community groups. Nonetheless, this does not require the abandonment of causal inquiry in qualitative evaluation. Inquiry into the ways in which professionals incessantly fabricate service forms and structures does promise a better way to understand causes.

By way of illustration, Shaw and colleagues describe a case study evaluation

Figure 1.
Stakeholder models of a Rural Activity centre

	Training for Work	Personal/social growth	Education for life
Aims	Credible work skills for independent/ sheltered work	Personal and social growth	Alternative occupation to enhance the quality of life
Target group	Demonstrable ability to benefit; younger	Wide range of age and ability	Wide ability range; younger
Programme	Time limited stay; skill learning; assessment and review; contracts; move-one facility; integration into work	Open stay period; small project; small-group activities; counselling; liaison with carers and social work agencies	Loosely held time limits; the best learning context; interest-led contracts; community based activities and outside links; craftwork and homemaking skills
Staffing	Education and special needs employment skills; plus volunteers	Social and group work qualifications; plus expert consultants	Education and social work qualifications; plus volunteers
Outcome	Regular throughput; work placements; normalization of work patterns; skill learning	No clear distinction between programme and outcome	Wide range of social skills; integration into community networks; change of attitudes on the part of outside community members

of a rural activity centre for people with learning disabilities. They observed and interviewed project participants, parents, carers, management group members, key workers and other professionals. Project records were analyzed.

When describing and explaining the workings of the centre, the people who were interviewed appeared to draw on one or more of three different models of the scheme. These were a 'training for work' model, a 'personal and social growth' model, and an 'education for life' model. These operated in part as causal maps which entailed an array of model-specific positions on the aims of the project, optimal target groups, desirable programme patterns, staffing requirements, future development strategies, and likely or desirable project outcomes[4].

Choosing Methods

Qualitative evaluation methodology is not only well equipped to address the local outworking of cause and effect, but also enriches the choice of methods relevant to evaluative purposes. The inter-relationship of qualitative and quantitative methods is not only, nor even primarily, about choice of methods. It is about the questions in Figure 2 and is also inextricably relevant to issues of the politics and purposes of social work research, values, participatory forms of research, interdisciplinary research, and the uses of research.

Figure 2.
Qualitative and Quantitative Methodology

Single cases or comparison.
Cause and meaning.
Context as against distance.
Homogeneity and heterogeneity.
Validity and the criteria of quality in social work research.
The relationship of researcher and researched.
Measurement.

Qualitative inquiry may shed light on programme outcomes in ways that are less susceptible to quantitative methodology. Miller, for example, discusses ways that institutional texts constructed to explain past decisions inevitably gloss over the openness and complexity of the decision-making process (Miller, 1997). He gives the mundane example of evaluation research on a bowel-training programme in a nursing home. The evaluation consisted of counting when and how patients had bowel movements. The programme was judged to have a successful outcome if patients used a toilet or bedpan and ineffective for those who continued soiling beds. One patient had soiled her bed. However, observation methods enabled the researcher to view a nursing aide contesting the definition of this as 'failure' on the grounds that the patient knew what she was doing and had

4 Rethinking approaches to understanding causal processes leads naturally to rethinking the ways in which it is possible to generalize from one program and its evaluation to another program. Qualitative researchers and evaluators have developed partial answers to this question through ideas about vicarious experience (Stake and Trumbull. 1982), transfer (Eisner, 1991), 'thick description' (Geertz, 1973) and analytic generalisation.

soiled her bed as a protest act against staff favouring another patient. This illustrates how observing the context of text construction illuminates mundane, everyday life. This would not have found a way into the formal outcome record. Text production in institutions is 'micro-politically organized', and this includes textual outcome records.

A further illustration of the relevance of qualitative methodology for outcomes evaluation can be traced through the surprising impact of Denzin's interpretive interactionism (Denzin, 1989 and 2002; Mohr, 1997). Mohr, for example, extends Denzin's argument to the evaluation of clinical outcomes in health research. She argues that the method leads us to inspect the relationships between personal difficulties, experiences, policies, interventions, and institutions. 'Interpretive interactionism permits intensive scrutiny of the ramifications and outcomes of various interventions' (1997: 284). It can:

1. Sort out different ways problems are defined.
2. Show how patients experience care. What it is about interventions they find helpful or not, and in what circumstances.
3. Identify 'secondary causes' eg contexts, culture, and the meanings patients bring.

'Strategic points for intervention can be identified by contrasting and comparing patients' thick descriptions, and these can be used to change, to improve, or to negotiate and renegotiate interventions' (p.284). It is valuable when 'an outcome may not be readily apparent, and…the intervention is something that only the patient and not the professionals can define' (p.285).

Constructive, if cautious, dialogue regarding the relative merits and characteristics of quantitative and qualitative methodologies has emerged more recently. The social work literature provides a useful example. From the quantitative side of the case, Reid in the USA and Sinclair in Britain have developed mediating positions. Reid seeks to 'redefine the nature of the mainstream so that qualitative methodology is a part of it not apart from it'. He regards quantitative research as strong when dealing with linkages, control, precision, and larger data sets, while qualitative research is able to depict system workings, contextual factors, and elusive phenomena, and provide thorough description. 'Neither method is superior to the other, but each provides the researcher with different tools of inquiry' that can be set against a single set of standards (Reid, 1994: 477).

Sinclair adds to Reid's conclusion, in his discussion of randomized control trails (RCTs), when he says that qualitative methods are in many ways 'more adapted to the complexity of the practitioner's world than the blockbuster RCT'.

Qualitative research draws attention to features of a situation that others may have missed but which once seen have major implications for practice. It counteracts a tendency to treat the powerless as creatures with something less than normal human feelings. It contributes to an ethically defensible selection of outcome measures. And, in combination with simple statistical description, it can lead to an informed and incisive

evaluation of programmes in social services. (Sinclair, 2000: 8)

He turns common assumptions on their head when he concludes that,

Quantitative social work research does face peculiarly acute difficulties arising from the intangible nature of its variables, the fluid, probabilistic way in which these variables are connected, and the degree to which outcome criteria are subject to dispute. (pp. 9-10)

Evaluating professional practice

Qualitative methodology also provides a strong purchase on the evaluation of direct service delivery. The main point I wish to make is that thinking and practice in the evaluation field have been too much influenced by ideas of evaluation as being equivalent to programme evaluation (indeed, often as equivalent to programme outcome evaluation). I believe we need to distinguish more strongly in our thinking between evaluation of policies, programmes, projects within programmes, and direct practice and service delivery. I am especially interested in the last of these. Rather than see good practice as being subject to evaluation – whether internal or external evaluation is immaterial – I believe good practice should in and of itself entail evaluation for and with service users. Influences on my position include:

- Reflective practice.
- Qualitative methodology.
- Empirical evidence on evaluation as a dimension of professional practice.
- Advocacy evaluation and user-led research.
- Action research.

The model I have tried to develop, and have tested out in some measure with practitioners, is premised on a composite image of good practice as requiring evaluative evidence, evaluative learning, and evaluative justice. It draws, therefore on what in the UK is increasingly being called 'knowledge-based practice'; on the learning organization and reflective learning literature; and on advocacy models of evaluation. The main practical approach I have developed has been to seek to 'translate' and 'colonize' methods, especially, but not exclusively, qualitative ones (eg Shaw, 1996, 1997). Examples include life histories, simulations, focus groups, narrative methods, cultural reviews, inductive local logic models, and peer interviewing.

Empirical work has shed some light on how practitioners seek to make sense of and resolve evaluative issues in their day-to-day work. Elks and Kirkhart urge that 'an alternative research model is needed, one that is exploratory rather than confirmatory, building a model of evaluation from the practitioner's own accounts rather than superimposing an ideal model and testing for conformity' (Elks and Kirkhart, 1993: 555). They interviewed seventeen social workers asking them how they evaluated what they were doing, and how they knew whether they were doing a good or bad job. Practitioners acknowledged difficulty in knowing if they were

effective. They also perceived an incompatibility between the roles of evaluator and practitioner. The researchers suggest that practitioners hold an implicit model of practice evaluation which they describe as a 'pragmatic-professional model'. This included a reliance on intuition and experience, an internalised notion of an ideal practitioner, a dependence on feedback from colleagues, friends and family, and a model of an ideal client which always included growth and change.

Humphrey and Pease conducted a corresponding study in which they interviewed British Probation Officers regarding their perceptions of probation effectiveness. Probation officers tended to 'de-couple' the process of supervision from the out-turns of probation. Thus one person said, clients 'frequently get into more trouble but I don't think that in any way is a reflection on whether or not I have been effective'. Indeed, there was widespread belief that an element of luck operated in being effective. One might do 'brilliant' work but if the circumstances are against you they will still re-offend. Thus, 'if luck is seen to determine outcome, probation supervision becomes merely a matter of keeping an offender in the community for luck to strike' (Humphrey and Pease, 1992: 40).

Subsequent work by Shaw and Shaw suggests that social workers appear to have two contrasting models of evaluation in their heads - a formal 'evaluation proper' and self-evaluation. Formal evaluation is experienced as largely alien to the realities of social work and in almost complete contrast to social workers' evaluative 'maps' of their actual day-to-day evaluating (Shaw and Shaw, 1997b). Evaluation strategies were constructed from a 'game plan', the success of which was viewed - as in Humphrey and Pease' research - as partly contingent on the untoward operation of 'sheer luck'. Social workers judged their practice according to whether their work produced emotional rewards; the case was 'moving'; intervention won steady, incremental change; practice was accomplished without inflicting harm through the operations of the welfare system, and confirming evidence was available from fellow professionals. These practitioners were preoccupied with causes and reasons for outcomes of their work, held strongly worked views about the complexity and ambiguity of social work evidence, and were aware of the constant interplay of knowing and feeling in practice (Shaw and Shaw, 1997a). The significance of emotions echoes Erikson's remark of the clinician that

> The evidence is not »all in« if he does not succeed in using his own emotional responses during a clinical encounter as an evidential source and as a guide to action. (Erikson, 1959: 93)

The emphasis of these previous paragraphs has been on evaluation as endemic and taken-for-granted within professional practice, and for the considerable gains derived from rendering it visible as a step to embedding it at the core of good practice. This embedding will be further promoted if we avoid over-simple distinctions between insider and outsider evaluation. A drawback of much practitioner research

stems from a tendency to regard 'practice' as distinct from theory, and hence to regard 'being theoretical' as something that happens in the mind and the 'practical' as having a derivative, 'applied' relationship to these guiding ideas. Conversely to this, good practitioner evaluation ought to give attention to exploring different kinds of tacit knowledge. The significance of personal contact and practical knowledge sharing between practitioners will be brought out, and sources of trust and mistrust between social workers made clear (Collins, 2000). This is a big agenda, and one that social work and other occupations have only begun to tackle. One consequence is that we cannot hang on to a narrow distinction between practitioner research as being 'insider' evaluation and 'academic' research as being 'outsider' evaluation. In a recent analysis of qualitative social work research, several of the contributors reflect on these issues. Hall and White, for example, record how they held both insider and outsider roles in relation to their research participants. Hall 'arrived' as an outsider but became in different ways a partial insider (Hall, 2001). White started as an insider, yet found herself undergoing a fruitful, if potentially hazardous, process of de-familiarisation through which she became in some degree a marginal 'inside »out« member (White, 2001). In the same volume, Scourfield focuses his reflections on the research and practice relationship through his consideration of what it was like to interview expert professional social work interviewers (Scourfield, 2001).

In what ways do we expect evaluation to be useful?

Finally, a focus on qualitative methodology leads fairly directly to the wider question of how evaluation might be useful for policies, programmes, projects and professional practice. At the broadest level, evaluation would be judged useful if it demonstrably contributed to one or more of the following.

- Better policies, services and practice.
- Strengthened the moral purpose of professional practice.
- Promoted methodological rigour, scope, depth and innovation.
- Strengthened the sense of a profession's intellectual nature and location.

We tend to make simple distinctions between research that has a direct, applied purpose and research that is basic, and hence where direct use questions are less relevant. This does not work (just as the related sharp distinction between 'research' and 'evaluation' does not work either).

Conventional quantitative research on outcomes is linked to a confidence in the instrumental utility of research. The problem with this is that it does not square with evidence on how evaluation is actually used, and it misunderstands the nature of the policy making process. It is based on a rationalistic model. The rationalist model of policy making sees it as a series of discrete events, where each issue to be decided is clearly defined, and decisions are taken by a specific set of actors who choose between well-defined alternatives, after weighing

the evidence about the likely outcome of each (Finch, 1986: 149-150).

An important early figure for this question is Carol Weiss, whose work in the 1970s explored how political considerations intrude on evaluation. She addressed three related issues. First, she delineated the political context in which evaluation is located. Although she has been primarily concerned with evaluation and policy research at the federal level, her empirical work with policy and programme staff resonates throughout evaluation theory and practice. This underlines the importance of being clear about the audiences for any inquiry.

Second, she exposed the limitations of conventional instrumental views of the political use of information, through her conceptualisation of use as enlightenment. With her colleagues she interviewed 155 senior officials in federal, state and local mental health agencies. Officials and staff used research to provide information about service needs, evidence about what works, and to keep up with the field. However, it was also used as a ritualistic overlay, to legitimize positions, and to provide personal assurance that the position held was the correct one. At a broader conceptual level, it helped officials to make sense of the world. For all these purposes, 'It was one source among many, and not usually powerful enough to drive the decision process' (Weiss, 1980: 390). As for direct utilization of research, 'Instrumental use seems in fact to be rare, particularly when the issues are complex, the consequences are uncertain, and a multitude of actors are engaged in the decision-making process. (p. 397).

Research use was also reflected in officials' views of the decision-making process. Decisions were perceived to be fragmented both vertically and horizontally within organizations, and to be the result of a series of gradual and amorphous steps. Therefore, 'a salient reason why they do not report the use of research for specific decisions is that many of them do not believe that they make decisions' (p. 398). Hence the title of one of her papers - 'Knowledge creep and decision accretion'. This provided the basis for her conclusion that enlightenment rather than instrumental action represents the characteristic route for research use.

The enlightenment model 'offers far more space to qualitative research, through its emphasis on understanding and conceptualisation, rather than upon providing objective facts' (Finch, 1986: 154).

Third, Weiss imbued models of use with a realistic view of the public interest. More than anything she has struggled towards a realistic theory of use. Others subsequently have developed such realistic views. For example, Chelimsky suggests that evaluation may have a deterrence function. 'In other words, the mere presence of the function, and the likelihood of a persuasive evaluation, can prevent or stop a host of undesirable government practices' (Chelimsky, 1997: 105).

Several cautious remarks are in order. First, we should not become over-pre-occupied with models of research use. Chelimsky believes 'it is often the case that….evaluations are undertaken without any hope of use'. Expected non-use is characteristic of some of the best evaluations, including 'those that question widespread popular

beliefs in a time of ideology, or threaten powerful, entrenched interests at any time' (p. 105). Thus, 'there are some very good reasons why evaluations may be expert, and also unused' (p. 105). Chelimsky's comments are both sane and plausible.

To justify all evaluations by any single kind of use is a constraining rather than an enabling idea because it pushes evaluators towards excessive preoccupation with the acceptability of their findings to users, and risks turning evaluations into banal reiterations of the status quo. (Chelimsky, 1997: 106)

Second, the enlightenment model should not be adopted as universally appropriate. For example, practitioner research is likely to proceed on a more immediate instrumental view (see, for example, Figure 3).

Third, the adoption of enlightenment assumptions about research use can easily translate into a defensive posture, arising from the fact that they can readily be used to support an incrementalist approach to social change. The step from an empirical recognition that policy and practice change often proceed through incremental enlightenment to a tacit assumption that this is how social change ought to proceed may be logically insupportable, but it is deceptively easy.

Fourth, in counterpoint to the previous point, we should avoid being unduly sanguine about the ability of research to change practice. Hammersley has criticized some professionally-driven research approaches on the grounds that they are based on too narrow a concept of research relevance and an overly optimistic faith in the ability of research to influence policy and practice. He suggests two grounds for concluding 'there are good reasons to believe that research cannot routinely solve teachers' problems' (Hammersley, 1993: 430). 'There is no scientific method that guarantees results' (p. 430) and teacher circumstances are diverse and unlikely to be amenable to action in any routine sense. Rather, 'sound practice cannot amount to the straightforward application of theo-

Figure 3.
Evaluation models and information use

Model of use	Evaluation focus	Evaluationbase	Discipline links	Time span for action
Enlightenment	Policy evaluation Programme development Advocacy evaluation	Higher education	Strongerdiscipline links Commitment to theorising	Longer term
Instrumental	Projectevaluation; programme feasibility studies; practitioner evaluation	Agency sponsored; 'Insider'evaluation; self evaluation	Limited social sciencetheorising	Immediate 'applications'

retical knowledge, but is an activity that necessarily involves judgement and draws on experience as much as on…scientific knowledge' (p. 430).

Fifth, there is often an important connection between research use and ethics. Consider the ethical and political dilemmas about how research material is used. These issues are sometimes sharper in qualitative research. This arises partly from the greater closeness and consequent trust that may develop between evaluator and participant. In quantitative research the greater distancing may make these issues less agonising. The risk of betrayal is also increased because of the typical use of smaller samples, the consequent difficulties of protecting the confidentiality of individuals, and the emphasis on the details of how people live their lives. Finch describes from her playgroups research her 'sense that I could potentially betray my informants as a group, not as individuals' (Finch, 1986: 207). 'Where qualitative research is targeted upon social policy issues, there is the special dilemma that findings could be used to worsen the situation of the target population in some way' (Finch, 1985: 117).

But none of these caveats reduces the overall value of careful elaboration of what we mean when we speak of research being useful. For instance, everyday discussions of evidence-based practice frequently proceed on a misconception at this point. Practice is not and cannot be 'based' on evidence in the straightforward and unproblematic way envisaged by many of its advocates.

I have argued in this paper that the field of evaluation has much to gain by holding back from a premature consensus approach to evaluation methodology. I have touched on four areas where qualitative methodology enables evaluators to re-cast central aspects of evaluation practice, viz causal understanding, methodological choice, the evaluation of professional practice, and the uses of evaluation. I could as easily have developed the same argument for other domains – quality standards in evaluation, evaluation ethics and governance, evaluation synthesis, the relation of evaluation to mainstream social science, interdisciplinarity, and so on. In all these areas I am convinced by Paul Feyerabend's provocative aside that 'it is not the puzzle-solving activity that is responsible for the growth of knowledge, but the active interplay of various tenaciously held views' (quoted by Trend, 1979:84).

References

Atkinson, Paul, Delamont, Sara, Coffey, Amanda, Lofland, John and Lofland, Lyn (2000) Handbook of Ethnography London: Sage Publications.

Chelimsky, Elinor (1997) 'Thoughts for a new Evaluation Society', in Evaluation, 3 (1): 97-118.

Collins, Harry (2000) Tacit Knowledge, Trust and the Q of Sapphire Cardiff University School

of Social Sciences. Working Paper. <http://www.cf.ac.uk/socsi/>

Cronbach, Lee, Ambron, S., Dornbusch, S., Hess, R., Hornik, R., Phillips, D., Walker, D., and Weiner, S. (1980) Toward Reform of Program Evaluation. San Francisco: Jossey-Bass.

Denzin, Norman (1989) Interpretive Interactionism, Englewood Cliffs, NJ: Prentice Hall.

Denzin, Norman (2002) 'Social work in the seventh moment' in Qualitative Social Work: Research and Practice 1 (1): 25-38.

Eisner, Elliot (1991) The Enlightened Eye: Qualitative Inquiry and the Enhancement of Educational Practice. New York: Macmillan.

Elks, Marion and Kirkhart, K. (1993) 'Evaluating effectiveness from the practitioner's perspective' in Social Work, 38 (5): 554-563.

Erikson, Erik (1959) 'The nature of clinical inference' in Lerner, D. (ed) Evidence and Inference. Illinois: Free Press.

Finch, Janet (1986) Research and Policy: the Uses of Qualitative Methods in Social and Educational Research. London: Falmer Press.

Geertz, Clifford (1973) The Interpretation of Cultures. New York: Basic Books.

Greene, Jennifer and Caracelli, Valerie (1997) Advances in Mixed Method Evaluation: the Challenge and benefits of Integrating Diverse Paradigms. New Directions For Evaluation, No 74, San Francisco: Jossey-Bass.

Hall, Tom (2001) 'Caught not taught: Ethnographic research at a young people's accommodation project' in Shaw, I and Gould, N. Qualitative Research in Social Work London: Sage Publications.

Hammersley, Martyn, (1993) 'On the teacher as researcher', in Educational Action Research, 1 (3): 425-445.

House, Ernest (1991) 'Realism in research', in Educational Researcher, 20 (6) 2-9.

Humphrey, Chris and Pease, Ken (1992) 'Effectiveness measurement in the Probation Service: a view from the troops' in Howard Journal, 31 (2): 31-52.

Lather, Patti (1986) 'Issues of validity in openly ideological research', in Interchange, 17 (4): 63-84.

Lather, Patti (1991) Getting Smart: Feminist Research and Pedagogy with/in the Postmodern. New York: Routledge.

Lofland, John and Lofland, Lynn (1995) Analysing Social Settings. Belmont: Wadsworth.

McKeganey, Neil, MacPherson, I. And Hunter, D. (1988) 'How »they« decide: exploring professional decision making' in Research, Policy and Planning, 6 (1): 15-19.

Miles, Matthew and Huberman, A, (1994) Qualitative Data Analysis: an Expanded Sourcebook. Thousand Oaks: Sage.

Miller, Gale (1997) 'Contextualizing Texts: Studying Organisational Texts' in Miller G and Dingwall R (eds) Context and Method in Qualitative Research London: Sage.

Mohr, Wanda K. (1997) 'Interpretive Interactionism: Denzin's Potential Contribution to Intervention and Outcomes Research' in Qualitative Health Research 7 (2): 270-286.

Pawson, Ray and Tilley, Nick (1997b) 'An Introduction to scientific realist evaluation', in Chelimsky, E. and Shadish, W. (eds) Evaluation for the 21st Century. Thousand Oaks: Sage.

Reid, William J. (1994) 'Reframing the epistemological debate', in Sherman, E. and Reid, W. (eds), Qualitative Research in Social Work. New York: Columbia University Press.

Scourfield, Jonathan (2001) 'Interviewing interviewers and knowing about knowledge' in Shaw, I and Gould, N. Qualitative Research in Social Work London: Sage Publications.

Shaw, Ian (1999) 'Seeing the trees for the wood: the politics of evaluating in practice', in Broad, B. (ed), The Politics of Research and Evaluation. Birmingham (England): Venture Press.

Shaw, Ian (1996) Evaluating in Practice, Aldershot (England): Ashgate.

Shaw, Ian (1997) Be Your Own Evaluator: a Guide to Reflective and Enabling Evaluating. Wrexham (Wales): Prospects Publishing.

Shaw, Ian and Shaw, Alison (1997a) 'Keeping social work honest: evaluating as profession and practice', in British Journal of Social Work, 27 (6): 847-869.

Shaw, Ian and Shaw, Alison (1997b) 'Game plans,

buzzes and sheer luck: doing well in social work', in Social Work Research, 21 (2): 69-79.
Silverman, David (1997a) Discourses of Counselling: HIV Counselling as Social Interaction. London: Sage.
Silverman, David (1997b) 'The logics of qualitative research' in Miller, G. and Dingwall, R (eds) Context and Method in Qualitative Research. London: Sage.
Sinclair, Ian (2000) 'Methods and measurement in evaluative social work', paper from ESRC seminar series 'Theorising social work research'. National Institute for Social Work web site http://www.nisw.org.uk/tswr/sinclair.html
Stake, Robert and Trumbull, D. (1982) 'Naturalistic generalizations' in Review Journal of Philosophy and Social Science, 7, 1-12.
Trend, M.G. (1979) 'On the reconcilation of qualitative and quantitative analyses, in Cook, T and Reichardt, C. Qulitative and Quantitative Methods in Evaluation Research. Beverly Hills: Sage.
Weiss, Carol (1980) 'Knowledge creep and decision accretion', in Knowledge, Creation, Diffusion, Utilisation, 1 (3): 381-404.
White, Sue (2001) 'Auto-ethnography as reflexive inquiry: the research act as self surveillance' in Shaw, I and Gould, N. Qualitative Research in Social Work London: Sage Publications.

Några problem i utvärdering av sociala interventioner och utfallsstudier

haluk soydan & bo vinnerljung

Med ökat intresse i studier av interventioners effekter har i större utsträckning frågor kring utfall blivit tydliga. Denna artikel beskriver några av de problem som författarna mött i utvärderingsverksamhet. Metodproblem i utfallsstudier kan vara mycket allvarliga och påverka analys och slutsatser på ett oberäkneligt sätt.

Inledning

Utfall kan definieras som avsedda och icke-avsedda förändringar hos enheter till följd av påverkan från sociala interventioner. Enheter som är föremål för interventioner och sociala program kan vara personer, familjer, grupper, bostadsområden, institutioner och organisationer.

Inom verksamhetsområden som socialtjänst, hälso- och sjukvård, kriminalvård– har kravet på empiriska kunskaper om interventioners effekter ökat i Sverige under 1990-talet. Internationellt är detta inget nytt. Särskilt i USA har beslutsfattare och policy-ansvariga sedan 1960-talet ställt krav på professionella grupper att ta reda på och redovisa utfall av sociala interventioner och program (se Albaek 1988). Uppföljning av insatser, avläsning av insatsers effekter, informationsåterföring till insatsansvariga har också blivit ett kvalitetskriterium för professionellt arbete. Sammantaget har detta skapat en omfattande forskning kring utvärdering som främst under 1990-talet började sprida sig till Europa. Den amerikanska utvärderingsforskningen har använt sig av en rad olika designmodeller, från randomiserade kontrollerade försök till kvalitativa studier (Rossi & Freeman 1993, Oakley 2000).

Haluk Soydan är professor i socialt arbete och arbetar som forskningsledare vid Centrum for utvärdering av socialt arbete.
Bo Vinnerljung är docent i socialt arbetet och arbetar som forskningsledare vid Centrum för utvärdering av socialt arbete.

Identifiering, kartläggning och analys av sociala interventioners utfall är förknippade med en lång rad kunskapsteoretiska, metodologiska och mättekniska utmaningar. Syftet med föreliggande artikel är att lyfta fram några vanliga metodologiska problem som genereras vid studier av sociala interventioners utfall.

Efter att ha berört några begreppsliga aspekter av utfall i sociala interventioner kommer vi mer specifikt att behandla fyra teman:

1) I utvärderingsstudier fokuseras ofta, och många gånger uteslutande, på ett utfallsmått. Vi kommer att argumentera för nödvändigheten av mätning och analys av flera utfallsvariabler i utvärderingar av sociala interventioner.
2) När en experimentell design inte kan användas i utvärderingsstudier och när särskilda grupper som t ex adopterade barn följs upp och jämförs med normalpopulationen kan en snedfördelning av specifika problem i undersökningsgruppen skapa problem i utfallsanalysen.
3) Tidens roll i utfallsstudier är av kritisk betydelse. Vi kommer att diskutera några besvärande problem med tidsfaktorn i utfallsstudier.
4) Att klienter inte fullföljer påbörjade sociala program – att de blir drop-outs, eller avvisade från programmet – skapar avgörande bortfallsproblem i utvärdering av sociala interventioner. Vi kommer att belysa problemet med avbrutna insatser i utvärderingsstudier.

Utfall i sociala interventioner – några begreppsliga aspekter

Innan dessa fyra teman diskuteras vill vi fästa uppmärksamheten på mångsidigheten av begreppet »utfall« i utvärderingar av sociala interventioner för att ge ett perspektiv på de frågor som mer ingående belyses i artikeln.

När sociala interventioner genomförs vill man naturligtvis åstadkomma en positiv förändring hos enheten (personen, gruppen, organisationen osv.). Exempel är målsättningen om att få personer att sluta med missbruk eller att återfalla i brott etc. I en stadsdel vill man kanske minska arbetslösheten och/eller minska kriminaliteten. När en social intervention genomförs aktualiseras minst tre centrala aspekter vad gäller förändringen hos enheten: 1) Har sociala interventionen genererat ett eller flera utfall? 2) Är effekterna avsedda och icke-avsedda? 3) Är effekterna resultat av den genomförda sociala interventionen och/eller externa faktorer? Frågan om flera utfall behandlas i nästa avsnitt. I det följande vill vi beröra de två senare aspekterna.

Avsedda effekter av sociala interventioner är lätta att identifiera i den bemärkelsen att sociala program vanligtvis brukar ange målet/målen för interventionen. Icke-avsedda effekter kan vara såväl icke önskvärda som önskvärda men är inte självklara och omedelbart synliga för utvärderaren. De kan spåras på flera olika sätt, exempelvis genom mätning av vad som rimligen är relevanta utfallsvariabler men som inte är angivna som direkta mål för interventionen. I KrAmi-utvärderingen (se annan

artikel i denna tidskrift) omfattas flera avsedda effekter: förbättrad arbetsmarknadsförankring och egenförsörjning samt avståndstagande från kriminalitet. Förändringar i programdeltagarnas familje- och sociala relationer, och förändringar i hälsotillstånd är icke-avsedda utfall i samband med exponering för KrAmi-programmet. Förbättrade familjerelationer och hälsoförhållanden är inte en del av KrAmi-programmets uttalade mål, men identifierades tack vare mätinstrumentets (Addiction Severity Index) förmåga att rutinmässigt mäta dessa två utfallsvariabler.

Ett centralt problem i utvärdering av sociala interventioner är frågan om ett givet utfall är en effekt av själva interventionen. Utfall som utvärderaren mäter hos enheter som exponeras för en intervention kan vara såväl interventionseffekter som effekter av externa faktorer. Det centrala problemet för utvärderaren är att om möjligt särskilja dessa två grupper av utfall. Det klassiska metodologiska greppet för att klara detta är användning av randomiserade experiment med kontrollgrupp (i andra hand kvasiexperimentella studier med jämförelsegrupp som uppstått utan forskarens medverkan) med för- och eftermätningar (se Boruch 1997).

Med hjälp av design och olika statistiska metoder försöker forskaren att kontrollera en mängd oönskade effekter av olika art: selektionseffekter, bortfall, statistiska effekter etcetera. Aspekter av selektionseffekter och, bortfall diskuteras särskilt i denna artikel. Vi vill dock nämna ytterligare två som försvårar tillvaron för forskare som arbetar med utfallsstudier. Individers livshistoria är ett »mätproblem«. Interventioner i behandlings- och socialt förändringsarbete löper ofta under en längre tid. Samtidigt händer också en rad andra saker i dessa individers liv. Någon eller några av dessa livshändelser som äger rum utanför interventionen kan ha avgörande betydelse för de utfall som utvärderaren är intresserad av, exempelvis att personer får barn, förälskar sig eller att de får en ny omvälvande livssyn. Lika »besvärligt« för utvärderingen är ett annat känt problem som brukar benämnas »personlig mognad«. Inom brottspreventionsområdet till exempel har det visat sig väsentligt att åtskilja interventionseffekter från individens personliga mognad över tid som bland annat har samband med biologisk ålder (refMcGuire 1995).

Nödvändigt att arbeta med flera utfallsmått

Sociala interventioner generar i regel flera utfall samtidigt eller i en given tidssekvens, medan forskare, praktiker och beslutsfattare tenderar att intressera sig för ett utfall i taget. Ett typiskt exempel är forskningen om brottsåterfall. Medan metaanalyser huvudsakligen sysslar med utfallsvariabeln 'brottsåterfall' genererar interventioner som avser att begränsa eller förhindra brottsåterfall även andra (positiva) utfall.

Frågan om vad som är adekvata – och rimliga – mått på utfall av insatser är ständigt närvarande i utvärderingsstudier. Inom området insatser för asociala barn/ungdomar är återfall i brott (»recidivism«) det traditionella utfallsmåttet när olika behandlingsprogram jämförs med varan-

dra. Framförallt gäller detta i metaanalyser eller systematiska forskningssynteser (se t ex Lipsey & Wilson 1998; Lipsey 1992, 19956). Huvudskälet till denna starka fokusering på återfall är förmodligen det starkt uttalade intresset bland politiker och andra beslutsfattare för interventioner som leder till reducerad brottlighet. Inom detta område kom först i början 1980-talet kvalificerade litteraturöversikter (se t. ex. Gendreau & Ross 1979; Ross & Gendreau 1980). Under 1990-talet började metaanalyser publiceras, särskilt pionjärinsatser av Mark Lipsey (se t. ex. Lipsey 1992, 1995). Även om dessa metaanalyser fokuserar på återfall i brott som den huvudsakliga utfallsvariabeln finns möjligheter i underlaget till systematiska synteser av andra utfallsvariabler. Eftersom kunskapsdatabasen för Lipseys metaanalyser rutinmässigt registrerar ett stort antal variabler inklusive olika utfallsvariabler var det möjligt att studera några av dessa i en särskild studie (Wilson, Lipsey and Soydan 2002). Denna metaanalys bygger på 142 studier av interventioners effekter på ungdomsbrottslighet. Forskarna arbetar i analysen inte bara med utfallsvariabeln »återfall i brott« utan även med andra resultat som »anpassning till skolan«, »reducering av beteendeproblem« och attitydförändringar.

Andra studier har använt betydligt mer sammansatta mått på »hur det gått för klienterna«. Ibland har detta tvingat fram frågor om hur mycket som är rimligt att kräva för att en genomgången behandling ska anses som framgångsrik. Ett bra exempel är det svenska SWEDATE-projektet, som fann att om man bara hade kravet »drogfri« senaste 6 månaderna, då hade det gått bra för 51 procent av 387 institutionsbehandlade missbrukare ett år efter behandling. Om man dessutom ställde krav på att inget annat missbruk (t ex av alkohol) skulle förekomma, sjönk siffran till 37 procent. När forskarna dessutom lade till kraven »inte mer än två månaders institutionsvård under uppföljningsperioden« och »ingen kriminalitet« blev det bara 22 procent »lyckade fall« kvar. Efter att ha lagt till ytterligare en rad krav (inkomst från arbete eller utbildning, ordnat boende, inga allvarliga psykiska problem vid uppföljningen, inget behov av socialt stöd m m (alla vanliga mål för behandling av missbrukare) återstod till sist bara 10 procent som klarade sig bra (Berglund et al. 1991).

Ett annat exempel på kombinationer av utfall är Maja Andersons (1976) klassiska registeruppföljning av Gustav Jonssons och Anna-Lisa Kälvstens 222 Stockholmspojkar (1964) och 100 pojkar som varit på institutionen Skå i sin barndom (Jonsson, 1967). Hennes rapport är fortfarande en lärorik läsning idag, 25 år efter att den publicerades. Uppföljningen genomfördes när »pojkarna« var 20–34 år. Anderson konstruerade sina utfallskriterier efter vad hon tolkade var den dominerande uppfattningen om en önskvärd utveckling i Sverige på 1960-talet. Hon använde en sjugradig skala på social anpassning. Längst ner på skalan (1-2) fanns de som ej levde upp till samhällets normer för att man »klarar sig«. De varken arbetade eller studerade, en del var omhändertagna av samhället eller var till och från föremål för samhällets åtgärder. Mitt på skalan (3-4) fanns de som klarade sig »fast lite knackigt« (i Ander-

son, 1976, s 22). De hade brister i vissa avseenden men det fanns också positiv information i registren om deras liv. Högst upp (5-7) fanns de som lyckats bra åtminstone i något avseende, exempevis skaffat sig ett yrke. Likaledes fanns en liknande skala för självförsörjning med tre grupper. »Självförsörjande« var män som inte registrerats för kriminalitet eller fylleri (förutom lindriga förseelser) och som ej varit ideligen sjukskrivna och som 1971 hade en minimiinkomst. »Någon brist« var de som hade en registerindikation på problem i något av dessa avseenden. »Ej självförsörjande« var en grupp som klart inte uppfyllde de tre villkoren för »självförsörjande«. Även om kategoriseringarna innebar en del kvalitativa tolkningar som inte kan spåras i detalj i rapporten (jfr Börjeson & Håkansson, 1990), framstår Andersons sammansatta utfallsmått både rimliga och trovärdiga även idag.

David Magnussons och Håkan Stattins (med fleras) banbrytande forskning om barns utveckling över tid, särskilt risk för asocialitet i tonåren och i vuxen ålder, har långtgående implikationer för frågan om utfallsmått i utvärderingar av insatser för barn och ungdom. Deras arbeten baseras på flera decenniers studier av longitudinellt material, främst den s k IDA-undersökningen i Örebro. I den redan klassiska artikeln »Antisocial development: A holistic approach « (Stattin & Magnusson 1996) sammanfattas några centrala fynd. Författarna visar att anpassningsproblem över tid tenderar att samlas och förstärkas hos en liten grupp med dyster prognos (problemgravitering och -aggregering). De som är kriminella tenderar också att vara de samma som använder droger etc. Ett illustrativt exempel finns i en studie av mobbande barn/ungdomar (Andershed, Kerr & Stattin 2000). Forskarna fann att det till stor del är samma barn som mobbar andra på skolgården, som bär vapen på stan och utövar våld i gatumiljön.

Extremgruppen med multiproblem är så dominerande i longitudinella material, att om de exkluderas i uppföljningar får enskilda riskfaktorer under barns uppväxt ett svagt prediktivt värde för senare utveckling (Stattin & Magnusson, 1996). Detta gäller för flera utfall, till exempel brottsbelastning och missbruk. Det är rimligt att anta att denna grupp även har en stark påverkan på utfall av interventioner riktade mot asociala ungdomar.

Av detta följer för det första att utvärderingar – när det är möjligt –bör arbeta med flera olika utfallsvariabler och med aggregeringar av dessa utfallsvariabler för att realistiskt kunna gradera resultatet av en genomgången behandling, exempelvis av asocialt beteende hos ungdomar. För det andra bör utvärderingar av insatser för barn och ungdomar med kvasiexperimentell design inte begränsas till att hålla kontroll av enskilda (risk-) variablers distribution mellan undersöknings- och jämförelsegruppen, utan forskaren bör även söka efter variabelmönster, d v s leta efter individer som kan antas tillhöra extrema riskgrupper.

Snedfördelningar

Många sociala insatser är svåra, nära nog omöjliga, att utvärdera med experimentell design. Inom barnavården gäller det exem-

pelvis fosterhemsplaceringar och adoptioner. Dessa insatser har dessutom mycket långtgående syften över tid, ofta att ge utsatta barn en trygg uppväxt under många år. Mycket av den »utvärderande« kunskap vi har om exempelvis långvarig fosterhemsvård kommer från studier av före detta fosterbarn i vuxen ålder (se översikt i Vinnerljung, 1996.). Man brukar här tala om »globalautfall«, vilket även gäller adoptionsforskningen. I dessa studier jämförs vuxna före detta fosterbarn och adopterade med jämnåriga i normalpopulationen (se t ex Vinnerljung 1995), alternativt med andra grupper som haft jämförbara utgångslägen i tidig barndom (t ex Vinnerljung, 1996; Vinnerljung & Ribe, 2001). Ofta undersöks förekomst av relativt sällsynta oönskade utfall, såsom mortalitet, psykisk sjukdom och allvarlig kriminalitet (t ex Hjern, Lindblad & Vinnerljung, 2002).

Flera adoptionsforskare har under senare år uppmärksammat problemet med sneda fördelningar (»skewed distributions«). I en diskussion om tillsynes motstridiga resultat i adoptionsforskningen – högre förekomst i kliniska populationer av adoptivbarn än icke-adopterade men avsevärt mindre skillnader i populationsbaserade undersökningar – visade Jeffery Haugaard (1998) på hur dessa resultat teoretiskt kunde förklaras av en snedfördelning inom adoptionsgruppen jämfört med normalpopulationen. Om det fanns en liten grupp med mycket dåliga utfall inom adoptionsgruppen, kunde en stor del av skillnaderna jämfört med en normalgrupp förklaras av detta fenomen, exempelvis om forskarna huvudsakligen använde olika former av medelvärden i sina analyser. Minst två studier har visat att det verkar finnas empiriskt stöd för denna hypotes (Sharma et al,. 1998; Miller et al, 2000). Skillnaderna mellan adoptivbarn avseende exempelvis skolprestationer, skolproblem, psykiskt välmående, drogbruk m m var i dessa undersökningar störst bland de fem procent adoptivbarn som hade sämsta värden i mätningarna. Det fanns med andra ord en tydlig »puckel« av mycket belastade individer bland adoptivbarn i ena ändan av den antagna normalfördelningskurvan. Problemet återkommer i olika typer av social forskning som använder sig av sällsynta, negativa utfall – exempelvis mortalitet – som indikationer på insatsens långsiktiga effekter. Eftersom forskaren då arbetar med dikotoma data (förekomst/ej förekomst) blir det svårt att veta vad resultaten av en jämförelse med exempelvis normalbefolkningen säger oss. Är det något allmänt om risker för undersökningsgruppen, exempelvis före detta samhällsvårdade barn? Eller speglar resultaten istället att en liten del av undersökningsgruppen har en rejält dålig livssituation, medan skillnaderna för resterande delen av gruppen inte är så stora jämfört med andra jämnåriga?

Detta problem pekar på att studier av sällsynta utfall har sina uppenbara begränsningar i utfallsstudier, även de som syftar att undersöka »globala utfall«. Ofta behövs det data om hela undersökningsgruppen för att kunna säga något mera konklusivt. Erfarenheterna från adoptionsforskningen pekar på betydelsen av analysmetoder där även fördelningen av utfall i undersöknings- och en jämförelsegrupper undersöks och analyseras.

Tiden ställer till med problem

Sociala problem och beteendestörningar som blir föremål för interventioner genereras, utvecklas och får sitt »naturliga« förlopp i ett tidsperspektiv. Alkoholism, drogproblem, kriminalitet, skolk etc uppstår inte över en natt utan utvecklas över en tid. På samma sätt pågår behandling av sociala problem och beteendestörningar över tid, ibland i flera år. I studier av utfall efter avslutad intervention uppstår då en rad frågor som handlar om tid: när uppnås vad man kan kalla för »en effekt« i samband med en intervention? När är det bäst att avläsa en sådan effekt? Under eller efter behandlingen? Om svaret är »efter behandlingen«, uppstår en ny fråga: strax efter avslutad behandling eller först efter en tid? Om det senare gäller– ska det vara efter sex månader, ett år, tre år, tio år efter behandlingen? Med andra ord : hur beständig över tid är behandlingens effekter?

Den kloka forskaren inser att det inte finns ett standardiserat och receptboksbaserat svar på frågan när en intervention kan tänkas ha effekt, om hur länge interventionens effekter varar och när mätningar av utfall ska ske. Man kan kanske prata om interventioners kontextualitet i bemärkelsen att det finns en mängd skillnader när man jämför olika interventionsområden, som till exempel alkoholbehandling, brottsprevention och försörjningsstöd. Observera att med kontextualiteten avses inte här att refererar till attvarje individuell intervention är unik och därmed kontextuell med den konsekvensen att det inte går att generalisera till andra liknande interventionssituationer. Snarare är det fråga om skillnader av en annan karaktär. Beakta exemplet med att interventionen »blindtarmsoperation« har en radikal engångseffekt som är omedelbar vad gäller upphörande av den akuta inflammationen och som dessutom är beständig i tiden under individens livsförlopp. Blodstryckspatienten, ceteris paribus, måste däremot medicineras dagligen eftersom blodstryckskontrollerande preparat verkar under begränsad tid (exempelvis ett dygn) även om effekten är nästan omedelbar. Det finns med andra ord en mångfald av samverkansmönster mellan intervention och utfall med tiden som mellanliggande faktor. Här följer fler exempel som antyder vikten av dessa problem.

Ett färskt exempel visar hur snabbt effekterna av behandling av alkoholproblem interventioner eroderas. I en nyligen utgiven volym redovisar Statens beredning för medicinsk utvärdering metaanalyser om intervention mot riskfylld alkoholkonsumtion (s. k. sekundär prevention av alkoholproblem). I denna studie har man kommit fram till slutsatsen att det är »osäkert om insatser ger effekter efter mer än två år eller om man då måste upprepa interventionen« (SBU 2001:I s 55). Om denna slutsats är riktig och kan bekräftas av ytterligare studier är det givet att sekundära preventiva interventioners positiva utfall har i genomsnitt en beständighet på högst två år och att interventionerna måste upprepas efter denna period om man vill upprätthålla preventiva effekter.

Ett annat exempel kommer från drogbehandlingsområdet. I det tidigare nämnda SWEDATE Projektet, som genomfördes mellan 1981 och 1983, följde forskarna upp effekterna av drogbehandlingsprogram

hos drygt 1100 missbrukare åtta år efter insatsen. Ett för vår diskussion intressant resultat var att hela 75,2 procent av populationen hade tolv eller flera registrerade återfall i brott eller var avlidna. Dessutom var de tidigare noterade skillnaderna mellan olika behandlingsprogram nu utjämnade (Bergmark et al, 1996). Återigen ser vi här hur effekter sakta men säkert eroderas med tiden.

I en annan och betydligt mindre studie (Nyström & Soydan 1997) exponeras ett överraskande resultat. I denna utfallsstudie av ett socialt rehabiliteringsprogram för kriminella och arbetslösa, KrAmi, följdes 29 programdeltagare upp för andra gången 1996, dvs fyra år efter avslutat eller avbrutet program. I den första uppföljningen studerades två grupper, de som fullgjorde programmet och de som var drop-outs (Lindberg & Soydan 1993). Denna visade bättre resultat för gruppen som hade fullföljt programmet. Fyra år senare hade den första gruppen fortfarande positiva utfall i termer av arbete, bostad, familj, missbruk och självförtroende. Detta resultat hade alltså bestått i fyra år efter avslutat program. Men därutöver hade ett flertal i drop-out gruppen också förbättrat sin situation sedan första uppföljningen, vilket kan tolkas på olika sätt – inte minst i termer av social mognad som vi tidigare nämnt.

Sammanfattningsvis menar vi att det finns reella problem vad gäller när och hur interventioners effekter gör sig gällande och hur beständiga de är. Det återstår dock att systematiskt kartlägga typer av problem som empiriska studier visar och att utveckla strategier som kan hantera dessa problem. En rimlig »basrekommendation« är att utvärderingar av program/insatser som har långtgående ambitioner om effekters beständighet över tid, måste innehålla upprepade mätningar vid flera tillfällen efter behandlingens avslutning. Samma gäller utvärderingar av program/insatser som enbart på ideologisk (ej empiriskt baserad) grund påstår att effekterna av behandlingen har hög beständighet över tid. Det är viktigt att komma ihåg att vissa interventioner med goda intentioner – och mycket resurser –faktiskt också kan ha en skadlig inverkan över lång tid, (se t ex McCord 1978).

Drop-outs

Drop-out – avbrutna insatser – är ett mycket vanligt problem i socialt arbete (se diskussion i Vinnerljung, Sallnäs & Kyhle-Westermark 2001). Deltagare hoppar av behandlingsprogram i förtid, blir avvisade från familjehem och institutioner, avviker från behandlingshem etc.

Inom psykiatrin och psykoterapin är problemet välkänt sedan länge. Joel Fisher konstaterade i en klassisk översiktsartikel 1978 (med titeln Does anything work?) att upp till 50 procent av psykoterapiklienter kom aldrig tillbaka efter första behandlingssamtalet, och upp till 80 procent föll bort före sex planerade samtal. Hans slutsats var närmast brutal: – Thus, psychotherapy appears to reach – and hold – only a very small percentage of those people who may need it (Fisher 1978 s 221). Även senare forskningsöversikter över såväl vuxen- som barnpsykiatrisk terapeutisk behandling har funnit att höga dropout-siffror är snarare

regel än undantag (se t ex Garfield, 1994; Kadzin, 1994a, 1994 b).

Problemet är dock avsevärt mindre uppmärksammat i socialt arbete, trots tydliga bevis om problemets omfattning i behandlingsforskningen. I en översikt av öppenvårdsinsatser för föräldrar som misshandlat sina barn redogör Corcoran (2000) för flera studier där 70-80 procent av föräldrarna aldrig genomförde behandlingen. Det svenska SWEDATE-projektet som utvärderade institutionsvård för närmare 1.200 drogmissbrukare fann att hela 60 procent genomförde aldrig behandlingen. För vuxna missbrukare var siffran 70 procent (Berglund et al, 1991). Detta är inte en extremsiffra. En norsk utvärdering av ett Phoenix-house-projekt för närmare 150 narkomaner fann att hela 87% av de manliga deltagarna avbröt behandlingen i förtid (Ravndal & Vaglum 1995). I SBUs rapport (2001) om behandling av alkohol- och narkotikaproblem redovisas retentionssiffror (kvarstannande i behandling) för behandling av narkotikamissbrukare på mellan nära nog 0 och 100 procent. Just kvarstannande i behandling används till och med som »ett primärt mått på behandlingseffekt« (vol. 1,s 21), trots att drop-out-processen rimligtvis måste spegla en slags anrikning av den ursprungliga undersökningsgruppen (jfr Edwards & Rollnick 1997).

När Patricia Chamberlain och hennes medarbetare i USA startade försök med fosterhemsplaceringar av kriminella tonåringar i kombination med bland annat familjeterapi[1] var deras första fråga: är det överhuvudtaget möjligt att genomföra familjehemsplaceringar av denna grupp, utan höga sammanbrottsiffror?[2] (Chamberlain & Reid 1998; Hansson 2001 m fl; jfr Baker 1989)? I en nyligen publicerad svensk studie undersökte Vinnerljung med medarbetare en nationell kohort av 13-16-åringar som placerades i dygnsvård 1991. Ungdomarna följdes i akter under fem år, eller till 18-årsdagen. Några exempel från resultaten: i »vanliga familjehem«, (ej släkt med barnet) havererade 41-51 procent av placeringarna i förtid, siffran beroende på val av definition av sammanbrott. I en grupp små enskilt ägda institutioner var sammanbrottsfrekvensen 52 procent. Den vanligaste placeringen i kohorten var att en tonåring med asocialt beteende sattes i ett »vanligt fosterhem«. 57-67 procent av dessa placeringar slutade med sammanbrott (siffran återigen beroende på val av definition). Resultaten för samma målgrupp i institutionsvård var inte mycket bättre: 40-50 procent, beroende på definition och ägandeform. (Vinnerljung, Sallnäs & Kyhle-Westemark, 2001). De enda grupperna med jämförelsevis låga sammanbrottssiffror (<20 procent) var släktinghem [3], jourfosterhem och s k §12-hem (tidigare benämnda ungdomsvårdsskolor).

1 S k Multi-dimensional treatment foster-care. Metoden beskrivs utförligt i Hansson 2001.

2 Med sammanbrott avses att placeringen »spricker«, d v s avslutas på ett uppenbart oavsiktligt sätt. Vanligaste skälet är att fosterhemmet/institutionen inte vill fortsätta med placeringen, eller att barnet/den unge rymmer. För diskussion om definitioner och för översikt av forskning om sammanbrott i barn- och ungdomsvården se Vinnerljung, et al , 2001).

3 Fosterhem där fosterföräldrarna är släkt med barnet, vanligtvis mormor och morfar.

I utvärderingar och till och med i översikter av utvärderingsresultat är det dessvärre inte ovanligt att drop-out-frekvenser ej redovisas eller förs bort från diskussionen, ibland med motivet »data saknas för gruppen«. Detta är självfallet mycket otillfredsställande och leder sannolikt till felaktiga slutsatser i många fall. Om vi exempelvis ska utvärdera ett enskilt behandlingshem för tonåringar där majoriteten av placeringar aldrig blir genomförda – ungdomarna avvisas av personalen efter disciplinkonflikter, rymmer etc. – blir det grovt missvisande att i en uppföljning bara redovisa hur det går för den minoritet som fullföljer behandlingsprogrammet. Låt oss konstruera ett hypotetiskt exempel: säg att vi jämför två behandlingshem, och i analysen bara tar med de som fullföljt behandlingen. Resultaten visar inga skillnader mellan programmen. Hem A har avvisat 50 procent av eleverna under behandlingens gång (t ex med motivet »de passade inte in i gruppen«) medan hem B har en drop-outfrekvens på 20 procent. Denna stora, avgörande skillnad kommer inte att synas i resultaten om utvärderaren bara tar med de som fullföljt behandlingen. Då gynnas hem A medan slutsatsen »ingen skillnad« blir orättvis och felaktig för hem B.

Vi vet också att drop-out/sammanbrott är en selektionsprocess, där de som avbryter eller avvisas från pågående behandling är mer problembelastade än de som stannar kvar (se t. ex. Kadzin, 1997, Newton et al, 2000 för exempel från barn- och ungdomsvården; Edwards & Rollnick, 1997 om interventioner för alkoholmissbrukare). Men även efter kontroll för detta, visar en mycket stor metaanalys av insatser för asociala barn/ungdomar att behandlingsprogram med höga drop-out-siffror tenderar tydligt att visa sämre resultat (Lipsey & Wilson, 1998). I ljuset av den kunskapen kan man faktiskt se ungdomar som »röstar med fötterna« och avviker från olika behandlingsprogram som välinformerade konsumenter.

Hur bör det vara? I den »klassiska« utvärderingsforskningen där s k randomiserade experimentella studier är den »gyllene« designen, gäller följande vedertagna regel: once randomized – analyzed (Boruch 1997). Det finns två huvudsakliga motiv för denna fundamentala regel. Det ena är vetenskapligt baserad, och berör utvärderingens validitet. Gruppen som avbryter behandlingen skiljer sig sannolikt från den som fullföljer, alltså speglar dropouts en selektionsprocess som inte kan antas vara densamma i experimentgruppen som i jämförelsegruppen. Särskilt självklart blir detta om jämförelsegruppen består av individer på en väntelista där man har slumpmässigt valt ut vilka som ingår i experimentgruppen. Det andra motivet är policy baserat och berör utvärderingens anknytning till »real-world«-förhållanden. I verkligheten utanför studien är det sannolikt att en behandling som erbjuds, föredras av policyskäl eller som ingår i lagstiftade insatser, kommer att avvisas av en del klienter och andra kommer att avvika eller avvisas från behandlingen. Det som händer under mer kontrollerade former i en utvärderingsstudie kan då sägas vara bevis på vad som skulle hända om dessa behandlingsprogram skulle implementeras i stor skala i ett land eller region, utanför

studien. Drop-out speglar med andra ord ett verkligt förhållande, inte ett av forskarna konstruerat, och det är därför av central vikt att även omfattningen av avbrutna behandlingar studeras och analyseras i en undersökning av behandlingens effektivitet.

Vi upprepar: den klassiska regeln säger att utfall ska studeras, analyseras och redovisas för alla individer som inledningsvis placeras i en behandlingsgrupp (experimentgrupp) och likaledes för alla fall som från början ingår i en jämförelsegrupp. Detta, menar vi, bör även gälla i utvärderingsstudier oavsett designform, så långt som det är praktiskt möjligt. Om det inte är möjligt att följa denna huvudregel bör forskaren klart redovisa detta och motivera varför det inte har gått. Han/hon bör även försöka bedöma konsekvenserna av att dessa data saknas, och slutligen diskutera hur detta påverkar studiens generaliserbarhet.

Avslutning

Den svenska inomvetenskapliga debatten om utvärderingar av sociala interventioner finns många kritiska röster. Ett fåtal menar till och med att det inte går att göra meningsfulla utvärderingar av exempelvis socialt arbete, eftersom metodproblemen är oöverstigliga. Även om vi anser att sådana resonemang i princip är ovetenskapliga (hur kan vi så säkert veta att det aldrig kommer att gå?), menar vi att det är viktigt att öppet och konkret diskutera de reella metodproblem som finns finns. I denna artikel har vi försökt belysa aspekter av några konkreta problem, och pekat på ställningstaganden och utvägar ur de dilemma som uppstår. Vi vill betona att det främsta syftet med denna artikel har varit att uppmärksamma problem som sällan diskuteras i handböcker om utvärderingsmetodologi. De slutsatser som vi drar av det som presenterats i artikeln är:

Visst – det finns problem med enskilda utfallsmått. I verkligheten åstadkommer praktiskt socialt arbete flera utfall genom en given intervention. Även forskarna själva mäter i många undersökningar flera utfallsvariabler, utan att sedan explicit arbeta med dessa i analyser., och vi bör när det är möjligt arbeta med mer sammansatta kriterier på utfall

Var vaksam mot förekomst av sällsynta fenomen som utfallsmått i uppföljningar. De kan spegla snedfördelningar som vilseleder både forskaren och konsumenten av den gjorda genomförda utvärderingen. Sök i stället efter utfallsmått som ger information om alla samtliga studieenheter i både undersöknings- och jämförelsegruppen. Undersök dessutom om en liten multi-belastad problemtyngd extremgrupp har ett högt förklaringsvärde för resultaten.

Frågan om hur man ska hantera tidsfaktorn på ett rimligt sätt i uppföljningar har inte ett givet svar. Här bör varje forskare vara öppen för såväl kliniska erfarenheter som konsensusuppfattningar i den internationella forskningen.

Drop-outs och avbrutna behandlingar är ett stort och reellt problem i utvärderingar av sociala interventioner, och är sannolikt större ju svårare problem som interventionen avser att påverka (se t ex sammanställ-

ning i SBU, 2001). Grundprincipen »once randomized – analyzed« – överförd även till studier med annan design än det randomiserade experimentella försöket – bör vara utgångspunkten. Det gäller både för analys och för diskussion om de resultat som undersökningen redovisar.

Vi tackar Annika Puide och Marie Sallnäs som lämnat värdefulla kommentarer på en tidigare version av denna artikel.

Referenser

Albaek, E. (1988) Fra sandhed till information. Evalueringsforskning i USA – förr och nu. Viborg: Akademisk Forlag.

Andershed H., Kerr, M. & Stattin, H. (under tryckning) Bullying in school and violence in the streets: are the same people involved? Journal of Scandinavian Studies in Criminology and Crime Prevention.

Anderson, M (1976) Hur går det för 50-talets Stockholmspojkar? En uppföljning av 222 vanliga skolpojkar och 100 Skå-pojkar. Stockholm: Monografier utgivna av Stockholms kommunalförvaltning, nr 38.

Baker, J. (1989) Therapeutic foster parent: professionally or emotionally involved parent? Child and Youth Services , vol 12, s 149–157.

Berglund, G., Bergmark, A., Björling, B., Grönbladh,L., Lindberg, S,, Oscarsson, L., Olsson, B., Segraeus V., & Stensmo C. (1991) The SWEDATE-project: interaction between treatment, clinet background and outcome in a one-year follow-up. Journal of Substance Abuse Treatment, vol 8, s 161–169.

Bergmark, A., Lindberg, S., Olsson, B. & Oscarsson, L. (1996) A long-term follow-up of residentially treated drug abusers. Paper presenterat vid 22:nd Annual Alcohol Epidemiologial Symposium of the Kettil Nruun Society for Social and Epidemiological Research on Alcohol, Edinbugh June 3–7, 1996.

Boruch, R. (1997) Randomized experiments for planning and evaluation. A practical guide. Thousand Oaks, Ca: Sage Publ Inc, Applied Social Research Methods series vol 44.

Börjeson, B. & Håkansson, H. (1990). Hotade, försummade, övergivna. Är Familjehemsplacering en möjlighet för barnen? Stockholm: Rabén & Sjögren.

Chamberlain P & Reid J (1998) Comparison of two community alternatives to incarceration for chronic juvenile offenders. Journal of Consulting and Clinical Psychology, 66:4,s 624–633.

Corchoran, J. (2000) Family interventions with child psysical abuse and neglect: a critical review. Children and Youth Services Review 22: 7, s 563–591.

Edwards, A. & Rollnick, S. (1997) Outcome studies of brief intervention in general practice: the problem of lost subjects. Addiction 92:12, s 1699–1704.

Fisher, J. (1978) Does anything work? Journal of Social Service Research, 1:3, s 215–243.

Garfield, S. (1994) Research on client variables in psychotherapy. I Bergin, A. & Garfield, S. (red) Handbook of Psychotherapy and Behavior Change. New York: Wiley & Sons.

Gendreau, P. & Ross, B. (1979) Effective correctional treatment: Bibliotherapy for cynics. Crime and Delinquency 25: 463-489.

Hansson K (2001) Familjebehandling på goda grunder. En forskningsbaserad översikt. Stockholm: Gothia/CUS.

Haugaard, J. (1998) Is adoption a risk factor for the development of adjustment problems? Clinical Psychology Review 18:1, s 47–69.

Hjern, A., Lindblad, F. & Vinnerljung, B. Suicide, psychiatric illness and social maladjustment in intercountry adoptees in Sweden. The Lancet, 360, s 443-448.

Jonsson, G. (1967) Delinquent boys, their parents and grandparents. Akadem. avhandl. Acta Psychiatrica Scandinavica, vol 43, 1967, suppl 195.

Jonsson, G. & Kälvesten A-L. (1964) 222 Stockholmspojkar. Stockholm: Almqvist & Wiksell.

Kadzin, A. (1994a) Methodology, design and evaluation in psychotherapy research. I Bergin, A. & Garfield, S. (red) Handbook of Psychotherapy and Behavior Change. New York: Wiley & Sons.

Kadzin, A. (1994b) Psychotherapy for children and adolescents. I Bergin, A. & Garfield, S. (red) Handbook of Psychotherapy and Behavior Change. New York: Wiley & Sons.

Kazdin, A. (1997) Practioner review: psychosocial treatments for conduct disorder in children. Journal of Child Psychology and Psychiatry 38: 2, s 161–178.

Lipsey, M. (1992) Juvenile delinquincy treatment: A metaanalytic inquiry into the variability of effects in T. D. Cook, H Cooper, D. S. Cordray, H. Hartmann, L. V. Hedges, R. J. Light, T. A Louis, F. Mostellar, red., Meta-analysisi for explanation. A Casebook. New York: Russell Sage Foundation.

Lipsey, M. (1995) What do we learn from 400 research studies on the effectiveness of treatment with juvenile delinquents? I J. McGuire, ed., What works: Reducing reoffending. Guidelines from research and practice. Chichester: John Wiley & Sons.

Lipsey, M. & Wilson, D. (1998) Effective intervention for serious juvenile offenders: a synthesis for Research. I Loeber, R. & Farrington, D. (red) Serious & violent juvnile offenders. Risk factors and successful interventions. Thousand Oaks, Ca: Sage Publications.

Lindberg, O. & Soydan, H. (1993). Att bli människa på nytt: En studie av socialt förändringsarbete vid KrAmi i Örebro. Högskolan i Örebro: Forskargruppen i socialt arbete, Rapport 2.

McCord, J. (1978) A thirty-year follow-up of treatment effects. American Psychlogist, March 1978, s 284–289.

McGuire, J. Red. (1995) What Works. Reducing reoffending. Guidelines from research and practice. Chichester: John Wiley & Sons.

Miller, B., Fan, X., Christensen, M., Grotevant, H. & van Dulmen. M. (2000) Comparisons of adopted and nonadopted adolescents in a large, nationally representative sample. Child Development 71:5, s 1458–1473.

Newton R, Litrownik A & Landsverk J (2000) Children and youth in foster care: disentangling the relationships between problem behaviors and number of placements. Child Abuse & Neglect, 24:10, s 1363–1374.

Nyström, S. & Soydan, H. (1997) Förändringars beständighet och social värdighet en uppföljning av KrAmi-verksamheten i Örebro . Socialvetenskaplig tidskrift. 1: 22-42.

Oakley, A. (2000) Experiments in Knowing. Gender and method in the social sciences. New York: The New Press.

Ravndal, E. & Vaglum, P. (1995) Psykiske belastninger og frafall blant stoffemisbrukere i behandling. Oslo: Sosial- og Helsedepartmentet – Rusmiddeldirektoratet.

Ross, R. & Gendreau, P. (red) (1980) Effective correctional treatment. Toronto: Butterworths.

Rossi, P. H. & Freeman, H. E. (1993) Evaluation. A systematic approach. Newbury Park: SAGE Publications.

Rutter M., Giller H. & Hagell A. (1998) Antisocial

behavior by young people. Cambridge: Cambridge University Press.
SBU (2001) Behandling av alkohol- och narkotikaproblem. En evidensbaserad kunskapssammanställning, volym 1+2. Stockholm: SBU.
Sharma, A., McGue, M. & Benson, P. (1998) The psychological adjustment of United States adopted adolescents and their nonadopted siblings. Child Development 69:3, s 791–802.
Stattin, H. & Magnusson, D. (1996) Antisocial development: A holistic approach. Development and Psychopathology, 8, 617-645.
Vinnerljung, B. (1995) Mortalitet bland fosterbarn som placerats före tonåren. Socialvetenskaplig Tidskrift 2:1, s 60–72.
Vinnerljung, B. (1996) Fosterbarn som vuxna. Akad avhandling. Lund: Arkiv Förlag.
Vinnerljung, B. & Ribe, M. (2001) Mortality after care among young adult foster children in Sweden. International Journal of Social Welfare 10:3, s 164–173.
Vinnerljung, B., Sallnäs, M. & Kyhle-Westermark, P (2001) Sammanbrott vid tonårsplaceringar – om ungdomar i fosterhem och på institution. Stockholm: CUS/Socialstyrelsen.
Wilson, S. J., Lipsey, M. & Soydan, H. (2002) Are mainstream programs for juvenile delinquency less effective with minority youth than majority youth? A meta-analysis of outcomes research . Research on social work practice (utkommer)

Summary

Some problems in evaluation of social interventions and outcome studies

The aim of this article is to discuss methodological problems that frequently occur in outcome studies of social interventions. Over the years we have observed a number of problems that disturb outcome studies. Four problems that evaluation researchers should be careful about are focused on.

1. Researchers (as well as practitioners who use results of those studies) often focus on only one outcome variable in intervention studies. However, social interventions usually generate a set of outcomes in a given point of measurement. Also, researchers often measure several outcomes without necessarily analyzing those registered outcomes. It seems more efficient to work with a multi-outcome approach trying to understand impact of social interventions.

2. Sometimes when experimental design in evaluation is not an option, follow up studies are used instead (e.g. for adopted children). Comparisons with the normal population may then lead to problems in the final analysis as specific groups may have skewed distributions. We draw the reader's attention to these problems especially when using rare phenomena as outcome variables.

3. Social and behavioral problems emerge and develop over a time span rather that instantly. Interventions also take time. A

crucial question is then raised: when does an effect occur in the course of an intervention and when should effects be measured? We find no self-evident guidance in the literature and recommend openness to clinical experiences and consensus among researchers.

4. Dropouts are a very common phenomenon in social work practice. Participants in social intervention programs leave the program at an early stage, are expelled. This problem affects outcomes of the intervention program. We draw the reader's attention to the importance of including all individuals who once were allocated in treatment as well as control or comparison groups.

Multi-site Evaluation and Research

edward mullen

This article examines multi-site evaluation and research. Differing from single-site research, multi-site research is appropriate for reasons that distinguish it from single-site research. This article examines forms and types of multi-site research to illustrate a variety of applications. The article presents examples of multi-site research conducted in the United States, at the national level with reference to mental health services and alcoholism treatment research applications. Additional mental health, child welfare, and social work practice examples are provided from research conducted at the Center for the Study of Social Work Practice in New York City. Advantages and disadvantages of multi-site research are described with suggestions for the conduct of multi-site research.

Edward J Mullen, B.A., M.S.W., D.S.W.; Willma & Albert Musher Chair Professor for Life Betterment through Science and Technology, Columbia University School of Social Work; Director, Center for the Study of Social Work Practice, a joint program of Columbia University School of Social Work & the Jewish Board of Family & Children's Services; Director, NIMH funded PhD Training Program in Mental Health Services Research.
The assistance of Haluk Soydan, Gretchen Borges, and Chito Trillana in the preparation of this paper is acknowledged.

Introduction

This article examines multi-site evaluation and research.[1] Differing from single-site research, multi-site research is appropriate for reasons that distinguish it from

1 The term "research« is used and it is meant to include both "program evaluation« as well as other forms of intervention and survey research.

single-site research. In this article I examine forms and types of multi-site research to illustrate a variety of applications. I describe examples of multi-site research conducted in the United States, at the national level with reference to mental health services and alcoholism treatment research applications. Drawing from my experiences as director of research at a social work research center, the Center for the Study of Social Work Practice (CSSWP), I describe additional applications in mental health, child welfare, and social work practice.[2] I conclude with reflections on advantages and disadvantages of multi-site research drawing out suggestions for the conduct of multi-site research.

Little has been written about the methodology of multi-site research. There are notable exceptions. An informative article published in French by Irene Elkin was written in 1992 based on her ten-year experience as Coordinator of the U.S. National Institute of Mental Health (NIMH) Treatment of Depression Collaborative Research Program (Elkin, I., 1992). Publications based on the experiences of Project MATCH, a multi-site study examining alternative alcoholism treatments, are another important contribution (Fuller RK, Mattson ME, Allen JP, Randall, CL, Anton, RF, Babor, TF, 1994;

2 As Director of the Center for the Study of Social Work Practice I have had the opportunity to observe a wide range of multi-site research applications in social welfare. The Center has lent itself to multi-site research because of its organizational sponsorship as well as its mission. The Center is located at Columbia University in the City of New York. It is a joint program of the Columbia University School of Social Work (CUSSW) and the New York City based Jewish Board of Family and Children's Services (JBFCS). Research studies have been conducted by principal investigators who are faculty members at the university. Accordingly, their interests have been national and international in scope supporting studies located at sites in New York City as well as in other locations throughout the United States and to some extent in other countries. Because of the Center's affiliation with the JBFCS the Center has tended to carry out research at locations served by that agency's programs, oftentimes through multi-site research projects. Indeed, the population served by JBFCS and, therefore, the focus of the Center's work is comparable in size and diversity with that of many national service organizations in other countries. The agency has a target population of over eight million New York residents spread over five boroughs. Over the period of one-hundred years the JBFCS has grown into one of the United States' largest nonprofit mental health and social service agencies. Now, JBFCS is a comprehensive agency that serves over 54,000 New Yorkers annually from all religious, ethnic, and economic backgrounds through 140 community-based programs, clinics, residential facilities, and day-treatment centers. JBFCS employs 1,400 staff including professional social workers, licensed psychologists, and psychiatrists, as well as a cadre of clinical support personnel in continuing day treatment and residential treatment centers. In addition services are provided by approximately 1,700 volunteers. Many of the Center's studies have been carried out at one or more of the ten JBFCS' New York State licensed outpatient mental health clinics, which provide mental health services for a wide range of emotional and social problems. Services for adults and children include evaluation and assessment; crisis intervention; and time-limited, time-effective, and ongoing individual, couple, family, and group therapy.

Zweben A, Donovan DM, Randall CL, et al, 1994; Del Boca, FK, 1994). Another informative paper has been prepared by Robert Boruch and Ellen Foley, scheduled to appear in a book edited by Leonard Bickman (Boruch and Foley, in preparation). Also, Boruch and Lawrence Hedges have examined one type of multi-site evaluation research in their article which appears in this special issue of Socialvetenskaplig Tidskrift. Indeed, there appears to be a scarcity of published analyses of multi-site research and evaluation methods. While multi-site studies have been common in recent years, it is as if there has been an assumption that it is sufficient to follow methods developed for use in single-site research. As a result, mistakes have been made and opportunities have been lost in many past multi-site studies.

Many questions need to be examined regarding multi-site research. What is multi-site research? Why conduct multi-site research? What is to be gained through multi-site studies? What is lost through multi-site investigations? What infrastructures are needed to successfully implement multi-site investigations? Elkin notes that when the NIMH Treatment of Depression Collaborative Research Program was considered a range of questions was posed (Elkin, I, 1992). These are important questions to consider more generally when undertaking multi-site research: Would it be possible to find researchers willing to undertake a lengthy collaborative effort? Would it be possible for individual researchers to place their research interests secondary to the general shared goals? Would it be possible to achieve uniformity across sites so as to consider the study a replication across sites? Would it be possible to maintain the interests and commitment of the research groups through such a lengthy enterprise? Would it be possible to analyze, interpret, and write up findings in a collaborative and mutually satisfactory fashion? Would the collaborative multi-site model prove worthwhile and superior to independent, single-site studies regarding drawing inferences about the effectiveness of treatments?

In some areas of research multi-site studies have become commonplace. However, because of their expense - in terms of time, money, and staff - as noted by Fuller, et al (1994) multi-site studies are generally undertaken only after preliminary data or policy considerations indicate the need for large, representative samples. As reported by Fuller, et al, (1994) funding has followed this trend. The United States National Cancer Institute and the National Eye Institute fund multi-site studies almost solely and the National Heart, Lung, and Blood Institute allocates about half of its clinical trial resources to multi-site studies. Other United States' National Institutes of Health units, such as the National Institute of Mental Health and the National Institute on Alcohol and Alcoholism, are increasingly encouraging cooperative or collaborative research which typically involves multiple sites. As noted by Elkin multi-site, large scale, collaborative research is not new to the field of mental health treatment research. This type of research has been called collaborative clinical trials. While collaborative clinical trials have been common in the field

of psychopharmacology and various areas of medicine for many years their use in psychotherapy and in social work research is relatively new (Elkin, I, 1992). Also, as noted by Fuller, et al (1994) in their discussion of Project MATCH, multi-site clinical trials have been used in medicine and psychiatry for some time, but only recently have they been used in alcoholism treatment research.

Differences between multi-site and single site research

What differentiates single-site and multi-site research? The most obvious distinction arises from geography. Single-site studies are carried out in one geographical location. Multi-site research is carried out in more than one site.

Single-site research

In single-site research units for study are drawn from that location. Generalization is achieved through replication at the same site using different samples or subpopulations. Also, generalization and external validity is achieved through subsequent replication at other geographical sites. Replication studies attempt to use identical or, at least, similar methods. Replications focus on the same questions and variables as those in the original study. However, subsequent single-site replications are often redesigned. As a result of the redesign original research questions as well as original study variables are most often modified based on what was learned in prior single-site applications. Subsequent single-site studies may be better thought of as elaborations or specifications of prior studies in a research program, since the aim is most often not replication but application with different sub-samples, settings, organizational contexts, and so forth. This latter objective is pursued in an attempt to advance generalization. Indeed, such programs of progressive single-site studies may suffer from a frequent failure to replicate because of the continual modifications that occur along the way. In addition, while modern methods of meta-analysis have been developed which permit the aggregate analysis of data from multiple single-site studies, when such single-site studies are combined in meta-analytic reviews, less certainty is achieved than would be the case with true concurrent replications resulting from multi-site investigations.

Single site evaluation studies often suffer from small sample sizes due to the constraints of how many subjects are available at any given geographical location.[3] For example it is my impression that this may be the case in Sweden wherein studies conducted in one municipality may have a built-in constraint on sample size. Unfortunately, replications of single-site studies at other locations often face the same constraint due to small samples used in subsequent replications. The repeated use of a series of small samples does not overcome problems inherent in small sample size research.

3 This is often the case for evaluation and intervention research studies. However, survey research studies often include large samples.

Multi-site research: A working classification of multi-site research

Multi-site studies use more than one geographical location within the context of a single study. Unlike single-site studies, multi-site studies can increase sample size by drawing from more than one location. The number of locations can be influenced by statistical power considerations.[4] Accordingly, to achieve a desired level of statistical power, projections can be made regarding how large a sample is required, how many subjects can be made available in each site, and, accordingly, how many sites would be required. This is an important advantage over single-site studies, each one of which is limited by the constraints of the site.

Also, unlike single-site studies, multi-site studies can provide for replication, if that is the intent. Multi-site studies that include the replication aim can use identical or near identical procedures at various sites thus minimizing possible effects of procedure variation on outcomes. Furthermore, when designed so that the research is conducted concurrently at the various sites, confounding temporal differences can be minimized. Of course, if the purpose is replication at multiple sites, efforts can be made to assure that all other site related conditions are similar such as sample characteristics. In the case of intervention studies multiple site replications would need to include special provisions to assure that the interventions administered in the sites are similar.

Sometimes in multi-site studies the intent is, not to replicate, but rather to examine variations across sample characteristics or procedures so as to facilitate elaboration and specification. In these cases, for example, sites with different populations could be sought so as to study an intervention's effects with populations of different ethnic composition.

Three forms of multi-site research

While there are no classifications of multi-site research that have become generally accepted, Elkin notes that at least three models of multi-site research exist (Elkin, I, 1992). These models reflect varying roles that individual researchers may take relative to central decision making. Elkin describes these models as: (1) centrally designed and coordinated models with sites competitively selected through peer review wherein collaborators are variously involved in decision making; (2) models wherein researchers decide on their own to collaborate, jointly designing a common procedure at the outset; (3) models wherein researchers use a common data base but do not collaborate in implementation of the study.

I find it useful to distinguish among three forms of multi-site research, namely, simple multi-site research; complex multi-site research; and, multi-site allocation research. Next I briefly describe each of

4 Statistical power is necessary to detect hypothesized associations between variables when using inferential statistical methods.

these types providing examples drawn from our experiences at the CSSWP as well as from other research conducted in the United States. Of necessity these descriptions are brief, highlighting multi-site dimensions only. Furthermore, no attempt is made to present study progress or findings. The interested reader is referred to the cited references for further information and findings pertaining to each study.

Simple multi-site research

Often the same research questions and procedures are used across sites. I will refer to these as simple multi-site studies. These differ from what I term complex multi-site studies which address different research questions and/or may use different procedures at the sites studied. I provide two examples, the first provides somewhat more detail since it is the first example of multi-site research and it is taken from our experiences at the CSSWP. The second is a well-known multi-site study funded by the U.S. National Institute of Mental Health (NIMH).

From Research to Practice

From Research to Practice is an example of a simple multi-site study recently completed at the CSSWP in collaboration with investigators at the New York State Psychiatric Institute (NYS-PI). It is an example of a simple multi-site study because a single set of research aims, questions, hypotheses and a common procedure were applied at all research sites. A multi-site rather than a single-site study was indicated so as to increase sample size beyond what could be achieved at any one site and so as to increase the geographic and demographic diversity of the patient population sampled.

The study was directed by a Principal Investigator together with two co-investigators at the coordinating unit, namely the NYS-PI. NIMH funding was awarded to the Principal Investigator at the NYS-PI through the New York State Research Foundation for Mental Hygiene.[5] The study was designed by the NYS-PI research team. This team carried ongoing responsibility for project coordination, implementation, data analysis and reporting. The original plan called for subcontracts to a New Jersey site research team as well as to the New York City based CSSWP research team. The New Jersey site was composed of one outpatient mental health clinic. The CSSWP conducted the

5 This report is based on research conducted under subcontract to the Center for the Study of Social Work Practice - CSSWP (New York State Research Foundation for Mental Hygiene contract #SDMHCU00642601). The multi-site study was initially funded by the National Institute of Mental Health Grant #1R01MH052822-01 (02) (03). The Principal Investigator at the New York State Psychiatric Institute was David Shaffer, M.D. and the Co-Investigators were Prudence Fisher, Ph.D. and Christopher Lucas, M.D. The investigators for the Center for the Study of Social Work Practice subcontract were: Principal Investigator Edward J. Mullen, D.S.W.; Co-Investigators Robert Abramovitz, M.D., William Bacon, Ph.D. and Bruce Grellong, Ph.D. Prior investigators included Helene Jackson, Ph.D. and Jennifer Magnabosco, Ph.D.

research for the New York City sites. The New York City sites included eight community-based outpatient mental health clinics operated by the Jewish Board of Family and Children's Services (JBFCS). These clinics are located in four of the five boroughs of New York City.

Each subcontractor had a designated Principal Investigator and research staff. Each subcontractor carried responsibility for project implementation at their respective sites including human subjects review and approval, procedure administration, and data gathering. Publication of study findings pertaining to a site was the joint responsibility of the NYS-PI Principal Investigator and the individual subcontract site Principal Investigator.

From Research to Practice examined how a computerized, mental health, diagnostic assessment instrument, the Computerized Diagnostic Interview Schedule for Children (C-DISC), originally developed for use in epidemiological research affects clinical practice when used with child and adolescent outpatients. The C-DISC is a lay administered, computerized interview based on the American Psychiatric Association's Diagnostic and Statistical Manual for Mental Disorders, version IV (DSM-IV) and the tenth edition of the World Health Organization's (WHO) International Classification of Diseases (ISD-10). As originally planned, the evaluation was to use an experimental design. In each of the clinics data would be gathered during a prospective period detailing each clinician's normal assessment practice during routine practice prior to any experimentation. Following this prospective period, in each clinic, half of the clinicians were to be randomly assigned to an experimental exposure to the C-DISC, and the other half would continue with routine practice completing a simple data gathering checklist. At the end of a year clinicians in each of the two conditions were to be assigned to the alternate condition in what is called a crossover design (Mullen, 1998; Mullen, et al., in preparation).

NIMH Treatment of Depression Collaborative Research Program (simple - intervention)

The NIMH Treatment of Depression Collaborative Research Program is another example of a simple multi-site study. The study was conducted simultaneously at several research sites (University of Pittsburgh, George Washington University in Washington, D.C. and the University of Oklahoma), addressing a common set of research questions and using a common research procedure. The study involved the collaborative efforts of these research sites, as well as additional training sites, NIMH and the Veterans Administration Data Analysis Facility (Veterans Administration Hospital, Perry Point, Maryland, USA).

The study used an experimental design in which treatments were randomly assigned to subjects. The basic purpose of this NIMH funded collaborative research was to assess the efficacy, efficiency, and safety of two well-defined, short-term psychological approaches, cognitive/behavior therapy and interpersonal psychotherapy, for outpatient treatment of non-bipolar,

non-psychotic depression. These two psychotherapies were compared to a medication treatment previously shown to be effective for this study population. In addition a pill-placebo condition was included. Outcomes measured were symptomatology, general clinical status, and social functioning. These measurements were taken at various points including during treatment, at termination and at several follow-up points (Elkin, 1994).

Complex multi-site research

A multi-site study can have a common set of aims, but within that common set of aims the study can allocate different research questions to different sites. Accordingly, different procedures could be used in different sites, specific to the questions addressed at each site. Such efforts are considered multi-site studies because they are a planned component in a larger research undertaking which is focused on a specific set of aims. I will call these complex multi-site studies, referring to the level of complexity of the research questions and procedures used. Two recent examples of complex multi-site research are Matching Patients to Alcoholism Treatments (Project MATCH) and Mental Health Service Use, Needs, Outcomes, and Costs in Child and Adolescent Populations (UNOCCAP).

Matching Patients to Alcoholism Treatments - Project MATCH

Project MATCH is an example of complex multi-site research. This was a multi-site client-treatment matching trial involving nine geographically diverse clinical research settings and one coordinating center. The study was funded by the United States National Institute on Alcohol Abuse and Alcoholism. The purpose of Project MATCH was to assess the benefits of matching alcohol dependent clients to three different treatments with reference to a variety of client attributes. This is an example of a complex multi-site study since two parallel but independent randomized clinical trials were conducted, one with alcohol dependent clients receiving outpatient therapy in five sites and one with clients receiving aftercare therapy following inpatient or day hospital treatment in four sites. Clients were randomly assigned to one of three twelve-week, manual-guided, individually delivered treatments: Cognitive Behavioral Coping Skills Therapy; Motivational Enhancement Therapy; or, Twelve-Step Facilitation Therapy. Clients were then monitored over a one-year post-treatment period. In addition a three-year follow-up study was conducted. Individual differences in response to treatment were evaluated for ten primary matching variables and sixteen contrasts specified a priori. The primary outcome measures were percent days abstinent and drinks per drinking day (Project MATCH Research Group, 1997).

Mental health service use, need, outcomes, and costs in child and adolescent populations (UNOCCAP)

Another research program was titled

Multi-site Study of Mental Health Service Use, Need, Outcomes, and Costs in Child and Adolescent Populations but referred to as the UNOCCAP study. Although the UNOCCAP research program was terminated following the developmental phase, it serves as a good example of large scale, complex multi-site research. In 1994 the NIMH invited cooperative agreement applications for a five-year study of child and adolescent mental health services. Applications were invited for two types of studies. One type involved multi-site, collaborative, longitudinal, community-level studies of the types and patterns of mental health service use by children and adolescents, the extent of unmet need for services, and the ways in which the organization and financing of services influence access to, use of, and outcomes of mental health services. The other type of study involved a national survey to address issues related to the prevalence and incidence of specific mental disorders among children and adolescents, rates of mental health service utilization across major service sectors, and costs and financing of care. The UNOCCAP initiative encouraged and required the collaboration of multidisciplinary research teams at both the community and national sites; and, it sought to enable independent teams of investigators to work together to develop common study procedures. The National Institutes of Health (NIH) cooperative agreement mechanism (U01) funded this research program. The National Institute of Health (NIMH) staff worked jointly with the awardees in a partnership role, to support, coordinate, and facilitate the awardees' activities, and to assist in moving the study through its phases. Direction and principal responsibility for the conduct and implementation of the study remained with the awardees.

The UNOCCAP participating sites were the Johns Hopkins University, University of California at Los Angeles, University of Chicago, and Washington University, with additional collaboration from the Research Triangle Institute, Rand Corporation, Vanderbilt University, Yale University, and the NIMH staff. The NIMH budgeted $45 million dollars over five years for the project, which also received support from the United States Administration for Children, Youth and Families; the United States National Institute of Child Health and Human Development; the United States Substance Abuse and Mental Health Services Administration; the United States Department of Education; and, the MacArthur Foundation.

UNOCCAP was originally conceptualized as providing both a national probability sample for determining the prevalence and incidence of specific disorders in children, and as a set of community level studies to address the types and patterns of mental health service use by children and adolescents, the extent of unmet need for services, and the financing of these services. During the two-year developmental phase of the UNOCCAP study, the participating researchers designed a nationwide household sample of approximately 10,000 children and adolescents, and made major efforts in instrumentation development. The collaborators also planned to assess additional samples of children in both outpatient, such as specialty mental health

and school-based services, and inpatient/residential services.

The study underwent scientific review in 1997 by an Oversight Board Appointed by the NIMH Director. Based on the Oversight Board's recommendations, the NIMH Director decided not to carry out the UNOCCAP study. I return to this study in my concluding discussion since it illustrates important issues pertaining to the conduct of multi-site research (National Institute of Mental Health, 1998).

Multi-site allocation research

Multi-site allocation research differs sharply from simple and complex multi-site research. In this third form of multi-site study the geographical unit is not used as the location-site for the study, but rather the geographical unit is the object of study. Accordingly, Boruch and Foley describe »sites and other entities, rather than individuals, as the units of allocation, treatment and analysis« in randomized trials (Boruch & Foley, in preparation, abstract). The sites they describe are geographical locations used as units of allocation. Other examples of allocation units are families, communities, and organizations. Their focus is on experimental research in which sites or other units are randomly allocated to differing interventions so as to study causal associations. However, sites could also be studied in non-randomized research as well as in randomized research. I will refer to this form of multi-site study, whether using randomized or non-randomized designs, as multi-site allocation research. Randomized allocation studies are fully discussed by Boruch and Hedges in another article in this issue. Next I provide one example from our work at the CSSWP, namely the Sanctuary study.

Trauma Focused Intervention Targeting Risk for Violence (Sanctuary)

This multi-site intervention study is being conducted by the CSSWP at twelve JBFCS residential treatment programs for children and youth operated in Westchester County, New York. The twelve units are being randomly assigned to either an innovative milieu treatment or to a standard residential treatment condition without the innovative enhancement. This multi-site research is examining the implementation and proximal effects of an intervention designed to reduce trauma-related symptoms of youths that place them at high risk for violent behavior, poor adjustment, and serious mental health difficulties.

The aims will be achieved by using a 2 x 5 design. Two service delivery conditions are provided, namely the experimental Sanctuary Model enhanced milieu treatment or the Standardized Residential Services. Measurement of outcome variables will be taken at five data collection points, namely at baseline, three months, six months, nine months, and 12 months. The twelve residential units have been randomly assigned to either the Sanctuary Model or Standard Residential Services. The twelve residential units serve one-hundred and fifty youths and have ninety-six staff. Implementation and effects of the model will be measured

at the provider level (i.e., perceptions of change in the therapeutic environment, changes in interaction patterns between staff and youths) and at the youth level (i.e., perceptions of change in the therapeutic environment, change in youths' behavior and skills).

This is a collaborative, multi-site study involving the CSSWP, the JBFCS Westchester facilities, the JBFCS Center for Trauma Program Innovation, the Columbia University School of Social Work and the Columbia University New York State Psychiatric Institute. The study Principal Investigator is an affiliate of the CSSWP and a faculty member at the CUSSW. Co-Investigators include members from each of the collaborating organizations. The research is funded by the NIMH (Rivard, 2000).

Two types of multi-site research

In addition to these three forms of multi-site research, two broad types of multi-site research can be specified. The two types of multi-site research are: (1)multi-site intervention research; and, (2) multi-site survey research. These are described next with examples of each provided.

Multi-site intervention research

In multi-site intervention research the aim is to study and draw conclusions about intervention programs. Examples of multi-site intervention research are the previously described From Research to Practice, the NIMH Treatment of Depression Collaborative Study, Project MATCH, and the Sanctuary study. Examples of multi-site survey research are provided next.

Multi-site survey research

A second type of multi-site research has as its aim description of populations, using survey methods. Both probability sampling and non-probability sampling methods can be used. I will call this second type of multi-site research multi-site survey research. In survey research multiple sites are used for sampling to increase representativeness and to increase sample size. This is a well-know application of multi-site research with well-developed methodologies.

Probability sampling in multi-site research

Four examples of multi-site survey research are provided next including three that use probability sampling methods and one that uses non-probability methods. Both simple and complex multi-site research examples are provided. Examples include studies conducted by the CSSWP and as well as through NIMH funded research.

The Patient Profile Study

Multi-site survey research can use probability sampling methods to address research questions. Cluster sampling methods have long been used in survey research. These clusters can be geographi-

cal sites. An example of such a multi-site study is the Patient Profile Study. This is a simple multi-site study using probability sampling survey methods. This survey was conducted by the CSSWP. The geographical sites sampled were a number of JBFCS operated community mental health clinics located in four New York City boroughs. The research aim was to describe the child and adult population of patients who came to these clinics for mental health services. A sample of approximately 20% of adult and child clients were randomly selected from these clinics, stratified on the basis of clinic, ethnicity, and age. Accordingly, findings were specified by site as well as ethnicity and age. Because multiple sites were used conclusions were drawn pertaining to each clinic as well as to the total JBFCS clinic population (Mattaini, M. A., Grellong, B. A., & Abramovitz, R. (1992).

Epidemiologic Catchment Area Study (ECA)

The Epidemiologic Catchment Area Study (ECA) is a well known example of multi-site survey research. It was a simple multi-site study in the sense that common research questions and procedures were used across sites. The study used multistage probability sampling drawing from five geographical locations. The purpose of the ECA research was to collect data on the prevalence and incidence of mental disorders and on the use of and need for services by the mentally ill. Independent research teams at five universities (Yale University, Johns Hopkins University, Washington University, Duke University, and University of California at Los Angeles), in collaboration with the National Institute for Mental Health, conducted the studies with a core of common questions and sample characteristics. The sites were areas that had previously been designated as Community Mental Health Center catchment areas. Each site sampled over 3,000 community residents and 500 residents of institutions, yielding 20,861 respondents overall. The longitudinal ECA design incorporated two waves of personal interviews administered one year apart and a brief telephone interview in-between (for the household sample).

While the ECA used probability methods to sample within each of five sites, these sites can not be assumed to be representative of the United States population (Robins & Regier, 1991). To address this limitation another NIMH funded study, the National Co-morbidity Survey (NCS), was subsequently conducted drawing a probability sample from the total United States population (Kessler, et al., 1994). Currently, the NCS is being replicated.

Methods for the Epidemiology of Child and Adolescent Mental Disorders Study (MECA)

This NIMH funded study is an example of a complex multi-site survey using probability sampling methods. The MECA collaborative study was conducted to develop methods for surveys of mental disorder and service utilization in unscreened population-based samples of children and adolescents. Probability household samples of youths were selected at four sites and interviews were conducted with a total of 1,285 pairs of youths and their adult caretakers in their homes. Lay interviewers administered

a computer-assisted version of the NIMH Diagnostic Interview Schedule for Children Version 2.3 (DISC) and structured interviews to assess demographic variables, functional impairment, risk factors, service utilization, and barriers to service utilization. More than 7,500 households were enumerated at four sites. Since sites varied in terms of procedures used and research questions examined this can be considered a complex multi-site survey.

Non-Probability sampling in multi-site research

Non-probability survey research methods can also be used in an attempt to describe a population. Multiple sites frequently serve as the source for such research. An example from our work at the CSSWP is provided next.

Odyssey Project

The Odyssey Project is an example of a simple multi-site survey using non-probability sampling methods. The Odyssey Project is a collaborative multi-site study in which the CSSWP is a participant. The Odyssey Project is a descriptive and prospective study of children in residential group care, group homes, and therapeutic foster care in the United States. The Child Welfare League of America, in cooperation with and support of its members, is conducting this national, multi-site, descriptive and prospective study of children and youths in residential care. The purpose of the descriptive phase is to determine what types of settings and services are serving what kinds of children and youths. The purpose of the prospective phase is to determine what types of settings and services are related to what outcomes with what kinds of children and youths. The research questions were designed to explore the impact of residential group care, group homes, and therapeutic foster care on children and youths in care. Children who enter care between April, 1995 and July, 1999 were eligible to participate in the project. Twenty-eight agencies from 15 states participated providing approximately 3,100 children and youths to the study sample. Many of these participating agencies included multiple facilities so that the actual number of distinct sites was larger than 28. Accordingly, in this example multiple facilities (sites) are nested within agencies which in turn are nested within the larger CWLA study.

Oftentimes multi-site studies are referred to as »collaborative« in nature. This is the case with the Odyssey Project. As collaborating partners in the research, member agencies had clearly defined rights and responsibilities.[6]

Issues in multi-site studies

In identification of issues in multi-site studies I draw from several sources. At the outset I wish to give full attribution to these sources for the insights provided. I have been struck by the common observations about issues in multi-site research found in these sources. Elkin's paper presents her personal reflections on experiences with the previously described NIMH Treatment of Depression Collaborative

Research Program (Elkin, 1992). The papers authored by members of the previously described Project MATCH research team identify issues in the context of that multi-site study (Carrol, et al., 1998; Del Boca, et al., 1994; Fuller, et al., 1994; Zweben, et al., 1994). The observations of the previously described UNOCCAP Oversight Board are based on a careful review of early experiences with that research program (National Institute of Mental Health, 1998). Finally, as Director of the CSSWP I have drawn upon my own experiences with the many multi-site studies conducted by CSSWP (Practice & Research, Spring 2000). Given the limits of space available in this article I will only outline a number of issues.

Advantages of Multi-site Research

The advantages of multi-site research have already been mentioned and they are clear. The ability to draw subjects from multiple sites can increase sample size. Accordingly, a study's statistical power is increased. Also, multiple sites can increase the sample's diversity on variables of interest. An additional advantage of multi-site research pertains to replication. Properly designed multi-site research can provide for simultaneous replication limited only by the number of comparable sites included in the study. In addition to these advantages Elkin notes two additional benefits. One pertains to the potential for increasing resources, both financial and in terms of expertise. The potential availability of multi-site team members with a range of methodological, statistical and clinical expertise can be invaluable. Associated with this is the benefit of collaborative decision making, resulting in enriched discussion and improved outcomes. As noted by Fuller, et al., in multi-site research there is also the potential to create a centralized unit with resources for core functions such as data analysis.

Disadvantages of Multi-site research

If not properly managed there are a number of potential disadvantages associated with multi-site research. Elkin, Fuller, et al, Del Boca, and Zweben describe potential disadvantages:

- Multi-site studies can be costly and methodologically complex.
- The process of joint decision making, consideration of procedural and methodological problems and joint resolution of these problems requires considerable time and effort.
- Idiosyncratic site effects can create problems in management as well as in data analysis and interpretation.
- Logistic difficulties can be problematic when implementation spans across multiple sites. Examples include allocation of staff across sites, communication among geographically dispersed team members, and so forth.
- Statistical issues can be problematic such as how best to combine data collected from multiple sites. Site effects need to be addressed in the statistical analysis.
- Addressing new requirements after

the study has begun can be troublesome. For instance sites can experience policy and fiscal changes that may impact on the study requiring procedural adjustments. When there are many sites these changes can be difficult to track in a timely way.

- Issues of reliability of data collection across sites are often problematic. Sites may implement data collection differently creating difficulties in maintaining uniform procedures.
- Issues of subject recruitment and eligibility criteria in multi-site studies can be complex. Again, these issues can arise from a lack of uniformity among the sites as well as from the logistical difficulties of monitoring recruitment practices among many sites.
- Redundancy of staffing across sites can increase cost unnecessarily. Careful planning is required to assure that staff is used efficiently across sites.
- There can be problems establishing and maintaining cooperation and commitment across sites since no one site »owns« or is totally responsible for the research. Accordingly, in many multi-site studies individual sites have limited ownership and get limited credit.
- Multi-site studies oftentimes experience difficulty maintaining cross-site uniformity of procedures. Maintaining consistency across sites can be problematic (e.g., similarity of treatment implementation across sites, of data collection, etc). There is greater difficulty in maintaining integrity of the treatment and research procedures across sites in multiple locales.
- Mechanisms need to be established for how the data will be analyzed and the findings published that are protective of individual site interests as well as of the multi-site collaborative effort as a whole.

References

Allen, J., Anton, R.F., Babor, T.F., Carbonari, J., Carroll, K.M., Connors, G.J., Cooney, N.L., Del Boca, F.K., DiClemente, C.C., Donovan, D., Kadden, R.M., Litt, M., Longabaugh, R., Mattson, M., Miller, W.R., Randall, C.L., Rounsaville, B.J., Rychtarik, R.G., Stout, R.L., Tonigan, J.S., Wirtz, P.W., & Zweben, A. (1998). Matching alcoholism treatments to client heterogeneity: Project MATCH three-year drinking outcomes. Alcohol Clin Exp Res, 22 (6): 1300-1311 Sep 1998.

Allen, J.P., Mattson, M.E., Miller, W.R., Tonigan, J.S., Connors, G.J., Rychtarik, R.G., Randall, C.L., Anton, R.F., Kadden, R.M., Litt, M., Cooney, N.L., DiClemente, C.C., Carbonari, J., Zweben, A., Longabaugh, R.H., Stout, R.L., Donovan, D., Babor, T.F., Del Boca, F.K., Rounsaville, B.J., Carroll, K.M., Wirtz, P.W., Bailey, S., Brady, K., Cisler, R., Hester, R.K., Kivlahan, D.R., Nirenberg, T.D., Pate, L.A., Sturgis, E., Muenz, L., Cushman, P., Finney, J., Hingson, R., Klett, J., & Townsend, M. (1997). Matching alcoholism treatments to client heterogeneity: Project MATCH posttreatment drinking outcomes. Journal of Studies on Alcohol, 58 (1): 7-29 Jan 1997.

American Psychiatric Association (1994). Diagnostic and statistical manual of mental disorders. Washington, DC: Author.

Anon (1998). Matching alcoholism treatments to client heterogeneity: Treatment main effects and matching effects on drinking during treatment. J Stud Alcohol, 59 (6): 631-639 Nov 1998.

Bacon, W.F. (2000). From research to practice: The C-DISC in clinical services. Practice & Research, Spring 2000, 26-28.

Blazer, D., D. Huges, & George, L.K. (1987). The epidemiology of depression in an elderly community. The Gerontologist: 27 (1987), 281-287.

Boruch, R. & Foley, E. (in preparation). The honestly experimental society: Sites and other entities as the units of allocation and analysis in randomized trials. In Bickman, L. (ed.). Validity and social experimentation: Donald T. Campbell's legacy. Thousand Oaks, Ca: Sage.

Carroll, K.M., Connors, G.J., Cooney, N.L., DiClemente, C.C., Donovan, D.M., Kadden, R.R., Longabaugh, R.L., Rounsaville, B.J., Wirtz, P.W., & Zweben, A. (1998). Internal validity of project MATCH treatments: Discriminability and integrity. Journal of Consulting and Clinical Psychology: 66 (2): 290-303 Apr 1998.

Del Boca, F.K., Babor, T.F., McRee, B., (1994). Reliabiltiy enhancement and estimation in multi-site clinical-trials. J Stud Alcohol: 130-136 Suppl. 12 Dec 1994.

Eaton, W.W., et al. (1985). The epidemiologic catchment area program of the National Institute of Mental Health. Public Health Reports: 96 (1981), 319-325.

Eaton, W.W., & Kessler, L. (eds) (1985). Epidemiology field methods in psychiatry: The NIMH epidemiologic catchment area program. Orlando, FL: Academic Press, Inc., 1985.

Elkin, I. (1992). Multi-site studies: Advantages and disadvantages. In Gerin, P & Dazord, A., Recherches Cliniques "Planifees" sur Les Psychotherapies: Methodolgie. Paris: INSERM, 1992.

Elkin, I. (1994). The NIMH treatment of depression collaborative research program: Where we began and where we are. In Bergin, A.E. & Garfield, S.L., Handbook of psychotherapy and behavior change. New York: John Wiley and Sons, Inc.

Fuller, R.K., Mattson, M.E., Allen, J.P., et al. (1994). Multi-site clinical-trials in alcoholism-treatment research – organizational, methodological and management issues. J Stud Alcohol: 30-37 Suppl. 12 Dec 1994 .

Guterman, N.B. (2000). The Odyssey project and community violence exposure among children & youth living in residential treatment settings. Practice & Research, Spring 2000, 23-25.

Journal of the American Academy of Child and Adolescent Psychiatry (1996). Special Section on Epidemiology of Child and Adolescent Mental Disorders. 35:7.

Kessler, R.C., McGonagle, K.A., Zhoa, S., Nelson, C.B., Hughes, M., Eshleman, S., Wittchen, H. & Kendler, K.S. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Archives of General Psychiatry, 51, 8-19.

Lahey, B.B., Flagg, E.W., Bird, H.R., Schwab-Stone, M.E., Canino, G., Dulcan, M.K., Leaf, P.J., Davies, M., Brogan, D., Bourdon, K., Horwitz, S.M., Rubio Stipec, M., Freeman, D.H., Lichtman, J.H., Shaffer, D., Goodman, S.H., Narrow, W.E., Weissman, M.M., Kandel, D.B., Jensen, P.S., Richters, J.E., & Regier, D.A. (1996). The NIMH Methods for the Epidemiology of Child and Adolescent Mental Disorders (MECA) study: Background and methodology. Journal of the American Academy of Child and Adolescent Psychiatry: 35 (7): 855-864 Jul 1996.

Mattaini, M. A., Grellong, B. A., & Abramovitz, R. (1992). The Clientele of a Child and Family Mental-Health Agency - Empirically Derived Household Clusters and Practice Implications. Research on Social Work Practice, 2(3), 380-404.

Mullen, E.J. (in preparation). From research to practice: A preliminary report of the C-DISC on dimensions of clinician, patient and caretaker satisfaction (presented as an unpublished paper under title of: "Using assessment instruments in social work practice«. October

7-8, 1999, Inter-Centre Consortium of Social Work Research Centers, Stockholm, Sweden). In preparation as an on-line electronic pre-print to be available: London, UK: National Institute for Social Work, http://www.intsoceval.net/Stockholm/stockholmpapers.htm .

Mullen, E.J. (2001). Outcome measurement in social work: health and mental health. (plenary session, 3rd International Conference on Social Work in Health and Mental Health, July 1-5, 2001, Tampere, Finland; [On-line electronic pre-print]. Available: London, UK: National Institute for Social Work, http://www.intsoceval.net/Utrecht/utrechtpapers.htm.

Mullen, E.J. (1998). Linking the university and the social agency in collaborative evaluation research: Principles and examples. Scandinavian Journal of Social Welfare. nr2.

Mullen, E.J. and Magnabosco, J. (eds) (1997). Outcomes measurement in the human services. Washington:D.C.: National Association of Social Workers.

National Institute of Mental Health (1994). Multi-site study of mental health service use, need, outcmes, and costs in child and adolescent populations. NIH Guide: vol 23, number 15, April 15, 1994 RFA.

National Institute of Mental Health (1998): Report from the UNOCCAP oversight board to the National Advisory Mental Health Council: Charting the mental health status and service needs of children: Recommendations from the UNOCCAP oversight board.

Practice & Research: Journal of the Center for the Study of Social Work Practice (Spring 2000).

Project MATCH Research Group (1997). Matching alcoholism treatments to client heterogeneity: Project MATCH posttreatment drinking outcomes. Journal of Studies on Alcohol 58: 1, 7-29. (also see:, Journal of Studies on Alcohol, Supplement No. 12, December 1994.)

Regier, D. A. (1994). ECA contributions to national policy and further research. International Journal of Methods in Psychiatric Research. 4, 73-80.

Rivard, J.C. (2000).Evaluating the implementation and impact of an intervention designed to enhance residential treatment for adolescents with histories of trauma. Practice & Research, Spring 2000, 33-36.

Robins, L. N., & Regier, D. A. (Eds.) (1991). Psychiatric disorders in America: The Epidemiologic Catchment Area Study. New York: The Free Press.

Zweben, A., Donovan, D.M., Randall, C.L., et al. (1994). Issues in the development of subject recruitment strategies and eligibility criteria in multi-site trials of matching. J Stud Alcohol: 62-69 Suppl. 12 DEC 1994.

Summary and Conclusions

This article has examined multi-site research. Differing from single-site research, multi-site research is appropriate for reasons that distinguish it from single site research. In this article I have examined forms and types of multi-site research to illustrate a variety of applications. I have described some examples of multi-site research at the national level with reference to mental health services and alcoholism treatment research applications. Drawing from my experiences as director of research at a social work research center I have described additional applications in mental health, child welfare, and social work practice. I have outlined some advantages and disadvantages of multi-site research. In concluding this analysis I provide suggestions for those considering engaging in multi-site research. While

there are many matters to consider I limit my comments in this concluding section to three areas that strike me as most often neglected and of particular importance to successful multi-site research. Since multi-site research can be expensive, involving relatively more resources and time than single-site research, it should not be undertaken unless preconditions are met. Several of these preconditions are described next.

Concepts and methods previously developed and tested

An important precondition is that preliminary studies should have shown that key research questions are conceptually sound and that the research methods needed to examine these questions have been adequately developed. For example, the ambitious UNOCCAP research program was undertaken after extensive methodological work had been done in the MECA study (described above). Nevertheless, the UNOCCAP research program was halted by NIMH because it was determined that the original UNOCCAP aims and research questions were too broad, unevenly developed, and too extensive for any one study to address. The NIMH Review Board concluded:

»--- no one study can address each of these questions well simultaneously. There is no way to combine these questions into one method or design without sacrificing the quality of data. --- Some areas require additional conceptualization, better instrumentation, or more empirical work to generate or test hypotheses. This critical basic research must be conducted to lay the groundwork for answering all of these questions. --- More important, science is not ready for such an approach. Quite simply, the necessary conceptualization, tools, and designs are not uniformly available in all substantive areas. If resources were diverted into prematurely conducting such an immense and elaborate effort, the Board is concerned that it would generate data that policymakers should not use and that a substantial portion of the researchers would not accept as credible.« (National Institute of Mental Health, 1998, 2-4)

In addition to the lessons learned from the UNOCCAP experience additional lessons can be learned from the above described From Research to Practice study. In that study unusual problems were encountered securing the cooperation of sites as well as in engaging adequate numbers of clinicians and patients. Although during the planning stage the site central administration was supportive of the study, problems pertaining to motivation and commitment were encountered during study implementation at the local level. Also, while preliminary analysis indicated that sites had sufficient numbers of clinicians and clients, during implementation it was difficult to engage adequate numbers of both. In this example it could be argued that a multi-site study was premature. Rather, pilot testing the intervention and the research methodology at one site may have been prudent before launching an ambitious multi-site investigation.

Resources in place

A second precondition for multi-site research is that necessary resources should be secured before undertaking multi-site studies. Multi-site studies require adequate resources to support a central core as well as to support site-specific activities. The central core can carry out common functions such as data processing which can relieve sites of this responsibility. Basic funding and adequate personnel are required to make sure that all program and research functions can be properly implemented throughout the life of the study. However, often overlooked in planning of multi-site studies is the necessity of providing financial resources and qualified staff for the »relationship« side of multi-site research. Ongoing, frequent communication among all parties is essential so that commitment, motivation, and day-to-day problem solving can occur smoothly. While necessary in single-site research this aspect of multi-site research is critically important (Mullen, 1998).

Governance clarified and agreed to by collaborators

A third precondition is that a clearly defined and agreed-to governance and management structure and process be developed before implementation of multi-site research. Because of the complexity of such studies and because of the unusual nature of study »ownership« the rules and procedures should be clear from the outset. Differential responsibilities of the central coordinating team, the site teams, practitioners and researchers, program administrators, the funders, and other partners need to be specified and agreed-to. Since there are different models for multi-site governance consideration should be given to which model best fits the circumstances of a given multi-site study.

If these (and others not addressed) conditions are met then multi-site research should be considered if it is determined that single-site research cannot provide an adequate sample size, the desired sample heterogeneity, or an adequate basis for drawing conclusions regarding a population of interest. Multi-site research should also be considered when concurrent replication is desired. As discussed elsewhere in this issue by Boruch, et al. when geographical sites or other complex units are the units of study, allocation multi-site research should be considered.

Finally, it is important for the success and development of future generations of multi-site research that a solid and detailed literature be generated regarding the planning, implementation, analysis and reporting of multi-site research. Researchers engaged in multi-site studies need to take responsibility for contributing to this literature by reporting not only multi-site findings but also what they have learned about the conduct of multi-site research. Currently, with the exception of survey research, little has been written about the methodologies of multi-site research as applied to social interventions.

Meta-Analysis and Program Outcome Evaluation

mark w. lipsey

Meta-analysis is a technique for statistically representing and analyzing the findings from a set of empirical research studies. In application to program evaluation research, it provides a means for systematically synthesizing knowledge about the characteristic and outcomes of effective programs. Six lessons learned from meta-analysis of evaluation research illustrate the application and findings of this approach: (1) many social programs are more effective than generally realized; (2) individual evaluations can easily produce erroneous results; (3) the methods used in an evaluation play a large role in the program effects found in the evaluation; (4) program effectiveness is a function of identifiable program characteristics; (5) there is much room for program improvement; and (6) the most credible evidence about program effects comes through integration of multiple evaluation studies.

Introduction

Evaluation provides an assessment of how a particular social program is performing in the context of its mission and the expectations of its stakeholders. However, when designing a new program or reforming an existing one, the responsibility of the evaluation field is to provide evidence about what program approaches have proven most effective in prior evaluation studies. To accomplish this task, evaluators must be able to learn from prior studies what kinds of interventions work for what purposes under what conditions. This, in turn, requires that at least some researchers in the evaluation field systematically gather and integrate the evaluation findings for a

Mark W. Lipsey is Professor of Public Policy at Vanderbilt University and he serves as Director of the Center for Evaluation Research and Methodology at the Vanderbilt Institute for Public Policy Studies.

wide range of programs and program variations. As a fringe benefit, such endeavors also provide opportunity to examine the methods evaluators use and how they relate to the results generated by those methods so that the evaluation field may learn how to improve its methodology.

The central issue raised here is one of generalization how to go from the particulars of individual program evaluations to a broader understanding of the differential effectiveness of different programs for different social problems (Cook, 2000). Valid generalization is the means by which we are able to derive evidence-based principles about what characterizes more and less effective programs. A well developed set of such principles in a given program area is a critical tool for designing, improving, and understanding effective interventions.

Unlike more academic social science fields, where research literature reviews and other forms of knowledge synthesis are commonplace, relatively little attention has been paid to systematic synthesis in the evaluation field. This is primarily due to the nature of evaluation research itself, not because such synthesis is useless. By their nature, evaluations tend to focus on the program under scrutiny and develop in ways that are tailored to the particulars of that program, the concerns of its stakeholders, and the specific purposes of the evaluation. When the findings are analyzed and reported, little or no effort is typically devoted to consideration of the generalizability of the results, how they might apply to other program situations, what has been learned that would be of interest to those who have not yet embarked on a program of that type, and so forth. As applied research, evaluation is organized around application to a specific program context and, correspondingly, evaluators, upon finishing one such project, generally move on to the next without much concern for extracting and reporting the broader lessons of the project for others in the field.

An especially interesting and important area in which the evaluation field would benefit greatly from systematic synthesis of the nature and findings of prior evaluation studies is with regard to outcome evaluation. For most programs, having the intended ameliorative effects on the target problem they address is of paramount political and practical concern. For purposes of program planning and improvement, however, it is of equal importance to know what kinds of programs have meaningful effects on such problems and, among those, which are most effective. More specifically, we might want to know which characteristics of the programs, the target populations, and the evaluation methods are associated with findings of larger and smaller program effects on major outcome variables.

What is Meta-Analysis?

Outcome evaluation is generally conducted using experimental or quasi-experimental research designs with quantitative outcome measures and results that are reported in statistical terms. For research of this type, the technique known as meta-analysis is especially well suited to the task of synthesizing the findings of multiple studies (Cooper, 1998; Cooper

& Hedges, 1994; Lipsey & Wilson, 2001). Meta-analysis revolves around a statistic called an effect size that represents the findings about the program effect on an outcome variable as estimated, for instance, from a comparison between outcomes for a sample of program participants and those for a control sample that does not receive services. The most commonly used effect size statistic for representing the results of intervention research is the standardized mean difference, defined as the difference between the mean value on an outcome variable for the treated group and that for the control group, divided by the pooled standard deviation of the two samples. Division by the standard deviation standardizes the effect size so that, no matter what the original units of the outcome measure, the effect size represents it in standard deviation units. An effect size of .50, for example, indicates that the outcome for the program group on a particular measure was one-half a standard deviation better than that for the control group, irrespective of the measurement scale actually used. Suppose one evaluation study measures depression outcomes on the Beck Depression Inventory and finds that the mean score for the treated group is .40 standard deviations lower (better) than that for the control group. Another study of similar treatment might measure the depression outcome on the Hamilton Depression Scale and find a difference equivalent to .25 standard deviations between the treatment and control group means. We could then compare these, noting that the first study showed a larger effect of treatment on depression. Also, if we wished, we could combine these effect sizes with similarly expressed depression outcomes from many more evaluations of treatment effects into a data set that would allow us to assess the distribution of outcomes, their overall mean, which types of interventions produced the largest effects on depression and which the smallest, and so forth. At this point, we are doing a meta-analysis.

Other types of effect sizes are also used in meta-analysis to represent the outcomes of different studies in a common metric. When the outcome variable is binary, e.g., sick or well, dead or alive, housed or homeless, and so forth, a useful effect size statistic is the odds ratio-- the odds of someone in the program group having the favorable outcome divided by the odds of someone in the control group having that outcome (Haddock, Rindskopf, & Shadish, 1998). Thus an odds ratio of 1.5 means that the odds of a good outcome in the sample receiving service were one and a half times as great as the odds of a good outcome in the control group. Odds ratios are widely used as effect size statistics for representing the outcomes of biomedical interventions and appear frequently in evaluations of medical treatments.

A synthesis of evaluation results using meta-analysis techniques involves computing an effect size for every outcome variable of interest for a collection of evaluations involving the same or similar interventions. These effect sizes are best referred to as the observed effects of the interventions, that is, the effects observed using the measures and methods applied in the evaluation research. Other information

about the nature and circumstances of the intervention, the characteristics of the persons receiving the interventions, the study methods and procedures, and the like are also usually coded for a meta-analysis. All of this information for all the studies included in a meta-analysis is then organized into a database that permits statistical analysis of the distribution of observed effects resulting from those studies.

The typical statistical analysis of a meta-analytic database would first sort the effect sizes according to the type of outcome variables they represent. For example, if the evaluation studies included in the meta-analysis assessed the effects of family therapy on such outcomes as marital satisfaction, quality of communication, and childrens problem behavior, the effect sizes for each of these outcome categories would be analyzed separately. Then, the mean effect size across all the studies would be calculated for each outcome, then the variation of the effect sizes around that mean would be assessed. If the variance of the effect sizes was no larger than expected from the sampling error associated with the samples of persons for whom outcomes were measured in the studies, the mean effect size would provide a good summary of the intervention effect. Because this effect size mean averages over whatever number of studies are included in the meta-analysis, it provides a more representative estimate of the effect of the particular type of intervention on the outcome represented in the effect sizes than estimates derived from any one outcome study.

Frequently, however, the effect sizes from different studies show more variation than likely to result from subject-level sampling error. In that situation, the task of the meta-analysis is to determine if there are systematic relationships between the characteristics of the different studies and the effect sizes they produce. The observed effects of a set of intervention studies can be viewed generally as a function of the efficacy of the treatment, the characteristics of the samples receiving treatment, the methods used to study the effects, and some amount of statistical noise. One useful way of summarizing the information generated by a meta-analysis is to depict the proportion of the variation in the observed effects that is associated with each of these different aspects of the evaluation situation. Further examination can than be made of the specific characteristics of the interventions, treatment recipients, and methods that are most closely associated with larger and smaller observed effects. The results of this process provide the evidence on which we can support useful generalizations about which treatments are most effective on which outcomes for which types of recipients.

Lessons from Meta-Analysis

Meta-analysis has been widely applied to outcome evaluation findings since the pioneering work of Smith and Glass (1977). Though in many ways still not fully developed, it has already generated important lessons about social programs and the methods evaluators use to study them. To illustrate the nature and results

of meta-analysis, and its potential for further enhancing the field of program evaluation, we will describe six lessons we have learned from meta-analysis. The findings that support these insights derive to greater or lesser extent from the work of many meta-analysts. We will not attempt to review the relevant meta-analysis literature here, however. Instead, we will simply summarize what we view as the significant conclusions to be drawn using examples conveniently at hand from our own work over the last decade.

1. Many Social Programs Are More Effective Than We Thought.

One of the troublesome facts of outcome evaluation is that it often finds no significant effects produced by the social programs assessed. It is not unusual for the results of outcome evaluation to be so weak that we cannot be confident the program has meaningful impact. What Rossi and Wright (1984) once called the parade of null results in evaluation can lead to the pessimistic conclusion that nothing works in the world of social programs. The usual basis for such conclusions is a body of outcome evaluations using experimental and quasi-experimental designs that show relatively few statistically significant positive effects on the outcome variables of greatest interest.

One of the distinctive characteristics of meta-analysis is that it focuses on the magnitude of the effects observed in each study, not their statistical significance. Moreover, by combining these magnitude estimates from numerous outcome evaluations, it can reveal the actual distribution of effect sizes that characterize a certain type of intervention. When this is done, it often becomes evident that many of the program effects observed in the original evaluation studies are larger and more consistently positive than they appeared when only those reaching statistical significance were counted. The reason for this, in brief, is that statistical significance is influenced by both the magnitude of an intervention effect and the size of the sample upon which it is measured (Cohen, 1988). Thus effects large enough to be of practical significance may, and in evaluation often do, fall short of statistical significance in an individual evaluation study because the research is conducted with small samples and correspondingly low statistical power.

It is relatively easy to demonstrate the different, and more positive, image of program effects that is revealed by meta-analysis in contrast to the vote-counting approach of assessing the proportion of effects that are statistically significant. Lipsey and Wilson (1993), for instance, assembled all the meta-analyses of the effects of psychological, educational, and behavioral interventions that could be located at the time, more than 300. Many of these were conducted in program areas marked by a history of controversy over whether the interventions produced any positive effects. However, when examined, the distribution of mean effect sizes across this wide range of interventions, and the hundreds of studies and thousands of participants included in the studies meta-analyzed, revealed that the vast majority

of outcome effects were positive and of nontrivial magnitude.

Figure 1.
Distribution of Mean Effect Sizes for 302 Meta-Analyses of the Effects of Psychological, Educational, and Behavioral Interventions

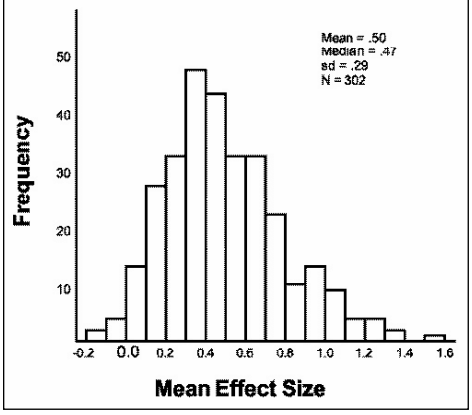


Figure 1 shows the summary distribution of mean effect sizes from all those meta-analyses. The vast majority of the meta-analyses found positive effects on the outcomes of interest (mean effect sizes greater than zero) and the average over these means was about .50. That is, on average across all the interventions represented, the outcomes for the individuals receiving program services were about a half standard deviation better on whatever scale was used for measurement than the outcomes for those in the control conditions who did not receive the program. To put this into perspective, suppose that, on their own, 50% of the individuals in the control group would end up with acceptable outcomes. An effect size of .50 means that, by comparison, nearly 70% of those in the program group would have acceptable outcomes. In many program areas, even smaller effects than this would be of great practical significance.

These meta-analysis results do not mean that all social interventions have positive effects, of course. Nevertheless, they do indicate that to reach any generalization about program effectiveness we should analyze the actual quantitative effect size estimates generated by the available outcome evaluations. The obvious wisdom of this approach, operationalized in meta-analysis techniques, reveals the full range of evaluation findings, and that often proves to represent a wider and more positive set of outcomes than otherwise evident.

2. Individual Outcome Evaluations Can Easily Produce Erroneous Results

The situation described above, in which many outcome evaluations show positive effects, and sometimes relatively large ones, that nonetheless fall short of conventional levels of statistical significance has sobering implications for the design of individual outcome evaluations. By examining the effect sizes over a number of evaluations, and thus in essence combining all their individual study samples, meta-analysis can focus directly on the distribution of observed effect sizes without much consideration of whether each is statistically significant. What we see when we do this, however, is that many of the individual evaluation studies do not show statistically significant effects, even when the meta-analysis reveals that the actual magnitude of the effects for that intervention are

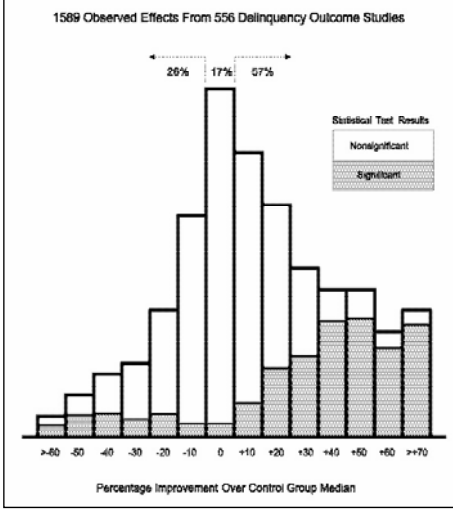
generally positive. In other words, the estimates of the effect sizes for key outcomes in individual studies yield positive values, but fall short of statistical significance and thus cannot be confidently identified as beneficial program impacts within the context of an individual evaluation study.

As noted earlier, this can easily happen when the sample size used in the evaluation design is too small to provide sufficient statistical power for attaining statistical significance even when the effect estimates are of meaningful size. Meta-analysis has revealed that insufficient statistical power is quite common in evaluation research (Lipsey, 2000). An underpowered evaluation design applied to an effective program will usually yield findings that fall short of statistical significance and thus commit what is called Type II error, failing to reject the null hypothesis (of no effect) when, in fact, it is false. From a scientific perspective, effects that fall short of statistical significance in an individual study for whatever reason have little credibility. By definition, they have an unacceptably high likelihood of being spurious, that is, representing statistical error rather than actual intervention effects.

Technically, failure to attain statistical significance in an underpowered outcome evaluation means only that the research has failed to reject the null hypothesis of no effects, not that it has confirmed the absence of effects. However, this is a subtle distinction easily lost on policy makers, program stakeholders, and many researchers. Statistically nonsignificant results are widely interpreted as indications that the program is not effective, with the associated political and practical implications. In this regard, the program is blamed for failing when it is the evaluation research that has failed to use a design with sufficient statistical power to find meaningful effects when they are there to be detected.

The relationship between observed effect sizes, as computed in a meta-analysis, and the statistical significance of those effect sizes found in the individual evaluation studies included in a meta-analysis is illustrated in Figure 2. That figure shows the distribution of effect sizes on all outcome variables reported in over 500 evaluation studies intervention programs for juvenile delinquents. For ease of interpretation, the effect sizes are represented in terms of the percentage improvement shown by the treatment group relative

Figure 2.
Reported Statistical Significance for Different Effect Sizes Observed in Evaluation Studies of Program for Juvenile Delinquents



to the control group median. Thus +30 means that, on whatever outcome variable was measured, the treated juveniles showed a 30% improvement compared to the control group. As can be seen, over half of the observed effect sizes are positive (greater than zero) and many are relatively large (e.g., representing 20% and greater improvement with treatment. Overall, there is little doubt that the interventions evaluated in these studies had positive effects on a majority of the outcomes measured (thought note that 17% of the outcomes were zero and about one-fourth were negative; that is, the control groups did better).

Within each effect size range, Figure 2 shows the proportion of effects found statistically significant in the original evaluation studies. Because the sample sizes in these evaluation studies tended to be modest (a median of about 60 each in the intervention and control groups), they do not have a great deal of statistical power. Figure 2 shows that the majority of the positive effects were not found statistically significant in the individual studies until they were out in the range where treated juveniles were showing improvements of 40% or more compared to those in the control groups. In practical terms, meaningful effects occur below this level, of course. Many programs would be pleased with a 10-20% improvement among the juveniles they served. Moreover, the many positive effects in that range are quite evident in the meta-analysis. But, as can be seen, the individual evaluation studies had a diminishing likelihood of detecting them at a statistically significant level as they got smaller.

It is interesting to note that a similar pattern appeared on the negative end of the continuum. Effects for treated juveniles had to be 50% worse than for control juveniles, or more, before the majority was statistically significant. The decreased proportions of statistically significant results in the negative direction compared with the positive direction that is evident in Figure 2 also represents a problem of small sample sizes. The samples on which negative effects were found tended to be especially small, raising the possibility that, even among those found significant, many may represent no more than sampling error.

The practical limitations imposed on outcome evaluation in field settings is such that it is often quite difficult to enroll samples large enough to ensure a high degree of statistical power. Given the substantial role of statistical noise in such research that has been demonstrated by meta-analysis, outcome evaluation on individual programs can easily fail to attain statistical significance for what are, nonetheless, meaningful program effects. It follows that the results of such evaluation, taken alone, may be misleading. One important contribution meta-analysis can make to this situation is to provide a context of results from other similar interventions within which to interpret the potentially ambiguous findings of an individual outcome evaluation. For example, effect sizes from an outcome evaluation could be compared to the distribution of effects found in a relevant meta-analysis. Their magnitude relative to those found in similar programs could then be assessed as a supplement to assessment by statistical significance testing.

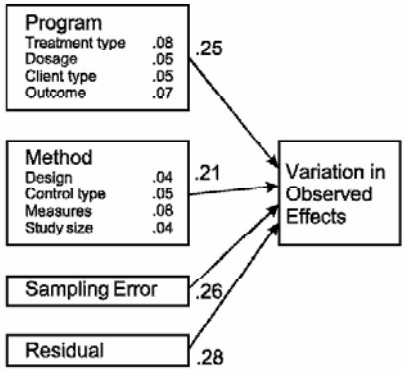
3. Method Matters

Ideally, the experimental and quasi-experimental research designs and procedures typically used for outcome evaluation would generate estimates of actual program effects that were relatively undistorted by the methods themselves. For example, we expect random assignment experiments to produce unbiased estimates of intervention effects, but it is not always possible to use such designs in practical outcome evaluations. It would be comforting to know that a range of more manageable nonrandomized designs would provide results reasonably similar to those from a randomized design. Similarly, when there are various reasonable ways to measure an outcome variable, it would be desirable for them to yield comparable results when applied to the same intervention.

One of the advantages of meta-analysis is that it can investigate the extent to which variation across studies in the methods and procedures of outcome evaluations are related to the effects those studies find. A simple approach is to assess the proportionate variation in observed effects that is associated with the methodological characteristics of the studies in contrast to that associated with such substantive aspects of the programs as the characteristics of the participants, the type of intervention, and the amount of treatment. If most of the effect size variation is associated with differences across studies in program-related characteristics, it is a good indication that the observed outcomes indeed mostly convey information about actual program effects. If, on the other hand, a very large portion of the effect size variation is associated with methodological differences among the studies, it tells us that the outcomes found in those studies may be heavily influenced by the manner in which the program was studied rather than the outcomes it actually produced.

When we have analyzed effect size variation this way with large meta-analytic data sets, we have been dismayed to find that about as much effect size variation is associated with methodological differences among studies as with program characteristics. Summarized over the 300 meta-analyses of psychological, educational, and behavioral interventions we mentioned earlier, for instance, we found the pattern of associations shown in Figure 3 (drawn from Lipsey & Wilson, in press). We have already commented on the large role of sampling error, reflecting typically small

Figure 3.
Sources of Between-Study Effect Size Variance Averaged Over 300 Meta-Analyses



Sources of Between-Study Effect Size Variance Averaged Over 300 Meta-Analyses

sample sizes. Comparing program-related and method-related sources of influence on effect sizes, however, Figure 3 shows that the variation in effect sizes associated with the methods used by the evaluators is larger than that associated with the characteristics of the interventions (21% vs 25%).

When different categories of methodological characteristics are broken out, there are additional surprises. Research design, representing mainly random vs. nonrandom assignment to intervention conditions, and, closely related, the type of control group (e.g., »no treatment« vs. placebo) are influential, as would be expected. There is a large methodological literature on the potential biases associated with design factors (e.g., Shadish, Cook, & Campbell, 2002). Aspects of the outcome measures, however, which have received much less attention in the literature on evaluation methods, also appear to have a substantial influence on the observed effect sizes. The measurement features represented in this category include the way in which the outcome constructs are operationalized (e.g., self-report measures, standardized test, official records) and the timing of measurement (e.g., immediately after intervention or lagged some time later).

Further exploration of evaluation studies with meta-analytic techniques should help determine which methods and procedures yield the most valid results and which create so much distortion that they are not appropriate to use in outcome evaluation. What meta-analysis has already demonstrated is that the neutrality of the typical range of methods for outcome evaluation cannot be taken for granted. What we observe as program effects may reflect as much influence from the methods with which the program was studied as the actual effects the program has on its intended beneficiaries.

4. Program Effectiveness is a Function of Identifiable Program Characteristics

Every social program is, in some regards, unique and the assessment of its impact must be tailored to its particular characteristics and situation. Nonetheless, there are commonalities among programs in a given intervention area that allow for generalizations across programs. It is useful for the evaluator to know what characteristics of an intervention tend to be associated with the most positive outcomes. Such information makes it easier to design an effective evaluation by highlighting the aspects of the program on which the evaluation should focus. In addition, for purposes of program design and improvement, identification of the characteristics of effective programs helps define the »best practices« in a particular intervention area that should be emulated.

Meta-analysis provides a probing way to analyze the characteristics of intervention programs that differentiate those which produce larger outcome effects from those producing smaller ones. Because of the relationship between the methods used in evaluation studies and the observed outcomes described above, however, it can be misleading to simply compare the effect sizes for programs with different characteristics. A potentially clearer picture

is provided by using meta-analysis techniques to statistically control for methodological differences between studies so that the program characteristics most closely associated with larger and smaller effects can be disentangled from methodological artifacts.

With such statistical controls, analyses like those shown in Figure 4 for the studies in our delinquency meta-analysis can be conducted. The details of this analysis are described elsewhere (Lipsey, 1992a,b, 1995), but the results demonstrate that there are consistent relationships between the type of program, how well the program services are delivered and implemented, and the outcomes. Figure 4, for instance, shows that different groups of delinquency intervention programs have quite different mean effects on the juveniles' reoffense recidivism. In particular, the more behaviorally oriented, skill-oriented, and multi-service programs tend to have larger effects.

Figure 4.
Mean Reoffense Recidivism Effect Sizes for Different Groups of Delinquency Intervention Programs with Different Levels of Treatment Implementation

Program Group	Integrity of Treatment Implementation: Low	Medium	High
Group I Skills training Cognitive-Behavioral Multi-service	.29	.45	.62
Group II Counseling Drug abuse Milieu therapy Community residential	.20	.37	.53
Group III Group counseling Challenge programs	.07	.24	.40
Group IV Employment-related Guided group Teaching family home	-.02	.14	.30

The largest effects, however, do not simply follow from using one of the more effective program models. Figure 4 also shows that the integrity of the treatment implementation has at least equal influence on the outcomes. Treatment implementation in this analysis encompasses the amount of treatment provided and the extent of the program efforts to guard against degradation or incomplete coverage in their services. Even programs in the generally most effective group do not have effects in the larger ranges if they are not implemented well. Conversely, programs of a generally less effective type can nonetheless have relatively large effects by implementing their services well. Our analysis has shown many other program characteristics that are also systematically related to their effects, but this example illustrates the general point. Program effectiveness depends upon particular combinations of program features that must be optimally configured to achieve the best outcomes. Moreover, the critical program features are not necessarily unique to any particular program but show general patterns across programs.

Generalizations about the characteristics of the most effective programs, and how they are best combined, cannot be identified in the evaluation of a single program. They are only evident when patterns across programs can be examined. Discovering such relationships, therefore, is a distinctive and important contribution of meta-analysis to the field of evaluation research.

5. There is Much Room for Program Improvement

The outcome evaluation research studies generally available for meta-analysis in any program area typically include a mix of ongoing »real world« programs for which an evaluation has been conducted and various demonstration programs or research-oriented tests of program concepts. One of the useful comparisons that can be made in meta-analysis is to contrast the magnitude of the effects for the best-designed and implemented programs with those of an everyday sort. Demonstration programs designed and implemented by researchers to test state-of-the-art intervention concepts would be expected to produce better outcomes than routine practical programs. Not only do they potentially use more effective intervention approaches, but they also generally have greater control over the consistency of their services and the nature of their clientele.

In this regard, demonstration programs explore the upper limit of program effectiveness attainable with available intervention techniques and thus show what practical programs might aspire to under optimal circumstances. A large gap between the effects of practical programs and those of demonstration programs in an intervention area suggests that the practical programs may be able to improve their effectiveness by modeling key features of the demonstration programs. Unfortunately, meta-analytic investigation of the effectiveness of demonstration programs in contrast to everyday practical programs has, to date, only been undertaken in a limited way. The early indications, however, show rather sizeable gaps in favor of the demonstration programs (e.g., Weisz, Weiss, & Donenberg, 1992, on childrens mental health programs).

The nature of the situation can be illustrated with data from the meta-analysis of programs for juvenile delinquents to which we have already made reference several times. We divided the programs into real world practical programs evaluated by a researcher who was not involved in designing the program or delivering the service and contrasted their outcomes with demonstration programs designed and implemented by the researcher. Simply comparing the overall effect sizes for reoffense recidivism outcomes revealed that the mean for the practical programs (.07) was only about half that for the demonstration programs (.13), though both were modest (but with much variation around them).

When the characteristics of the practical and demonstration programs were compared, a number of specific differences emerged. Among the most important and interesting were the following.

- Type of program: less likely to be one of the more effective types (skill-building, behavioral, multi-service) for practical than demonstration programs.
- Administered by juvenile justice personnel: more likely for practical than demonstration programs.
- Monitoring of the integrity of the service implementation: less likely for practical than demonstration programs.
- Difficulties in treatment implementa-

tion reported: more likely for practical than demonstration programs.

- Program duration: about 25 weeks for practical programs; about 38 weeks for demonstration programs.
- Intensity of treatment: rated lower for practical programs than for demonstration programs.

Although some of the advantageous characteristics of the demonstration programs may be difficult for practical programs to emulate (e.g., program types that require highly trained personnel), others are clearly feasible. The results of comparisons such as this, therefore, can be used to guide the improvement of practical programs in ways that should enhance the magnitude of their outcome effects. The validity of this perspective is supported by analysis of the considerable variation within the domain of practical programs themselves. Not surprisingly, practical programs have many of the favorable program features identified above while others have less favorable configurations. If we examine the mean outcome effects for the practical programs that are more favorable configured in these terms, we find that they are indeed more effective.

Figure 5 shows one such comparison for the juvenile delinquency programs that focuses on reoffense recidivism outcomes. The practical programs are categorized according to how many characteristics they have from the set found in the meta-analysis to be related to effect sizes. There is a clear trend for those with a greater number of favorable characteristics to produce greater mean reductions in recidivism among their juvenile clients relative to control cases. Indeed, those with none of the favorable characteristics actually show an increase in recidivism among the juveniles they treat.

Figure 5.
Improvement in Recidivism Rates Relative to the Control for 196 »Real World« Delinquency Programs with Different Numbers of Favorable Program Characteristics

Number of Favorable Characteristics*	Distribution of Programs	Percentage Reduction in Recidivism
0	7%	+12
1	50%	-2
2	27%	-10
3	15%	-20
4	2%	-24

*Favorable Program Characteristics:
Uses one of the more effective types of service, e.g., skill-oriented, multimodal
Juvenile justice administered program conducted in non-JJ facility
Good program implementation with relatively high amount of service
Works with juveniles with mean age >15 years or with mixed prior offenses.

Perhaps equally interesting is the distribution of the programs represented in the meta-analysis across the various categories shown in Figure 5. More than half of the programs evaluated had zero or one favorable characteristic and, correspondingly, minimal or counterproductive effects. On the other hand, only 2% of the practical programs had the full complement of favorable characteristics and achieved the higher levels of recidivism impact. Possibly the most favorably configured programs are not evaluated, or their evaluations not reported, so that they would be underrepresented in the research available for meta-analysis. It seems more likely, however, that most practical programs, in fact, are not configured for optimal impact and have considerable room for improvement.

6. There is Safety in Numbers

Perhaps the most significant lesson from meta-analysis is the one that encompasses all the others: Many factors influence the findings of an outcome evaluation and, even under the best of circumstances, the validity of those findings is uncertain. While there is, and will continue to be, an important role for outcome evaluation of individual programs, we must be very cautious in interpreting a single set of results, even from a well-designed evaluation study. Ultimately, the most credible evidence about effective programs will come through careful integration of evaluation results from many studies and programs. Correspondingly, one of the greatest challenges facing the evaluation profession is how to ensure that high quality, useful synthesis of evaluation studies are carried out and the results disseminated to relevant evaluators, practitioners, and policymakers.

An important recent initiative offers great promise for meeting this challenge. In 1999 an international group of evaluators, policymakers, and researchers met at University College in London and agreed to launch the Campbell Collaboration for developing and disseminating systematic synthesis of outcome evaluation findings for social programs. This endeavor is modeled on the Cochrane Collaboration, which organizes syntheses of medical health-related research, and was named in honor of the American psychologist and methodologist, Donald Campbell, a renowned advocate for rigorous program evaluation. The Campbell Collaboration (C2) has grown rapidly and currently has a membership drawn from 15 countries and coordinating groups in the areas of crime and justice, education, social welfare, synthesis methods, and dissemination. C2 aspires to sponsor and facilitate high-quality synthesis of outcome evaluations for social programs and make them readily available on the world wide web to interested parties (http://www.campbellcollaboration.org). Though still in its infancy, the Campbell Collaboration has numerous syntheses underway and holds great promise as a way to extract and share the lessons that can be learned from the thousands of studies conducted in the vigorous field of program evaluation.

References

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cook, T. D. (2000) Toward a practical theory of external validity. In L. Bickman (ed.), Validity & social experimentation: Donald Campbell's Legacy (vol. 1, pp. 3-43). Thousand Oaks, CA: Sage.

Cooper, H. M. (1998). Synthesizing research: A guide for literature reviews (3d ed.). Thousand Oaks, CA: Sage.

Cooper, H. M., & Hedges, L. V. (Eds.). (1994). The handbook of research synthesis. New York: Russell Sage.

Haddock, C. K., Rindskopf, D, & Shadish, W. R. (1998). Using odds ratios as effect sizes for

meta-analysis of dichotomous data: A primer on methods and issues. Psychological Methods, 3, 339-353.

Lipsey, M. W. (1995). What do we learn from 400 research studies on the effectiveness of treatment with juvenile delinquents? In J. McGuire (ed.). What works? Reducing reoffending (pp. 63-78). NY: John Wiley.

Lipsey, M. W. (1992a). The effect of treatment on juvenile delinquents: Results from meta-analysis. In F. Loesel, D. Bender, & T. Bliesener (eds.). Psychology and law: International perspectives (pp. 131-143). Berlin; NY: Walter de Gruyter.

Lipsey, M. W. (1992b). Juvenile delinquency treatment: A meta-analytic inquiry into the variability of effects. In T.D. Cook, H. Cooper, D.S. Cordray, H. Hartmann, L.V. Hedges, R.J. Light, T.A. Louis, & F. Mosteller (eds.). Meta-analysis for explanation: A casebook. NY: Russell Sage Foundation.

Lipsey, M. W. (2000). Statistical conclusion validity for intervention research: A ($p<.05$) problem. In L. Bickman (ed.), Validity and social experimentation: Donald Campbell's legacy (vol. I). Sage.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. American Psychologist, 48, 1181-1209.

Lipsey, M. W., & Wilson, D. B. (2001). Practical meta-analysis. Applied Social Research Methods Series, vol. 49. Thousand Oaks, CA: Sage.

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

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. American Psychologist, 32, 752-760.

Weisz, J. R., Weiss, B. D., and Donenberg, G. R. (1992). The lab versus the clinic: Effects of child and adolescent psychotherapy. American Psychologist, 47, 1578-1585.

Wilson, D. B., & Lipsey, M. W. (in press). The role of method in treatment effectiveness research: Evidence from meta-analysis. Psychological Methods.

Om och varför? Den potentiella nyttan av att inkludera programteoriutvärderingar i metaanalys.

anthony petrosino

Två envisa utmaningar för att förstå effekterna av socialpolitiken är ansamlingen av rön från tidigare utvärderingsstudier och att förstå varför ett program lyckas eller misslyckas. Under de senaste trettio åren har metaanalys och programteoriutvärdering lanserats som metoder för att svara på denna utmaning. Författaren visar hur till och med enkla programteoriutvärderingar kan komma till nytta om de används vid metaanalys.

Inledning

Att bestämma om och varför ett program fungerade är en riskabel sysselsättning. Ett slumpmässigt experiment – om det realiseras och fullföljs med totalt oberoende – kan ge det minst tvetydiga svaret på frågan »fungerade programmet?« Men om man inte planerar och analyserar ytterligare data förutom resultatmätningarna kan ett experiment sällan ge svar på frågan »varför fungerade interventionen?«[1] Även om vi skulle lyckas med att utforma en studie som svarar på både »om« och »varför«, vilket inte vore en liten bedrift i sig, finns det ytterligare ett problem. Enstaka, fristående utvärderingar kommer sällan att ge definitiva resultat. Resultaten från utvärderingar av samma program varierar ofta från fall till fall beroende på skillnader i kunder, personal och så vidare (Lipsey, 1997).

Två olika perspektiv har emellertid framträtt som tillsammans skulle kunna ge bättre svar på »om« och »varför«. Det första är programteori. Under de senaste trettio åren har utvärderare skrivit övertygande –

Anthony Petrosino is Research Fellow at the Center for Evaluation, Initiatives for Children Program of the American Academy of Arts & Sciences. He is also a Research Associate and Project Manager for the Study of Decisions in Education at Harvard University

1 Vissa människor vill kanske hävda att faktoriella experiment isolerar mekanismerna till varför ett program fungerar.

och ofta – om behovet av att oförbehållsamt pröva programteori vid utvärdering (t.ex. Bickman, 1987; Chen and Rossi, 1992; Weiss, 1972). Även om vokabulären skiftar mellan dessa författare är alla överens om att utvärderingar, när det är möjligt, bör tydligt formulera och pröva de underliggande antagandena om varför programmet borde »fungera«.[2] I konsekvens med andra artiklar i den här volymen, har jag använt programteoriutvärdering (program-theory evaluation – PTE) för att beskriva studier som använder sådana termer som teoribaserad eller teoristyrd för att beskriva liknande metoder.[3]

Intresset för PTE går hand i hand med framkomsten av en »vetenskap om litteraturöversikter«. Även om problemet med att tolka skilda men liknande studier har uppmärksammats ända sedan 1904 sköt inte vetenskapen om litteraturöversikter fart förrän på 1970-talet med metaanalysen (Hunt, 1997). Eftersom man insåg att de traditionella metoderna för att sammanställa forskningsresultat var bristfälliga framträdde metaanalysen som en ytterst noggrann metod för sammanfattning av resultat från tidigare forskning (t.ex. Lipsey, 1990). Till skillnad mot PTE har användningen av metaanalys varit utbredd. Lipsey och Wilson (1993) lyckades identifiera drygt 300 metaanalyser av behandlingar inom det sociala fältet och inom utbildningsområdet.

2 Observera att det finns avvikande synpunkter beträffande värdet av PTE (t.ex. Stufflebeam, 2001).

3 Jag föredrar termen «orsaksmodell« på grund av den allmänna förvirring som råder kring termen «teori« (Petrosino, 2000).

Författare som Cordray (1992) och Lipsey (1997) menar att kombinationen av PTE och metaanalys skulle kunna ha fördelar. Lipsey (1997) visade till exempel hur de skulle kunna användas tillsammans för att skapa teorier om social intervention. I den här artikeln bygger jag på dessa tidigare argument för att visa hur upplagringen av kunskap från PTE genom metaanalys skulle kunna ge nyttiga data för socialpolitiska beslut och beslut inom den sociala praktiken. Både hypotetiska illustrationer och faktiska exempel kommer att användas.

Utvärderingsmodell i ett steg

Många utvärderingar prövar bara effekten av en intervention på resultatmätningarna – vilket ibland kallas ettstegsmodeller

Tack
En äldre version av denna artikel har tidigare publicerats som «Om och varför: Den potentiella ömsesidiga nyttan av att inkludera programteoriutvärderingar i metaanalys« i New Directions in Evaluation, 87, 59-70 ©2000, Jossey-Bass, ett dotterföretag till John Wiley & Sons, Inc. Tidigare utkast har framlagts för Harvard Children's Initiative Evaluation Task Force (april 1998) och American Evaluation Association (november 1998). En docentur vid Harvard Children's Initiative finansierad med medel från Spencer Foundation samt ett stipendium från Mellon Foundation till Center for Evaluation underlättade detta arbete. Jag är tacksam för synpunkter på tidigare utkast av denna artikel från Mary Askew, Anne Barten, Len Bickman, Iain Chalmers, Jodi Delibertis, Tim Hacsi, Tracy Huebner, Mark Lipsey, Heather McMillan, Frederick Mosteller, Pamela Perry, Carolyn Petrosino, Patricia Rogers, Sean Riordan, Robert Rosenthal, Haluk Soydan, Carol Weiss, David Wilson och Stuart Yeh. Det är emellertid endast författarens åsikter som det ges uttryck för i den här artikeln.

(Weiss, 1997). Problemet med ettstegsmodellen är att den inte förklarar varför ett program skulle påverka resultatet (Chen och Rossi, 1992). Den behandlar inte den orsaksmässiga komplexiteten som finns i många program som riktar in sig på sådana resultat som till exempel kriminellt beteende. Vissa sociala interventioner fungerar genom indirekta processer: en behandling inleds i ett sammanhang och förväntas ta andra avgörande mekanismer i anspråk för att påverka resultatet i ett annat sammanhang (Donaldson, 2001). Skolbaserad drogprevention inriktar sig till exempel på drogmissbruk bland ungdomar, vilket förmodas ta i anspråk till exempel sådana mekanismer som motstånd från kamraterna. Dessa mekanismer förmodas öka motståndet mot att använda droger både i och utanför skolan. Ettstegsmodellen skulle inte ta hänsyn till mätningen av sådana underliggande mekanismer som kamratmotstånd utan bara inrikta sig på sådana resultat som droganvändning.

Hur utvärderingar i ett steg påverkar metaanalysen

Valet att utföra ettstegsutvärderingar har konsekvenser för forskningssyntesen eftersom det är de ursprungliga studierna som förser metaanalysen med data. Det är viktigt att utvärderare inser att det sätt på vilket de genomför och redovisar sina studier påverkar senare genomgångar.

I de flesta metaanalyser tar granskarna fram en effektstorlek för att uttrycka programmets inverkan på den aktuella resultatmätningen. I en utvärdering i Illinois, till exempel, utvaldes 94 ungdomsbrottslingar slumpmässigt för att delta i ett fängelsebesök, medan 67 ungdomar inte kontaktades alls (Greater Egypt Regional Planning & Development Board, 1979). I programmet ingick också ett tillfälle då ungdomarna fick möjlighet att prata med fångarna, som återgav realistiska berättelser om fängelselivet, inklusive våldtäkter och mord. Programmet, av typen »Scared Straight« (Skrämd till skötsamhet), hade för avsikt att avskräcka ungdomsbrottslingarna från vidare brottslighet. Efter sex månader fann utvärderarna att 17 % av experimentgruppen hade anhållits på nytt jämfört med kontrollgruppens 12 %, vilket ger en effektstorlek på -0,14.[4] Minustecknet beror på att programmets påverkan gick i motsatt riktning mot den förväntade.

I de flesta metaanalyser anger man genomsnittet av sådana här effektstorlekar för alla inkluderade studier (t.ex. »program som avser att minska ungdomsbrottsligheten hade en genomsnittlig effektstorlek på 0,10«). I vissa metaanalyser anges genomsnittet av effektstorlekarna för specifika behandlingar som använder vida beteckningar (t.ex. »yrkesutbildningsprogram för att minska ungdomsbrottsligheten hade en genomsnittlig effektstorlek på 0,05«). En sådan beteckning tjänar två syften: den tillhandahåller en rubrik som beskriver den grundläggande interventionen, men är så pass vid att den innefattar mer än bara ett fåtal studier. Dessa kategorier har konsekvenser för metaanalysen. Kognitiv terapi, beteendeterapi, individuella terapeutiska

4 I Lipsey (1990) finns omvandlingsformler för effektstorlekar.

samtal och gruppterapi är samtliga behandlingar för sexualförbrytare (Laws, 1989). Om utvärderingarna av dessa behandlingar hade redovisats, skulle metaanalysen kunna ta vid och effektstorlekarna beräknas. De flesta metaanalytiker sätter som behörighetskriterium att utvärderingarna skall ha med antingen en kontrollgrupp eller en jämförelsegrupp.

I tabell 1 visas de hypotetiska resultaten från en metaanalys av studier kring behandling av sexualförbrytare. Av tabellen framgår att effektstorleken för kognitiv behandling var 0,30. Skulle ett sådant rön kunna vara viktigt? Rosenthal och Rubin (1982) har utarbetat en metod att översätta effektstorlekar till skillnader i procent, mera känt som Binomial Effect Size Display (BESD). Genom att använda BESD kan man översätta effektstorleken 0,30 till en 15-procentig förbättring för den kognitiva gruppen jämfört med kontrollgruppen. En genomsnittlig minskning med 15 procent av återfallen för sexualförbrytare skulle vara viktig.[5]

Men den vida beteckningen »kognitiv behandling« döljer viktig information om programmen i denna kategori. Det finns många olika slags kognitiva program för sexualförbrytare med stor variation i hur bara en enda typ av program genomförs på olika håll. Ansvariga för finansieringen eller

Tabell 1.
Hypotetisk metaanalys av ettstegsutvärderingar, behandling av sexualförbrytare (ordnade efter effektstorlek)

Program (N)	Effekt på återfall i sexualbrott	BESD
kognitivbaserat (11)	0,30	+15 %
beteende (10)	0,11	+5,5 %
individuellt (12)	0,01	+0,05 %
grupp (8)	–0,05	–2,5 %

införandet av kognitiva program anser dock kanske inte att resultaten i tabell 1 hjälper dem i deras beslutsfattande. Det finns många program som anses vara kognitiva och den vida beteckningen antyder inte vilket av de kognitiva programmen de bör använda.

Hur PTE skulle kunna användas vid metaanalys

Ett sätt att komma ifrån det här problemet är att öka antalet rigorösa utvärderingar med hjälp av programteorimetoden. I slutet av den här artikeln finns ett praktiskt förslag på hur man gör det. Till och med enkla PTE som bara inriktar sig på en enda nyckelmekanism och resultatet skulle visa att programmen fungerar genom en eller annan mekanism. Metaanalytiska rön skulle då kunna kategoriseras genom den nyckelmekanism som prövades i de ursprungliga utvärderingarna. Om sådana mekanismer prövades i ett antal PTE skulle metaanalysen bli bättre lämpad att ge vägledning om vad som är en effektiv intervention.

För att återgå till tabell 1, så visade rönen

5 Rosenthal och Rubins (1982) BESD kan omvandla effektstorlekar till skillnader i procent, vilket gör att även lekmän kan ta till sig resultaten. Enkelt uttryckt är BESD lika med hälften av effektstorleken d (om man t.ex. vid d=0,30 förutsätter en återfallsprocent (basline) på 50 %, återfaller experimentgruppen i 32,5 % av fallen och kontrollgruppen i 47,5 %).

att kognitiva program var den effektivaste strategin för att minska återfallen. Men det gavs ingen information om varför de kognitiva programmen var mer effektiva. Ännu mer problematiskt är det att den genomsnittliga effektstorleken innefattar program med olika effekter; vissa var antagligen mycket effektiva, medan andra förmodligen hade mindre effekt på återfallsrisken än till och med icke-kognitiva behandlingar (t.ex. gruppterapi).

I stället för de elva utvärderingarna av kognitiv behandling som redogjordes för i tabell 1, vad skulle hända om ett större antal enkla PTE inkluderades i metaanalysen? En enkel PTE är vad Lipsey och Pollard (1989) beskriver som en tvåstegsmodell: mätning och prövning av minst en förändringsmekanism samt ett resultat. Om man tillämpade en sådan modell på utvärdering av behandlingsprogram för sexualförbrytare skulle modellen pröva huruvida programmet först förändrade någonting som sedan i sin tur påverkade återfallsrisken.

I tabell 2 jämförs resultaten från tabell 1 med en hypotetisk metaanalys av en rad enkla PTE. Fördelen med PTE-metaanalysen är att den ger ledtrådar om förändringsmekanismer. Effektiva program känns lättare igen genom den nyckelmekanism de tar i anspråk. I tabell 2, till exempel, är kognitiva program som ökar färdigheten hos sexualförbrytare att identifiera och minska sina egna högrisksituationer mer effektiva när det gäller att minska återfallsrisken. Om kognitiva program skall användas är de till och med mer effektiva om de används samtidigt som man inriktar sig på empatin hos brottslingen. Sådana rön skulle vara mycket användbara när det gäller att ge vägledning inför ett beslut.

Wilson och hans kollegor (2000) har givit ett tidigt och äkta exempel på hur en sådan metaanalys kan gå till i praktiken. I sin studie analyserade de 33 studier som utvärderade effekten av arbete och yrkesutbildningsprogram på senare återfall från förövarnas sida. Bara nio studier redogjorde för effekten av ett program på både återfallsrisken och på den grundläggande nyckelmekanismen: sysselsättning. Med andra ord: Är det så att de program som ökar sysselsättningen (mekanismen) även minskar återfallsrisken? Trots att interna validitets-

Tabell 2.
Jämförelse mellan hypotetisk metaanalys av ettstegsutvärderingar och hypotetisk metaanalys av PTE

Med ettstegsutvärderingar:		Med enkla PTE:		
Program (N)	Återfall i sexualbrott	Kognitivbaserade program (N)	Effekt på mekanism	Återfall i sexualbrott
kognitivbaserade (11)	0,30	färdighet, empati med offer (7)	0,61	0,44
beteende (10)	0,11	färdighet i högrisksituationer (14)	0,55	0,38
individuella (12)	0,01	minska rationalisering (8)	0,28	0,12
grupp (8)	–0,05	öka empatin med offren (12)	0,25	0,09

hot var vanliga i de studier som Wilson et al. (2000) tog exempel ifrån rapporterar de att »… större skillnader i återfallsfrekvens hos programdeltagare jämfört med sådana som inte deltog står i samband med större skillnader i sysselsättningsstatus vid uppföljning. I program där man konstaterar en påverkan på sysselsättningen brukar också noteras en påverkan på återfallsrisken« (s. 361).

Minimikrav på PTE för metaanalys

Om PTE skall kunna användas vid metaanalys bör de leva upp till tre kriterier: (1) en tydlig orsaksmodell för hur programmet kommer att påverka resultaten, (2) prövning av minst en underliggande mekanism som interventionsvariabel tillsammans med resultaten och (3) kontrollgruppsdata redovisade för båda variablerna. Det första kriteriet kräver att en prospektiv och tydlig modell prövas vid utvärderingen. Förutom att tydligheten minskar antalet gissningar som läsaren måste ägna sig åt beträffande vilken programteori det rörde sig om, så förhindrar det prospektiva kravet att data efterhand anpassas till modellen.[6]

Det andra kriteriet kräver att PTE innefattar minst en mellanliggande variabel. Inom utvärderingsvärlden är en mellanliggande variabel något som programmet måste påverka eller förändra för att på ett positivt sätt påverka slutresultatet (Lipsey och Pollard, 1989). Vissa utvärderingar där programteori har ingått har inte försökt pröva några nyckelförbindelser eller -mekanismer i modellen (Petrosino, 2000). Om det bara är resultatdata som analyseras och redovisas spelar det ingen roll om en programteori är tydlig för då kommer inte sådana utvärderingar att ge mer information än ettstegsmodeller – programmets orsaksmodell prövades inte. Enligt det tredje kriteriet måste kontroll- och jämförelsedata redovisas för både de mellanliggande variablerna och resultatvariablerna. Om data om den mellanliggande variabeln bara redovisas beträffande behandlingsgruppen ger utvärderingen inte mycket belägg för att effekten på den mellanliggande variabeln skulle ha inträffat utan programmet (Cook, 2000).

Utvärderingar av processresultat motsvarar inte heller minimikraven för PTE eftersom de inte ger några data om de underliggande mekanismerna. Till och med när utvärderare kopplar samman processdata med resultaten i sina analyser återspeglar de programaktiviteternas påverkan och graden av trohet gentemot resultatet – inte den underliggande förändringsteorin (Weiss, 1997). Det är något med de här aktiviteterna som borde ta en avgörande mekanism i anspråk. Vad är detta något? Det är det som PTE måste ge uttryck för och pröva (se Weiss, 2000).

6 Vissa utvärderingar samlar ihop ansenliga mängder data om deltagarna och en del information skulle kunna konceptualiseras som mellanliggande variabler. Metaanalysen kan hantera sådana «vardagsutvärderingar« – om det finns tillräckligt många – och undersöka korrelationerna för både de mellanliggande variablerna och resultatmätningarna.

Hur skulle metaanalys av PTE kunna vara till ledning för mer omfattande sociala teorier?

Inom PTE är teori en förklaring till hur programmet kommer att åstadkomma de avsedda resultaten. Mer omfattande vetenskapliga teorier erbjuder generella förklaringar till sådana fenomen som kriminalitet, dålig inlärning och till och med om hur program genomförs. En metaanalys av PTE skulle potentiellt kunna vara till ledning för sådana mer omfattande teorier.

I Chandlers (1973) experimentella utvärdering av en förebildsintervention med brottsbenägna ungdomar prövade han till exempel en tvåstegsmodell: en minskning av självupptagenheten (dvs. bristen på empati med andra) skulle minska brottsligheten (Lipsey och Pollard, 1989). Chandler genomförde en uppföljning efter två år och kom fram till att behandlingen åstadkom statistiskt signifikanta minskningar i både självupptagenhet och brottslighet.

I stället för att bara visa ett experiment ger tabell 3 ett hypotetiskt exempel på hur 50 studier som Chandlers skulle kunna vara till ledning för teorier om brottslighet. I tabell 3 prövar tio PTE självupptagenhetsmodellen och fyra andra grupper av PTE prövar olika förändringsmekanismer. Tabellen visar hypotetiska effektstorlekar för var och en av fem kategorier. Intervention som inriktade sig på självupptagenheten åstadkom hypotetiskt sett större effekter på både den mellanliggande variabeln och den påföljande brottsligheten. Ett sådant rön antyder att självupptagenhet är en avgörande länk i uppkomsten och utvecklingen av brottslighet.

De hypotetiska rönen visar också att interventioner i allmänhet hade mindre effekt på mätningar av självkänsla, yrkesskicklighet, familjefunktion och rädsla för påföljder. Sådana rön skulle kunna få brottsteoretiker att på nytt undersöka förhållandet mellan sådana faktorer och den påföljande brottsligheten. Naturligtvis skulle effekterna kunna bero på att programmet genomfördes dåligt eller på en allmänt ineffektiv behandling (om de tio studierna grundades på en gemensam behandlingstyp). Men när allt kommer omkring skulle en generation av PTE för metaanalys kunna ge en del teorier om ledtrådar, inte minst om sådana problem som brottslighet.

Tabell 3.

Hypotetiska effektstorlekar för mellanliggande variabler och resultatvariabler (ordnade efter effekt på den mellanliggande variabeln)

Mellanliggande (N) variabel	Effekt på mellanliggande variabel	Effekt på brottslighet
självupptagenhet (10)	0,64	0,34
självkänsla (10)	0,48	-0,07
familjefunktion (10)	0,36	0,17
förbättring av yrkesskicklighet (10)	0,22	0,02
rädsla för brottspåföljd (10)	0,12	-0,15

Minimitröskelnivåer och »kaskadeffekter«[7]

Om programteori var väl utvecklad för en omfattande intervention skulle PTE-data kunna användas i metaanalys för att ge information till beslutsfattare. Till exempel skulle mellanliggande variabler och resultatvariabler kunna tillåta uppskattningar av minimitröskelnivåer, dvs. den förbättring som skulle krävas i den mellanliggande variabeln för att ge förbättrade resultat. Detta skulle kunna var till hjälp för program där de mellanliggande variablerna mäts vid något tillfälle före resultaten. Ett misslyckande med att åstadkomma en mellanliggande effekt skulle kunna tjäna som en varningssignal för beslutsfattare om att programmet är på väg mot dåliga resultat (Weiss, 1997).

Sådana data skulle också kunna skildra »kaskadeffekter«, eller sannolikheten för att det för varje påföljande mellanliggande variabel eller länk i en programteorikedja troligen kommer att redovisas mindre effekter. I till exempel kunskap-attityd-beteende (KAB)-modellerna (Lipsey, 1997) visar program som redovisar stor effekt på kunskap vanligen mycket mindre effekter på attitydmätningarna och ännu mindre på beteenderesultaten. Data om dessa kaskadeffekter skulle kunna användas för att tala om för beslutsfattare att ett program behöver nya verktyg för att kunna åstadkomma större effekter på kunskaper eller attityder för att mer effektivt kunna förändra beteendet i ett senare skede. Eller också behöver kanske modellen revideras för att få med andra mellanliggande variabler mellan länkarna attityd och beteende.

Mervärdet av PTE

Metaanalys skulle kunna ge en metod för uppskattning av det mervärde som PTE ger. Mervärde betyder ofta någonting som kan mätas matematiskt, men här syftar det på huruvida PTE ger någon fördel förutom att tillhandahålla andra sätt att gripa sig an utvärdering eller inte. Trots att fördelarna med PTE länge har antytts har de inte blivit empiriskt påvisade. Ett sätt att pröva mervärdet är genom metaanalys. Lipton (1995) och hans kollegor genomför till exempel en metaanalys av utvärderingar av kriminalvårdsprogram som redovisats sedan 1968. Deras metaanalys kommer sannolikt att innehålla över 1 000 utvärderingar, varav vissa använder andra metoder såsom PTE.

Deras data skulle kunna användas för att jämföra PTE med dessa andra utvärderingsmetoder. Utvärderingarna kanske inte är så lätta att kategorisera, men skulle kunna graderas längs ett kontinuum för hur väl utvecklad den teori är som används för att styra utvärderingen (Lipsey, 1988). Graderingarna skulle kunna analyseras för att bestämma påverkan av teoriutvecklingen inom PTE på ett antal beroende variabler, inklusive effektstorlek, programframgång eller -misslyckande och så vidare. Ett mindre antal av studierna skulle kunna undersökas för att bestämma hur de användes i de påföljande besluten. Även om framtagna data bara är ungefärliga skulle de kunna ge vägledning beträffande

7 Mark Lipsey föreslog termen «kaskadeffekt« vid personlig kontakt i april 1998.

de reella fördelarna som PTE ger avseende processresultat eller utvärderingar enligt ettstegsmodellen.

Hinder och begränsningar

Det är inte många författare som skulle ha invändningar mot att ta med mellanliggande variabler vid utformningen av en utvärdering (t.ex. Cook, 2000). Men likaväl som det finns hinder mot att genomföra en enstaka PTE finns det vägspärrar uppsatta mot användningen av PTE i metaanalys. Bland dessa finns:

Det låga antalet PTE. Det största hindret är det låga antalet tillgängliga PTE som redovisats i facklitteraturen. Vårt eget sökande efter bra exempel på PTE var svårt (se Rogers et al., 2000). Till och med enkla PTE som kräver tvåstegsmodeller är svåra att hitta. Det är sällan som utvärderare tydligt och prospektivt redovisar en modell som kan prövas.

Betoningen på experiment och kvasiexperiment. De flesta metaanalyser kräver som ett behörighetskriterium att de ursprungliga utvärderingarna innefattar en kontroll- eller jämförelsegrupp. Det här är ren kohandel eftersom man ökar den interna validiteten men utesluter potentiellt användbara studier som använder andra metoder för att utvärdera program.

Dålig redovisning. Granskare är allmänt missbelåtna med den dåliga redovisningskvaliteten i originalforskningshandlingar. Kombinationen av PTE och metaanalys skulle kräva att mer data samlades ihop, analyserades och redovisades av utvärderarna. Alla rekommenderar en förbättring av redovisningskvaliteten men det har varit svårt att hitta lösningar.

Förenklade programteorier. Den här artikeln har inte tagit hänsyn till komplicerade modeller. De enkla PTE som diskuteras här är linjära och förutsätter en dominoeffekt – en förändring i en variabel kommer att resultera i en påföljande förändring i nästa mätta variabel. Som Rogers (2000) noterar fungerar kanske inte världen på det sätt som dessa modeller antyder. Modeller kan vara utförliga även i linjära teorier. Weiss (1997) förtecknar 17 länkar i sitt yrkesutbildningsexempel. Som hon noterar om utvärderare som genomför originalstudier (2000), kan metaanalytiker också tvingas bestämma vilka länkar i vilken teori de skall koda och undersöka vid sina genomgångar.

En rekommendation att främja PTE

Som nämndes tidigare är det huvudsakliga hindret för den här metoden bristen på PTE. Sherman och hans kollegor (1997) föreslår en metod för att öka antalet högkvalitativa utvärderingar. I deras genomgång av brottsförebyggande studier för den amerikanska kongressen undersökte de också de utvärderingskrav som ställs av den federala regeringen och delstatsregeringarna när de finansierar program som är relaterade till rättskipning i brottmål. Trots att det som en förutsättning för finansieringen vanligtvis krävs att alla anslagsmottagare skall genomföra en utvärdering fann Sherman och hans kollegor att bara ett fåtal finns redovisade. Ett problem är att det som

kallas för utvärdering ibland bara är inputdata eller information om betjänade kunder. I det fåtal resultat- eller påverkansstudier som genomförs ingår sällan kontroll- eller jämförelsegrupper. Kontentan är att mycket litet är känt om vad det är som fungerar inom brottsförebyggandet (Sherman et al., 1997). En bidragande orsak är den i allmänhet otillräckliga mängden pengar för utvärdering som finns med i programbudgeten. De föreslår en annan metod: I stället för att kräva utvärdering av alla finansierade program bör det administrerande organet (vanligen på federal nivå eller delstatsnivå) slå samman utvärderingsmedlen för att stödja ett mindre antal utvärderingar av hög kvalitet på bara några få platser (t.ex. Sherman et al., 1997). En sådan metod skulle kunna bidra till att främja en ökning av antalet mycket noggrant styrda PTE.

Ett exempel på hur den här strategin skulle kunna fungera i praktiken kommer från ett program finansierat av federala medel.[8] Titel V-anslagen för lokala brottsförebyggande program från Office of Juvenile Justice and Delinquency Prevention (OJJDP) stödde 477 särskilda interventioner i USA under räkenskapsåren 1994-1997 (OJJDP, 1998). I stället för att kräva en utvärdering i alla de 477 fallen (budgeten uppgick vanligtvis till 10 000 dollar per anslagsställe) skulle de 4,7 millioner dollar som avsatts till utvärdering kunna användas till finansiering av PTE i 20 av fallen. Varje PTE skulle kunna innefatta randomiserade kontrollerade studier eller någorlunda meningsfulla jämförelsegrupper. Varje utvärdering skulle kunna få 200 000 dollar per anslagsställe (till en total kostnad av 4 miljoner dollar). De andra medlen (700 000 dollar) skulle kunna användas för att bygga in en begränsad samling av data för kontroll av de ställen som inte har PTE. Om man använde sig av det här tillvägagångssättet skulle en metaanalys av de 20 PTE:na kunna utföras på en rimligt kort tid. En systematiskt och mycket noggrant utförd granskning av 20 PTE skulle med all säkerhet kunna ge mycket bättre information än de 477 lågkvalitativa och utspridda utvärderingar som annars ofrånkomligen skulle komma att redovisas.

8 De statliga byråerna förmedlar vanligtvis anslagen i olika «finansieringsriktningar«. USA:s justitiedepartement har t.ex. många finansieringsriktningar (t.ex. lagen om kvinnovåld). Anslagen fördelas blockvis till delstaterna för de olika riktningarna och sedan fördelar delstaterna dem i sin tur som underanslag (Sherman et al., 1997).

Referenser

Bickman, Leonard, Editor (1987) Using Program Theory in Evaluation. San Francisco, CA: Jossey-Bass.

Chandler, MJ (1973) Egocentrism and antisocial behavior: The assessment and training of social perspective-taking skills. Developmental Psychology 9, 3, 326-332.

Chen, Huey T. and Peter Rossi. Editors (1992) Using Theory to Improve Program and Policy Evaluations. New York: Greenwood.

Cook, Thomas D. (2000) The false choice between theory-based evaluation and experimentation New Directions in Evaluation 87, 27-34.

Cordray, David S. (1992) Theory-driven meta-analysis: Practices and prospects In Chen, H.T., Rossi, P. ed.: Using Theory to Improve Program and Policy Evaluations. New York: Greenwood.

Davies, P., A. Petrosino and I. Chalmers (1999) Proceedings of the International Meeting on Systematic Reviews of the Effects of Social and Educational Interventions. July 15-16, London: University College-London, School of Public Policy.

Donaldson, Stuart I. (2001) Mediator and moderator analysis in program development In Sussman, S, Editor: Handbook of Program Development in Health Behavior Research and Practice. Newbury Park, CA: Sage.

Greater Egypt Regional Planning & Development Commission (1979) Menard Correctional Center. Juvenile tours impact study. Carbondale, IL: Greater Egypt Regional Planning & Development Commission.

Hunt, Morton (1997) The Story of Meta-Analysis. New York: Russell Sage Foundation.

Laws, Richard, Editor (1989) Relapse Prevention with Sex Offenders. New York: Guilford.

Lipsey, MW (1997) What can you build with thousands of bricks? Musings on the cumulation of knowledge in program evaluation. New Directions in Evaluation 76,7-24.

Lipsey, M. W. (1990) Design sensitivity: Statistical power for experimental research. Newbury Park, CA: Sage.

Lipsey, MW and Wilson, DB (1993) The efficacy of psychological, educational and behavioral treatment. Confirmation from meta-analysis. American Psychologist 48, 12, 1181-1209.

Lipsey, MW and Pollard JA (1989) Driving toward theory in program evaluation: More models to choose from. Evaluation and Program Planning 12,317-329

Lipsey, MW (1988) Practice and malpractice in evaluation research. Evaluation Practice 9, 4,5-24.

Lipton, DS (1995) CDATE: Updating the Effectiveness of Correctional Treatment 25 years later. Journal of Offender Rehabilitation, 22,(1/2),1-20.

Petrosino, A (2000) Answering the why question in evaluation: The causal-model approach. Canadian Journal of Program Evaluation 12, 1, 1-25.

Office of Juvenile Justice and Delinquency Prevention (1998) 1998 Report to Congress Title V Incentive Grants for Local Delinquency Prevention Programs. Washington, DC: OJJDP.

Rogers, Patricia (2000) Causal models in program theory evaluation. New Directions in Evaluation 87, 47-56.

Rogers, Patricia, Anthony Petrosino, Tim Hacsi and Tracy Huebner (2000) Program theory evaluation: practice, promise, problems. New Directions in Evaluation 87, 5-14.

Rosenthal, R, and Rubin, D (1982) A simple, general purpose display of magnitude of experimental effect. Journal of Educational Psychology 74,166-169.

Sherman, Lawrence W., Denise Gottfredson, Doris MacKenzie, John Eck, Peter Reuter and Shawn Bushway (1997) Preventing Crime: What Works, What Doesn't, What's Promising. A Report to the United States Congress. College Park, MD: University of Maryland, Department of Criminology and Criminal Justice.

Stufflebeam, Daniel, Editor (2001) Evaluation

Models. New Directions for Evaluation 89. San Francisco, CA: Jossey-Bass.
Weiss, Carol H. (1972) Evaluation. Englewood Cliffs, NJ: Prentice-Hall.
Weiss, CH (1997) Theory-based evaluation: Past, present, and future. New Directions for Evaluation 76,41-55
Weiss, Carol H. (2000) Which links in which theories shall we evaluate? New Directions in Evaluation 87, 35-46.
Wilson, DB, Gallagher, CA and MacKenzie, DL (2000) A meta-analysis of corrections-based education, vocation and work programs for adult offenders. Journal of Research in Crime & Delinquency 37, 4, 347-368.

Summary

Whether and why? The potential benefits of including program theory evaluations in meta-analysis

Over the past thirty years, two different methodologies have been promoted to resolve persistent challenges to those attempting to ascertain the effectiveness of social programs and policies. Many argue that it is not enough to know whether a program works or not, but why it succeeds or fails. What are the critical mechanisms or mediating influences involved? Over the past three decades, scientists have proposed theory-based or theory-driven (i.e. program theory) evaluation as a method for understanding more about why a program works or fails. Another challenge is how to handle the results from separate but similar studies. Meta-analysis, or the quantitative analysis of prior research studies, has been promoted as one method for resolving the difficulty of accumulating the results from prior evaluation studies. In this paper, the author proposes merging these two methodologies to provide even more benefit.

The author first describes simple one-step models that examine the effects of a program on an ultimate or distal outcome only. The paper then contains an illustration, using hypothetical data, on how such one-step models can affect meta-analysis. In response, program theory evaluations provide further insight on the critical mechanisms that mediate the relationship between the program and the ultimate outcome. The author uses both hypothetical and real world data to show how this information can be used to illuminate our understanding of social policy and program impact. For program theory evaluations to be used in meta-analysis, they must meet a set of minimum requirements, including the measurement of the impact of the program on at least one mediating variable and one outcome variable. Other benefits suggested by this merger include informing larger social science theories, identifying minimum threshold levels and cascading effects, and specifying the value-added by theory-based evaluation. Limitations are also discussed. The author concludes with one recommendation for increasing the number of high-quality program theory evaluations that can later be analyzed in research synthesis.

Estimating the Effects of Interventions in Multiple Sites and Settings: Place-based Randomized Trials[1]

robert boruch, ellen foley & jeremy grimshaw

1.Introduction

A place-based trial here means a study in which a number of places or organizations are randomly assigned to one of two or more interventions so as to learn which intervention works best. The »places« may be villages or neighborhoods, schools or juvenile facilities, housing projects, or other organizations. The places that are assigned to interventions will not differ at the outset. They are statistically equivalent on account of the random assignment. This equivalence permits a fair comparison, i.e. an unbiased estimation of the relative effects of the intervention and a statistical statement of one's confidence in the results.

Trials in which individuals are randomly assigned to different interventions are familiar in medical and other research. Random allocation of units such as places and entities are less frequent. As Donald T. Campbell suggested in »Reforms as Experiments«:

Where policies are administered through individual client contacts, randomization at the person level may often be inconspicuously achieved....
But for most social reforms, larger administrative units will be involved, such as classrooms, schools, cities, counties or states. We

Robert Boruch is University Chair Trustee Professor at the Graduate School of Education, the Statistics Department of the Wharton School of Business, and the Fels Center for Government at the University of Pennsylvania.
Ellen Foley is the Senior Associate in District Redesign at the Annenberg Institute for School Reform, Brown University and serves as Research Director for School Communities that Work: A National Task Force on the Future of Urban Districts.
Jeremy Grimshaw is the Director of the Clinical Epidemiology Unit, Ottawa Health Research Institute and Director of the Centre for Best Practices, Institution of Population Health, University of Ottawa.

1 This paper is based on work supported by the U.S. Department of Education's Planning and Evaluation Service, The Rockefeller Foundation, and the Swedish Center for Evaluation of Social Services (CUS) in Stockholm

need to develop the political postures and ideologies that make randomization at this level possible. (Campbell, 1969; Campbell, 1988)

Campbell, did not consider deeply the use of places or entities in randomized trials because such trials, at the time, were rare. In what follows, we depend on Campbell's insight and build on others' more recent work. The topic is germane to evaluation of complex social programs that are designed to enhance health and well-being, welfare and, education, and to reduce crime and delinquency.

1.1 Definitions

The unit of allocation refers to who or what is randomly assigned to different interventions in a trial. Conventional textbooks in psychology and design of medical trials, for instance, typically handle experiments in which individuals are the units of allocation. Here, we focus on sites, administrative units or groups, rather than on individuals. We refer to »place-based randomized trials« in this paper. Such trials are also called »group randomized trials« (Murray, 1998) and »cluster randomized trials« (Donner and Klar, 2000).

The units of analysis are those for which data are available and used. Juvenile facilities may be the units of random allocation in a trial that compares two facility-wide approaches to reducing recidivism. The units of analysis may be the facilities or both facilities and individuals within facilities.

1.2 The Contents of this Paper

In what follows, we discuss assumptions about the use of randomized trials and their rationale. Further, we identify difficulties in their use. The examples in this paper are diverse, partly to demonstrate that useful trials can be carried out in a variety of settings.

1.3 Assumptions

The first assumption is that the government agencies and private foundations are interested in estimating the relative effect of new programs that they sponsor. Put another way, we assume that the public is interested in answering the question: »What works better, for whom, and for how long?«

A second assumption is that a defensible estimate of an innovation's effect depends on determining how sites or other entities would behave in the absence of an innovation. As a practical matter, one might, for example, develop such an estimate from time series forecasts. Kuusi's (1957) study on the effect of alcohol sales in Finland is a remarkable precedent in using administrative records in short time series. Here, we assume that time series data and ad hoc comparisons are insufficient to produce unbiased estimates of a program's effect. Some of these alternatives to randomized trials, including time series and their vulnerabilities, are covered by Campbell and Stanley (1963) and by Shadish, Cook, and Campbell (2002).

Most important, a simple, and scientifically defensible method of composing

a comparison group, one that permits fair estimates of the relative differences among programs, is the method of random assignment. For instance, a sample of juvenile facilities might be randomly selected from the pool of eligible facilities and engaged in a new intervention program. The outcomes at these facilities would then be compared to the eligible facilities that were randomly assigned to continue operating under the existing programs. The random assignment assures that the two groups of facilities do not differ systematically, apart from the influence of the intervention program under study.

A third assumption is the future of impact evaluation in the many countries lies with controlled trials that are mounted on a small scale so as to understand which programs work before such programs are mounted at the national or regional level. In fact, such experiments have been undertaken and their frequency has increased. Boruch and Foley (2000), for example, list over 50 different studies involving communities or geographic sectors, schools or classrooms, housing projects, and other kinds of organizations as the units of allocation in a randomized field trial. See Boruch (1997), Donner and Klar (2000), and Murray (1998) generally, and the Campbell Collaboration's Social, Psychological, Educational, and Criminological Trials Register (http://www.campbellcollaboration.org).

2. Rationale: Why Use Sites as the Units?

Why should we consider places or other entities as the units of assignment to programs in evaluating the effect of a program? The reasons include: program theory; law and ethics; policy; the counsel of advisory groups; and statistical theory and rules of evidence.

2.1 Program Theory

By »theory« here, we mean how an intervention is supposed to have the effects that we believe they will have. In other words, the theoretician proposes a »logic model« to explain tentatively what happens when a program is implemented. Or, the theoretician may outline a formal path model or a causal chain.

Numerous theories of societal change posit that a program will work if it is delivered by organizational elements acting in concert. Research on preventing sexually transmitted diseases for example depends on theories about what institutional and group factors influence risky behavior. See Wasserheit, Aral, Holmes, and Hitchcock (1991) generally and Hornik (1991) in particular. Randomized field trials undertaken in California and Texas have employed 20 schools as the unit of allocation and analysis so as to test programs based on several such theories (Coyle et al., 1996; Basen-Engquist et al., 1997).

A variety of place-based randomized trials have also used schools as units to assess theory-driven programs that were designed to prevent or reduce substance abuse. The Midwestern Prevention Project (Pentz, 1994), for example, was based on a theory that adolescents' drug use depends on their characteristics, such as prior

drug use, and on the adolescents' ability to handle peer pressure toward using drugs. The theory also recognized that environmental and situational factors beyond the individual are important because community norms, for instance, can influence adolescent behavior.

Theory has also driven multi-stage research on how to engage and encourage mental hospital practices shown in earlier research to be more effective for treating certain forms of mental illness. Such theory involved ideas about the level at which the hospital staff might first be engaged (top down or bottom up) and the best mode of engagement. The latter included involvement in workshops or demonstration projects as opposed to merely sending brochures. The expectation was that people would react differently to these various engagement strategies (Fairweather et al., 1974; Fairweather and Tornatzky, 1977).

2.2 Law, Ethics, and Culture

One reason why sites might be used as the units of random assignment in a trial is that the random assignment of individuals to alternative programs within a site may not be legal or ethical. Or, this kind of randomization may not be acceptable on cultural or political grounds. Random allocation of entire sites to alternative programs might be regarded as both legally and ethically responsible.

For instance, in a randomized trial testing the Drug Abuse Resistance Education model (D.A.R.E.), researchers randomly assigned entire schools to treatment and control groups partly because it would have been difficult to get cooperation from schools if some of their students received the program and some did not (Curtin, personal communication, April 3, 1996). A kind of institutional ethic or culture prevailed. Using schools as the units of random assignment helped insure the cooperation of control schools in the trials. Schools in the control group were promised access to the D.A.R.E. program the year after the completion of the study.

Similarly, each of the 80 or so juvenile facilities in Sweden, for example, may object to random allocation of their clients to different programs so as to discern which program is most effective in reducing recidivism. Other ethical values in the local facility may take precedence, e.g., giving the »same« service to everyone in the facility. A randomized trial in which eligible and willing facilities try out one of two different approaches may then be regarded as more just. This point was made by Karin Tengvald at Stockholm's meetings on evaluating social service programs (Soydan, 1998).

Again, the emphasis here is on comparing alternative interventions in different communities, not on giving one set of these groups a »treatment« and leaving the others high and dry. The focus, then, is not simply on whether a treatment works but on which treatment works better.

2.3 Policy and Politics

As a matter of policy and politics, the government agency or foundation that sponsors programs make rules that affect

organizations directly rather than individuals directly. Such rules require sites or organizations to take particular actions, create transactions, and so on. The implication is that a study of the effects of such a program has to recognize sites as the immediate targets in an evaluation design. The individuals within sites are the ultimate targets.

For example, federal policy on demonstration projects in the United States has emphasized, at times, that communities are essential in ameliorating certain social problems. Preventing substance abuse is a case in point. The Center for Substance Abuse Prevention (CSAP) was created to reduce the incidence of alcohol, tobacco, and drug use. It has tried to do so through efforts such as the Community Partnership Demonstration Program which has focused on learning how diverse community-based organizations can be engaged in effective intervention. Different ways to do so were described by Kaftarian and Hansen (1994). The emphasis was on communities as the units of allocation and analysis in randomized field trials (Pentz, 1994; Wagenaar et al., 1994; Ellickson, 1994; Murray and Wolfinger, 1994; Lorion, 1994).

Other examples of programs in which the most direct connection are between entities and government or foundation assistance rather than between individuals and such assistance are easy to identify. They appear in compensatory education and other programs sponsored by the U.S. Department of Education; and the U.S. Department of Health and Human Services. Loans made by the World Bank to governments are operationalized by organizations such as banks, agricultural stations, or schools. The World Bank rarely supports randomized trials, but there are a few examples of programs sponsored by bank loans that have been tested in place-based trials.

2.4 Statistical Theory and Analysis

Contemporary statistical analysis methods rely on the assumption that an observation on any given individual or entity is independent of observations on all the others. When the assumption does not hold, and the analyst fails to recognize this, the analysis will be compromised. For instance, difference in program effectiveness may be declared statistically significant because the analysis is wrong in failing to recognize non-independence. See for instance Donner and Klar (2000) and Murray (1998).

Assuming that the units of observation are independent is not plausible in many settings. For example, a particular gang member's response to a juvenile crime reduction program may not be independent of other gang members' responses even though the program involves only some members. A child's grade on a test of ability to work in teams presumably will not be independent of grades given to other children on the same team.

For the statistician, all this implies that it is not individuals who ought to be randomly assigned to programs. And it is not individual level data that must ordinarily be used to estimate the program's effect. Rather, allocation and analysis should focus first on entire groups or organizations and second on individuals within each group or entity.

2.5 The Counsel of Advisory Groups on Research and Evaluation Policy

At times, preventing dangerous diseases, including sexually transmitted ones, requires that the programs be deployed through organizations or geopolitical jurisdictions. As a consequence, the National Academy of Sciences Panel on Evaluating AIDS Prevention Programs suggested that diagnostic testing and counseling sites be considered as the units in controlled experiments to improve the services (Coyle, Boruch, and Turner, 1991). Multidisciplinary conferences on sexually transmitted diseases (STDs), sponsored by the National Institute on Allergy and Infectious Diseases (NIAID), have led to the observation that clinical practices, factories, churches, and other organizations, as well as communities, might properly serve as the units in randomized trials (Green and Washington, 1991).

In considering approaches to preventing abuse of controlled substances, the participants in the »Communities that Care« Evaluation Design Conference said:

rigorous evaluation of a comprehensive community intervention requires an experimental design whereby communities are randomly assigned to experimental and control conditions.

See Peterson, Hawkins, and Catalano (1992). England's Joseph Rowntree Foundation has been influenced by similar concerns (Farrington, 1997).

The National Research Council's Panel on the Understanding and Control of Violent Behavior offered the following:

Recommendation 4: The panel calls for a new multi-community program of developmental studies of aggressive, violent, and antisocial behaviors, intended to improve both causal understanding and preventive interventions… (p. 25).

Edited by Reiss and Roth (1993), this Panel's report argued that »Randomized controlled field experiments usually have important advantages as an evaluation strategy« (p. 320).

Finally, consider that »Design and Analysis Issues in Community Trials« was the primary topic on the agenda of a 1992 National Institutes of Health conference. The participants agreed that the use of the communities as the units of allocation and analysis presented challenges, but that there were a variety of techniques for overcoming these challenges (Murray et al., 1994).

3. Examples

People often do not realize that it is possible to execute randomized trials that use organizations or other entities as the units of random allocation in trials that permit fair comparisons. In what follows, we give evidence on the feasibility of such trials

3.1 Schools, School Districts, and Classrooms as the Units of Random Assignment

Schools and classrooms, for instance,

have been randomly assigned to different approaches in educating children about avoiding substance abuse (Schaps et al., 1982; Moskowitz, 1984; Botvin et al., 1995; Murray, Moskowitz, and Dent, 1996). In tests of the Drug Abuse Resistance Education (D.A.R.E.) program in Illinois, for example members of 12 pairs of schools were randomly assigned to different programs in the interest of fair comparison (Rosenbaum et al., 1991). Other entity-based experiments on this program were reviewed by Ennett et al. (1994). The Flay et al (1985) work in Canada is a remarkable precedent in this arena.

In efforts to evaluate a theory-driven program to reduce alcohol use by underage youth, Wagenaar et al. (1994) mounted a randomized field trial involving 15 school districts.

Seven of the willing districts in Minnesota and Wisconsin were randomly assigned to employ a special community-based prevention program. Eight of the willing districts were randomly assigned to the control group.

Schools have also been the units in at least two smoking prevention experiments. The Television, School and Family Smoking Prevention Project, used multi-attribute balancing to randomly assign 35 Los Angeles area schools to different media-based smoking prevention campaigns. Flay et al. (1985) randomly assigned 22 matched schools to experimental and control conditions in the Waterloo Study, a Canadian smoking prevention effort. Tests of school-wide cardiovascular risk reduction programs for children have been undertaken. For example, schools have been randomly assigned to such programs in four states (Killen et al., 1988; Hansen and Graham, 1991; and Perry et al., 1992).

In a mobile societies, it is important to understand how to reduce the psychological and educational risk of children who are moved from one education context to another. Jason et al. (1992, 1993a, 1993b) focused on children who transferred into new schools and who were, as a consequence, vulnerable. One project involved randomly assigning members of ten matched pairs of schools to an innovative treatment program or to a control condition in order to determine whether their special transition program worked.

Until the late 1990s, high quality evaluations of violence reduction programs in schools were rare. Among the notable exceptions is the Grossman et al. (1997) study of the effectiveness of violence prevention curricula for second and third graders. Six matched pairs of schools were randomly assigned to employ the curriculum or to serve in a control group. Differences in children's behavior were discernible and persisted for at least six months.

Until the 1970s, no controlled-field experiments of any scale appear to have been run to understand the effects of standardized testing on students in any country. In 1975, the Irish Republic decided to consider for the first time standardized testing for children in the Republic's elementary schools. Kellaghan, Madaus, and Airasian (1982) and their colleagues at St. Patricks' College (Dublin) mounted a study in which 175 eligible schools, matched and stratified, were allocated randomly to different conditions. The control condition involved

no standardized testing. The intervention was standardized testing, with and without feedback to teachers, on student performance.

Randomized trials have been mounted to understand what kinds of programs might be deployed in education settings as to enhance children's understanding of high risk sexual behavior and how to avoid it. In the U.S. for example, Gay's (1996) dissertation research involved matching eight middle school classrooms and allocating half to a new Red Cross program and half to a control condition in which no such education effort existed. In the Philippines, Alpasca et al. (1995) also targeted classrooms within schools. In a large-scale trial in California, Kirby et al. (1997a) randomly assigned 102 classrooms in six middle schools to a theory-driven risk prevention program that relied heavily on young »peer education« to implement the program. Another California based program, Postponing Sexual Involvement (PSI) was evaluated using a complex research design in which classrooms were randomized in one component (Kirby et al., 1997b). Over 50 schools were involved.

A different stream of health related work has concerned nutrition education. Woodruff (1997), for instance, described a San Diego experiment that involved eight intervention classes and nine control classes being randomly assigned to a new nutrition program from three community colleges.

Earlier examples to test different approaches in different countries to enhancing children's achievement deserve recognition. Consider examples from Nicaragua, El Salvador, and the U.S. Classrooms in Nicaragua have been randomly assigned to radio-based mathematics education and to conventional education so as to learn whether the former would enhance mathematics achievement and reduce education costs relative to the latter (Dean et al., 1981; Jamison, Searle, and Suppes, 1980). A similarly designed randomized trial in El Salvador disintegrated; Hornik et al. (1972) gave an admirably candid description. During the 1970s, the U.S. Department of Education sponsored a large scale study to understand whether funding could be effectively employed by schools to reduce racial isolation and enhance the achievement of students. Eligible schools that were willing to participate in the experiment were randomly allocated to a special funding opportunity and to a control group that received no special treatment. See Coulson (1978), Reichardt and Rindskopf (1978), and Weissberg (1978).

3.2 Communities and Geopolitical Entities as the Units of Random Assignment

In a study of how to encourage voter registration in Chicago, Gosnell (1927) appears to have randomly assigned distinct neighborhoods in political precincts to treatment and control conditions. The »treatment« involved publicity, mail, and in-person contacts, provided at times in different languages to diverse ethnic neighborhoods. The intent was to provide information about voter registration and to encourage registration in different ways, and to test the treatment.

Communities have been the units of

allocation in evaluations of health related programs. LaPrelle, Bauman, and Koch (1992), for instance, reported on a study of the relative effectiveness of three media campaigns to prevent cigarette smoking among adolescents. They screened, matched and then randomly assigned communities from a sample of ten communities to one of three treatments and to a control group. The Community Intervention Trial for Smoking Cessation (COMMIT), assigned eleven matched pairs of communities to its treatment and comparison groups (Freedman, Green, and Byar, 1990 cited in Peterson et al., 1992).

In randomized trials on fertility interventions in the Far East, communities and villages have been randomly assigned to different approaches to understand how to decrease birth rates (Freedman & Takashita, 1969; Riecken et al., 1974). Small numbers of communities have also been used as units in randomized studies of HIV risk prevention tactics (Kelly et al., 1991).
In media-based smoking prevention campaigns, standard metropolitan statistical areas (SMSAs) have been allocated randomly to the campaigns or to control conditions (Bauman et al., 1991). Federal statistical agencies specify these SMSA geographic areas in a uniform way so as to make clear what is meant by »metropolitan area« in contrast to a rural area, for example, and use these areas to design the census and national surveys. Education studies in Cali, Colombia involved randomly assigning very small geographic areas in the low-income barrios to a cultural enrichment and health enhancement program for preschoolers to determine its effect relative to randomly assigned control areas (McKay et al., 1978).

Some randomized trials have been mounted because the integrating multiple services at the community level are thought to be important to people who are mentally ill and live in the community. Access to Community Care and Effective Service Supports (ACCESS) involved eight cities, each of which contained two independent jurisdictions that were randomly assigned to the ACCESS or to the control condition (Randolph et al., 1997). About 50 agencies within each jurisdiction cooperated on the study.

Finally, consider early research on crime prevention. In the Kansas City patrol experiment, fifteen police beats were matched and randomly divided into three groups of five beats each. This precedent compared the relative effects of reactive, proactive, and control (normal) patrols on victimization (Kelling, Pate, Dieckman, and Brown, 1974). Twenty years later, Sherman and Weisburd (1995) executed a better-randomized trial in Minneapolis. The researchers identified over 100 »hot spots«, local areas of predictably high crime and randomly allocated half of these areas to more intensive police patrol or to a normal patrol activity.

3.3 Other Private and Public Organizations as Units of Random Assignment

In some countries, a sensible way to enhance the well-being of individuals is through private organizations. Programs

designed to reduce the risk of sexually transmitted diseases, for example, might be more effective if the program is directed toward all the workers in corporate factories rather than to individuals who may or may not work in the factories. It is partly for this reason that the National Institute of Allergies and Infectious Diseases in the U.S. has invested in tests of factory-based peer education (NIAID, 1997). No one knows whether peer education among factory workers will reduce infection. The project involved some 40 factories in Zimbabwe, half being randomly assigned to programs designed to reduce incident HIV infection and the remaining to a control condition. Other randomized trials have used work sites as units in assessing nutrition programs and weight control and smoking cessation programs (Simpson et al., 1995).

Non-profit service organizations have, at times, committed resources to randomized trials. For instance, Good Will Industries in the U.S. agreed to participate in controlled experiments on how to improve the management of the organization's stores (Glaser et al., 1967). In this instance, independent stores were the units of allocation.

In the medical arena, nearly forty Minnesota community hospitals agreed to participate in a trial to discover whether local medical opinion leaders and a formal feedback system could influence the rate at which the hospitals adopted new beneficial therapies for acute myocardial infarction patients (Soumerai et al., 1998). The theory underlying the program is that the entire hospital staffs' understanding, not just the physician's education, together with the monitoring of therapy, are necessary to produce change. Hence, allocating hospital physicians randomly to a program was not sensible. The trial's design involved the random allocation of 20 hospitals to this approach to clinical education and random allocation of 17 hospitals to a control condition.

Our final illustration involves a program designed to enhance employment of individuals at high risk of unemployment who live in low-income public housing developments in communities that need economic revitalization. In each of seven cities, the trials involved the random allocation of one public housing facilities to the program and one or two public housing facilities to a control condition. The presumptions underlying the program's design were that local collaboration and collective decisions are essential in transforming local communities in ways that affect, among other things, education, training and employment, and wage rates (Riccio, 1998; Bloom, Bos, and Lee, 1998).

4. Difficulties and Possible Resolutions

Challenges to using places or other entities as the units of allocation in a randomized trial are numerous. Strategies that have been invented to surmount obstacles are valuable and discussed in what follows.

4.1 Statistical Power

Consider a randomized field trial in which two literacy programs are compared to one

another to establish which is more effective and less costly. Statistical power refers to our ability to discern the relative effectiveness of the two literacy programs. This power depends, of course, on how literacy is measured. It also depends on how many literacy centers are randomly allocated to one or the other literacy program and on how many students there are in each program. The »statistical power« refers to our ability to detect a difference in the effects of the interventions if indeed there is a difference.

How many centers might be required in this experiment to assure that its statistical power is about .80? Assume, as is likely, that the true difference between the programs is small (.10) and fix the statistical threshold (alpha) at .05. If all the students within schools were independent, about 400 students for each plan would have to be sampled to discern the effect of the treatments under these conditions.

When the similarity among students within a school is substantial, a larger sample size will be necessary to assure that real differences between the intervention is detected. Assuming a low similarity rate (intraclass correlation) of .05, one might then use 85 schools with a sample of 10 students each, for each treatment (program) in a formal test. Or, one may use 44 schools with 40 students each.

In the opinion of LaPrelle et al. (1992), their trial on community-based substance use prevention in citywide programs was underpowered. Four treatments in an experiment were spread over ten communities. Their thoughtful post-trial analysis suggested that about 40 communities per group would have been required to detect an important difference in the effectiveness of smoking prevention programs.

Place-based randomized trials have relied successfully on at least three tactics to assure adequate statistical power. First, entities that are independent should be screened for eligibility and a reasonable level of homogeneity. Second, the entities should be matched and then randomized. A third tactic is implicit: engage as many entities as possible in the trial.

4.2 Measurement Systems and Theory

By a theory of »what should happen,« we mean laying out the way that the programs being compared are each expected to engage and affect the entities. That is, the logic of how the thing is supposed to work needs to be made plain. More to the point, the theory guides us in selecting what should be measured and, at its most sophisticated, whether and how well it might be measured.

Consider the multi-site Wagenaar et al. (1997) trial. It was designed to understand whether a community-based program could reduce the use of alcohol by underage youth. Mobilization of communities was regarded as theoretically important to creating alcohol use policy. Observations then were made of community power structures and the attitudes of students and youth. Analyses were undertaken of media coverage. Changes in community practice were also measured on the supposition that these would follow community mobilization. Among other efforts, this

stage included surveys of retail alcohol outlets to determine if indeed they failed to ask proof of the age of customers whose appearances were youthful. This was done because, in theory, decreasing youth access to alcohol would result in fewer alcohol-related traffic accidents. Further, the latter were assessed using state and local record systems.

4.3 Engaging Sites and other Entities

Engaging sites, administrative units, and other entities in a randomized field trial requires considerable skill. Walker et al (2000) provide an exceptionally detailed description of strategies for recruiting U. K. Hospitals into randomized trials. They focus attention on identifying stakeholders and gatekeepers, informing them, approaching gatekeepers to engage the hospital, negotiating the terms of engagement, conducting the study, and providing feedback of different kinds to gatekeepers and stakeholders. The process is time consuming an challenging. To judge from researchers success in mounting such trials. The strategies are worth serious consideration.

Consider next, Ellickson's (1994) paper on the conduct of Project ALERT, which involved 30 schools being randomly assigned to ALERT or to a control condition. Its object was to determine how well the ALERT project worked to prevent substance abuse among children and how long the project's effects last. Recruiting entire schools into a RFT must recognize natural limits on their capacity to participate. Ellickson (1994) reported that eleven schools out of about 60 schools that were invited to participate declined to do. One school, for instance, could not participate on account of a court order demanding considerable resource allocation on racial equity. Four of the eleven schools declined to participate because they already had prevention programs in place. The reasons for other declinations concerned their capacity, e.g., inability to assure community support for engaging in the experiment.

4.4 Temporal and Structural Stability

We expect sites not to change much over a short period of time. Nonetheless, the stability of certain characteristics of sites may be low or trends may reverse direction. Bauman et al. (1991), for example, found high positive correlation over a two-year period (r = .77) for adolescents' reported rates of recent smoking in a sample of 10 cities. The researchers found a negative correlation (r=-.31) for adolescents' rates of experimentation with smoking in the same cities. Reasons for this finding are unclear. The instability is clear.

One normally assumes that the places or other entities that are targeted for a program will be structurally stable over the study's course. A school in year 1, for instance, is expected to be a school in year 2. To judge from experience, it is prudent to anticipate some change. For example, the Midwestern Prevention Project involved randomly assigning schools to different conditions. Pentz (1994) reported that 8 of the initial 50 targeted middle schools and high schools »closed or consolidated with

other schools over the first three years of the study« (p. 44). Further, feeder schools changed as a consequence of changes in busing patterns and the creation of magnet schools that drew students from areas outside the original catchment area schools.

Similar problems have occurred elsewhere. In the Irish Standardized Testing experiment, after matching and randomly assigning schools based on census data, the researchers found that many important school characteristics had changed (Kellaghan, Madaus, and Airasian, 1982). Tennessee's experiment on school incentives encountered difficulties because schools were closed or consolidated with other schools (Bickman, 1985). All this engenders complex problems in designing randomized trials and in their analysis.

4.5 Regional Variation

To produce a good estimate of the effect of smoking prevention programs, Bauman et al. (1991) focused attention on only one geographic region. Despite this attempt to work in a homogeneous context, the experiment was underpowered. That is, the sample of organizations within the region may have been too small to discern a real effect of programs because there was considerable variation within the region. For instance, the rates of recent smoking among adolescents across ten cities in one region reported by Bauman et al. (1991) were in the range 2-7% in 1985 and 13-20% in 1987. Rates of smoking in 1987 among 1985 nonsmokers were in the range of 3-14% across the cities.

Stratification or blocking by region in a place-based trial makes sense. But the definitions of region and the implications of a choice have not been investigated deeply. In any event, reconnaissance prior to mounting a randomized experiment—pilot tests and analyses of extant date—are warranted.

4.6 Unbalanced Groups and Restricted Randomization

Consider a randomized trial in which a sample of communities that is provided with increased literacy resources is compared to a sample of communities that has been allocated to a waiting list, i.e., have not yet been given the resources. The number of communities involved in such a study must often be relatively small, say 20 to 40, in each of the groups. For the analyst, this raises a concern that the two groups that are randomly composed will not be »equivalent« at the outset. That is, there is an imbalance between the groups that is attributable to chance. This »unhappy random configuration« will complicate comparisons. One approach used to reduce the problem in multi-site RFTs is restricted randomization.

In restricted randomization, some configurations of the random allocation of sites to different treatments are defined as undesirable a priori. That is, all possible randomized configurations under a particular experiment's design are laid out beforehand. The »unhappy« ones are then eliminated from eligibility. A random selection is then made from the remaining eligible configurations. For the applied researcher, constraining the randomization

options to sensible configurations prevents badly unbalanced groups of institutions from being assigned to different program variations. For instance, Ellickson and Bell (1992) linked »unlike schools from districts into pairs and randomly (assigned) the pairs to the experimental conditions...« to achieve balance (p. 85).

The implication is that when a small number of sites are the units of allocation in randomized trials, we can enumerate all possible allocations of sites in advance of the trial. Further, we can eliminate the possible allocations that are strange, out-of-line, and so on. Having eliminated the allocations that are out-of-line, we can randomly select a configuration, allocate institutions in accord with it, and develop a comparison of programs that is fair.

4.7 Implementation Fidelity and Measurement

It makes no sense to estimate the effect of a new program unless one can verify that the program activities occur and can be described. »Implementation fidelity« here refers to the degree to which a new program is actually delivered to target individuals. Its measurement refers to observing indicators of fidelity. We need to determine whether administrative actions are taken, information systems are emplaced, and so on. Learning that actions are indeed taken is a prerequisite for any impact evaluation.

Trials that attempt to evaluate interventions that involve »integration« or »coordination« of services across many agencies within an organization or community present special problems. Developing a coherent definition of integration and measurable indicators of integration is not easy. Consider studies of ACCESS' effect on the homeless and mentally ill, for instance. The various jurisdictional units may differ on: whether and how they employ interagency coalitions; interagency teams for service delivery; interagency management systems; interagency agreements and memorandums of understanding; finding arrangements; eligibility standards; and co-location of services (Randolph et al., 1997). Learning how to observe any of these reliably and to assure fidelity in implementation and its measurement is demanding.

Bibliography[2]

Aplasca, M., Siegel, D., Mandel, J. S., Santana, R., Paul, J., Hudes, E. S., Monzon, O. T, and Hearst, N. (1995) Results of a Model AIDS Prevention Program for High School Students in the Philippines. AIDS, Supplement 1, 7-13. (*)

Basen-Engquist, K., Parcel, G. S., Harrist, R., Kirby, D., Coyle, K., Banspach, S., and Rugg, D. (1997) The Safer Choices Project: Methodological Issues in School Based Health Promotion Intervention Research. Journal of School Health, 67(9), 365-371. (*)

Bauman, K. E., LaPrelle, J., Brown, J. D., Koch, G.

[2] The references in the bibliography that are marked with an asterisk (*) report on trials that involve places, organizations, and groups or other entities as the units of random allocation in randomized trials.

C. and Padgett, C. A. (1991) The Influence of Three Mass Media Campaigns on Variables Related to Adolescent Cigarette Smoking: Results of a Field Experiment. American Journal of Public Health, 1991, 81, 597-604. (*)

Bickman, L. (1985) Randomized Field Experiments in Education. New Directions for Program Evaluation, 28, pp. 39-54. (*)

Bloom, H., Bos, J. and Lee, S. W., (1998) Using Cluster Random Assignment to Measure Program Impacts: Statistical Implications for the Evaluation of Education Programs. New York: New York University, Robert F. Wagner School of Public Service (Research Report). (*)

Boruch, R. F. (1993(a) Multi-site Tests in the Civil and Criminal Justice Arena. Invited Presentation, Annual Meeting of the American Society of Criminology (October 30, 1993) Phoenix, Arizona. Available from: Author. University of Pennsylvania, Philadelphia, PA 19104. (*)

Boruch, R. F. (1993(b)) Multi-site Evaluation and the Children's Initiative. Paper prepared for the Pew Charitable Trusts, Philadelphia, PA. Available from: Author. University of Pennsylvania, Philadelphia, PA 19104. (*)

Boruch, R. F. (1997) Randomized Experiments for Planning and Evaluation: A Practical Guide. Thousand Oaks, CA: Sage.

Boruch, R. F. and Foley, E. (2000) The Honestly Experimental Society: Sites and Other Entities as the Units of Allocation an Analysis in Randomized Trials. In L. Bickman (Ed.) Validity and Experimentation: Donald Campbell's Legacy Volume 1. Thousand Oaks, CA, London, New Delhi: Age Publications.

Botvin, G. J., Baker, E., Dusenburg, L., Botvin, E. M., and Diaz, T. (1995) Long Term Follow-up Results of a Randomized Drug-Abuse Prevention Trial in a white Middle Class Population. Journal of the American Medical Association, 273, 1106-1112. (*)

Campbell, D. T. (1969) Reforms as Experiments. American Psychologist, 24(4), 408-429. (*)

Campbell, D. T. (1988) The Experimenting Society. Chapter 11 in S. Overman (Ed.) Methodology and Epistemology for Social Science: Selected Papers by Donald T. Campbell. Chicago: University of Chicago Press, pp. 290-314.

Campbell, D. T. and Stanley, J. C. (1963) Experimental and Quasi-experimental Designs for Research Teaching. In N. L. Gage (Ed) Handbook of Research on Teaching. Chicago, IL: Rand McNally, pp 171-246.

Coulson, J. E. (1978) National Evaluation of the Emergency School Aid Act (ESAA): A Review of Methodological Issues. Journal of Educational Statistics, 3(3), 1-60. (*)

Coyle, S. L., Boruch, R. F., and Turner, C. F. (Eds.) (1991) Evaluating AIDS Prevention Programs (Expanded Edition). Washington, DC: National Academy of Sciences. (*)

Coyle, K, Kirby, D., Purcel, G., Basen-Engquist, K., Banspach, S, Rugg, D., and Well, M. (1996) Safer Choices: A Multicomponent School Based HIV/STD and Pregnancy Prevention Program for Adolescents. Journal of School Health, 66(3), 89-84. (*)

Dean, J., Seare, B., Galda, K., and Heyneman S. P. (1981) Improving Elementary Mathematics Education in Nicaragua: An Experimental study of the Impact of Textbooks and Radio on Achievement. Journal of Education Psychology, 73(4), 556-567. (*)

Donner,. A. and Klar, N. (2000) Design and Analysis of Cluster Randomized Trials in Health Research. New York: Oxford University Press.

Ellickson, P. L. (1994) Getting and Keeping Schools and Kids for Evaluation Studies. Journal of Community Psychology (Monograph Series: CSAP Special Issue), pp. 102-116. (*)

Ellickson, P. L. & Bell, R. M. (1992) Challenges to Social Experiments: A Drug Prevention Example. Journal of Research in Crime and Delinquency, 29(1), pp. 79-101. (*)

Ellickson, P. L. & Bell, R. M. (1990) Drug Prevention in Junior High: A Multi-site Longitudinal Test. Science, 247, pp. 1299-1305. (*)

Ennett, S. T., Tobler, N. S., Ringwalt, C. L., and Flewelling, R. L. (1994) How effective is Drug Abuse Resistance Education? A Meta-analysis of Project DARE's Outcome Evaluations.

American Journal of Public Health, 84(9), 1394-1401. (*)

Fairweather, G. W., Sanders, D. H., & Tornatsky, L. G. (1974) Creating Change in Mental Health Organizations. New York: Pergamon. (*)

Fairweather, G. W. and Tornatzky, L. G. (1977) Experimental Methods for Social Policy Research. New York: Pergamon Press. (*)

Farrington, D. P. (1997) Evaluating a Community Crime Prevention Program. Evaluation, 3. (*)

Flay, B. R., Ryan, K. B., Best, J. A., Brown, K. S., Kersell, M. W., d'Avernas, J. R. & Zanna, M. P. (1985) Are Social-psychological Smoking Prevention Programs Effective? The Waterloo Study. Journal of Behavioral Medicine, 8(1), pp. 37-59. (*)

Freedman, R. & Takashita, J. T. (1969) Family Planning in Taiwan: An Experiment in Social Change. Princeton, NJ: Princeton University Press. (*)

Gay, K. E. M. (1996) Collaborative School-based Research: The Creation and Implementation of an HIV/AIDS Prevention Curriculum for Middle School Students. PhD Dissertation, University of Pennsylvania, Philadelphia, PA. (*)

Glaser , E. M., Coffey, H. A., and others (1967) Utilization of Applicable Research and Demonstration Results. Los Angeles, CA: Human Interaction Research Institute.(*)

Gosnell, H. F. (1927) Getting Out the Vote: An Experiment in the Stimulation of Voting. Chicago: University of Chicago Press. (*)

Green, S. B. and Washington, A. E. (1991) Evaluation of Behavioral Interventions for Prevention and Control of Sexually Transmitted Diseases. In: J. N. Wasserheit, S. O., Aral, K. K., Holmes, and P. J. Hitchcock (Eds.) Research Issues in Human Behavior and Sexually Transmitted Diseases in the AIDS Era. Washington, D.C.: American Society for Microbiology, pp. 345-352.

Grossman, D. C., Neckerman, H. J., Koepsall, T. D., Liu, P., Asher, K. N., Beland, K., Frey, K., and Rivara, F. P. (1997) Effectiveness of a Violence Prevention Curriculum among Children in Elementary School: A Randomized Controlled Trial. Journal of the American Medical Association, 277(20), 1605-1611, (*)

Hansen, W. B. & Graham, J. W. (1991) Preventing alcohol, marijuana and cigarette use among adelscents; Peer pressure resistence training versus establishing conservative norms. Preventive Medicine, 20, 414-430.

Hornik, R. (1991) Alternative Models of Behavior Change. In J. N. Wasserheit, S. O., Aral, K. K. Holmes, and P. J. Hitchcock (Eds.) Research Issues in Human Behavior and Sexually Transmitted Diseases in the AIDS Era. Washington, D.C.: American Society for Microbiology, pp. 201-218.

Hornik, R. C., Ingle, H.T., Mayo, J. K., McAnany, E. G., and Schramm, W. (1972) Television and Education Reform in El Salvador. (Report No. 14) Stanford University, Institute for Communication Research. (*)

Jamison, D., Searle, B., & Suppes, P. (1980) Radio Mathematics in Nicaragua. Stanford, CA: Stanford University Press. (*)

Jason, L. A., Weine, A. M., Johnson, J. H., Donner, K. E., Kuraski, K. S., & Sohlberg, L. (1993a). The school transitions project: A comprehensive preventive intervention. Journal of Emotional and Behavioral Disorders, 1(1), pp. 65-70. (*)

Jason, L. A., Weine, A. M., Johnson, J. H., Sohlberg, Filippelli, Turner, E., & Lardon, C. (1992) Helping Transfer Students: Strategies for Educational and Social Readjustment. San Francisco: Jossey-Bass. (*)

Jason, L., Johnson, J. H., Danner, K. E., Taylor, S., and Krasaki, K. S. (1993b) A Comprehensive, Preventive, Parent-Based Intervention for High Risk Transfer Students. Prevention in Human Services, 10(2), 27-37. (*)

Kaftarian, S. J. & Hansen, W. B. (1994) (Eds.) Community Partnership Program: Center for Substance Abuse Prevention. CSAP Special Issue/Monograph Series. Journal of Community Psychology. (*)

Kellaghan, T., Madaus, G. F., Airasian, P. W. (1982) The Effects of Standardized Testing. Boston/ The Hague/London: Kluwer-Nijhoff. (*)

Kelling, G. L., Pate, T., Dieckman, D., & Brown, C. E. (1974) The Kansas City Preventive Patrol Experiment: A Summary Report. Washington, DC: Police Foundation. (*)

Kelly, J.A., Lawrence, J.S., Diaz, Y. E. and others. (1991) HIV Risk Behavior reduction Following Intervention with Key Opinion Leaders: An Experimental Analysis. American Journal of Public Health,81, 168-171. (*)

Killen, J. D., Telch, M. J., Robinson, T. N., Maccoby, N., Taylor, C., & Farquar, J. W. (1988) Cardiovascular Disease Risk Reduction for Tenth Graders: A Multiple Factor School-based Approach. Journal of the American Medical Association, 260(12), pp. 1728-1733. (*)

Kirby, D., Korpi, M., Adivi, C. and Weismann, J. (1997a) An Impact Evaluation of Project SNAPP: An AIDs Prevention and Pregnancy Middle School Program. AIDS Education and Prevention, 9 (Supplement A), 44-61. (*)

Kirby, D., Korpi, M, Barth, R. P., and Cagampang, H. H.(1997b) The Impact of Postponing Sexual Involvement Curriculum among Youths in California. Family Planning Perspectives, 29, 100-108. (*)

Kuusi, Pekka (1957) (WestPhaler, A. Translator). Alcohol Sales in Rural Finland. Volume 3 Publication of the Finish Foundation for Alcohol Studies. Stockholm, Sweden: Almqvist and Wiksell.

LaPrelle, J., Bauman, K. E. & Koch, G. G. (1992) High intercommunity variation in adolescent cigarette smoking in a 10-community field experiment. Evaluation Review, 16(2), pp. 115-130. (*)

Leviton, L., Valdiserri, R., Lyter, D., Callahan, C., Kingsley, L., Huggins, J., and Rinalde, C. R. (1990) Preventing HIV Infection in Gay and Bisexual Men: Experimental Evaluation of Attitudes Changes from Two Risk Reduction Experiments. AIDS Education and Prevention, 2(2), 95-108. (*)

Lorian, R. P. (1994) Epilogue: Evaluating the Community Partnership Program. Reflections on a Name. Journal of Community Psychology (Monograph Series: CSAP Special Issues), pp. 201-205. (*)

McKay, H., McKay, A., Sinnestera, L., Gomez, H., and Lloreda, P. (1978) Improving Cognitive Ability in Chronically Deprived Children. Science, 200(4), 270-278. (*)

Moskowitz, J. et al. (1984) The Effects of Drug Education and Follow-up. Journal of Alcohol and Drug Education, 3, pp. 45-49. (*)

Murray, D. (1998) Design and Analysis of Group Randomized Trials. Oxford and New York: Oxford University Press.

Murray, D. M., McKinlay, S. M., Martin, D., Donner, A. P., Dwyer, J. H., Raudenbush, S. W., & Graubard, B. I. (1994). Design and Analysis Issues in Community Trials. Evaluation Review, 18(4), pp. 493-514. (*)

Murray, D. M. and Wolfinger, R. D. (1994) Analysis Issues in the Evaluation of Community Trials: Progress Toward Solutions in SAS/STAT Mixed. Journal of Community Psychology (Monograph Series: CSAP Special Issue), pp. 140-154. (*)

Murray, D., Moskowitz, J. M., and Dent, C. W. (1996) Design and Analysis Issues in Community-Based Drug Abuse Prevention. American Behavioral Scientist, 39(7), 853-867. (*)

Pentz, M. A. (1994) Adaptive Evaluation Strategies for Estimating the Effects of Community Based Drug Abuse Prevention Programs. Journal of Community Psychology (Monograph Series CSAP Special Issue), pp. 5-25. (*)

Perry, C., Parcel, G. S., Stone, E., Nader, P., McKinlay, S. M., Leupker, R. V., and Webber, L. S. (1992) The Child and Adolescent Trial for Cardiovascular Health (CATCH): An Overview of Intervention Program and Evaluation Methods. Cardiovascular Risk Factors, 2(1), pp. 36-43. (*)

Peterson, P. L., Hawkins, J. D., & Catalano, R. F. (1992) Evaluating Comprehensive Community Drug Risk Reduction Interventions. Evaluation Review, 16(6), pp. 579-602. (*)

Randolph, F, Basinsky, M., Leginski, W., Parker, L., and Goldman, H. H. (1997) Creating Integrated Service Systems for Homeless Persons with Mental Illness: The Access Program. Psychiatric Services, 48(3), 369-373. (*)

Reichardt, C. S. & Rindskopf, D. (1978) Randomization and Educational Evaluation: The ESAA Evaluation. Journal of Educational Statistics, 3(1), 61-68. (*)

Reiss, A. J. & Roth, J. A. (Eds.) (1993) Understanding and Preventing Violence. Washington, DC: National Academy of Sciences Press.

Riccio, J. A. (1998) A Research Framework for Evaluating Jobs-Plus, A Saturation and Place-Based Employment Initiative for Public Housing Residents (Working Paper). New York, Manpower Demonstration Research Corporation. (*)

Riecken, H. W., Boruch, R. F., Campbell, D. T., Caplan, N., Glennan, T. C., Pratt, J. W., Rees, A., & Williams, W. (1974) Social Experimentation: A Method for Planning and Evaluating Social Programs. New York: Academic Press. (*)

Rosenbaum, D. P., Ringwalt, C., Curtin, T. R., Wilkinson, D., Davis, B., & Taranowski, C. (1991) Second Year Evaluation of D.A.R.E. in Illinois. (Available from: D. P. Rosenbaum Center for Research in Law and Justice, University of Illinois at Chicago, Chicago, Illinois 60607). (*)

Schaps, E., Moskowitz, J., Condon, J., & Malvin, J. (1982) A Process and Outcome Evaluation of a Drug Education Course. Journal of Drug Education, 12, pp. 245-454. (*)

Shadish, W. R., Cook, T. D., and Campbell, D.T. (2002) Experimental and Quasi-experimnetal Designs for generalized Causal Inference. New York: Houghton Mifflin.

Sherman, L. and Weisburd, D. (1995) General Deterrent Effects of Police Patrol in Crime »Hot Spots«: A Randomized Controlled Trial. Justice Quarterly, 12(40, 625-648. (*)

Simpson, J. M., Klar, N. and Donner, A. (1995) Accounting for Cluster Randomization: A Review of Primary Prevention Trials, 1990 through 1993. American Journal of Public Health, 85(10), 1378-1383. (*)

Soumerai, S. B., McLaughlin, T. J., Gurwitz, J. H., Guadgnoli, E., Hauptman, P. J., Borbas, C., Morris, N., McLaughlin, B., Gao, X., Willison, D. J., Asinger, R. and Gobel, F. (1998) Effect of Local Medical Opinion Leaders on Quality of Care for Acute Myocardial Infarction. Journal of the American Medical Association, 279(17), 1358-1363. (*)

Soydan, H. (1998) (Issue Editor) Evaluation Research and Social Work. Scandinavian Journal of Social Welfare, 7 (2).

Wagenaar A. C., Murray, D. M., Wolfson, M., Forster, J. L., & Finnegan, J. R. (1994) Communities mobilizing for Change on Alcohol: Design of a Randomized Community Trial. Journal of Community Psychology (Monograph Series/CSAP Special Issue), pp. 79-101. (*)

Wagenaar, A. C., Murray, D. M., Gehan, J. P., Wolfson, M., Forster, J. L., Toomey, T. L., Perry, C. L., and Jones-Webb, R. (1997) Communities Mobilizing for Change on Alcohol (CMCA): Outcomes from a Randomized Trial. Report. University of Minnesota. Submitted. (*)

Walker, A. E., Campbell, M. K., Grimshaw, J. M.,, and the TEMPEST Group (2000) A Recruitment Strategy for Cluster Randomized Trials in Secondary Care settings. Journal of Evaluation in Clinical Care Settings, 6(2), 185-192.

Wasserheit, J. N., Aral, S. O., Holmes, K. K., and Hitchcock, P. J. (Eds.) (1991) Research Issues in Human Behavior and Sexually Transmitted Diseases in the AIDS Era. Washington, D.C.: American Society for Microbiology. (*)

Weisberg, H. (1978) How Much does ESAA Really Accelerate Academic Growth. Journal of Educational Statistics, 3(1), 69-78. (*)

Weisburd, D., Sherman, L., & Petrosino, A. J. (1990) Registry of Randomized Criminal Justice Experiments in Sanctions. Washington, DC: National Criminal Justice Reference Service (SRO 19000-00/129725).

Woodruff, S. I. (1997) Random Effects Models for Analyzing Clustered Data from a Nutrition Education Intervention. Evaluation Review,21 (6), 688-697. (*)

Summary

Randomized trials have yielded good evidence about which programs work better, for whom, and how long in medicine, criminology, welfare reform, education and other sectors. Trials that involve the random assignment of places such as communities, housing projects, organizations, neighborhoods, schools or other entities, to different interventions so as to generate fair comparison are not yet common. But they can be justified for theoretical, statistical, policy, political and ethical reasons.

The theoretical rationale for place-based trials is that programs work when organizational elements in a place concert, e.g., community-wide programs. A basic statistical rationale for focusing on places or institutions as the units of random allocation in a trial is that conventional statistical analyses of the effect of programs can be wrong when analyses are based on individuals rather than on institutions.

The policy and political rationale for focusing on organizations and other sites as the units for study is that organizations are the immediate target for a government agency and foundation action. Individuals are not. The ethical and cultural rationale is that, at times, the random allocation of organizations to alternative regimens, in the interest of a fair comparison, is more acceptable and desirable than random assignment of individuals.

The feasibility of using places, and other entities as units in controlled randomized trials is demonstrable. Entities have been allocated at random to different interventions in trials on fertility control methods, welfare enhancement, education reform, law enforcement, health-risk reduction programs and others. The units of random allocation have been neighborhoods, factories, classrooms and schools, hospitals, saloons, and so on.

There are difficulties in executing such trials, of course. Able administrators, researchers, civil servants, and foundation people have met the challenges at times. Statisticians and methodologists who understand the design of place based randomized trials can tailor the trials design at times so as to meet the challenges.

Regardless of the difficulties, the future of place-based randomized trials is promising. They are being run more frequently. Place-based trials have been mounted in diverse areas such as education, crime and delinquency, mental health, employment, health risk redution and welfare. They are an important tool in generating evidence about which programs work and for whom, which do not work, and which programs are promising.

En investering i socialt arbete

kari jess & siv nyström

Är insatser som leder till förbättringar för klienterna också samhällsekonomiskt lönsamma? Det är en av de frågor som behandlas i denna artikel där resultatet av en klienteffektstudie jämförs med resultatet av en samhällsekonomiska utvärdering av samma insatser och samma klientgrupper.

Inledning

Brottslighet och missbruk förorsakar samhället stora problem och för också med sig ett personligt lidande för dem som själva begår brott eller missbrukar. För att lösa problemen erbjuder kriminalvård, socialtjänst och andra offentliga aktörer olika slag av behandling och rehabilitering. Några sådana samhällsinsatser är KrAmi, Knuff och frivården som alla arbetar med att hjälpa sina klienter tillbaka till ett socialt accepterat liv. Dessa tre står också i centrum för de utvärderingar som här kommer att diskuteras.

KrAmi-programmen riktar sig till unga frivårdsklienter som behöver hjälp med att komma in på arbetsmarknaden. Förutom problem med arbetslöshet, kriminalitet och brott, har många klienter också problem med missbruk. KrAmi-programmen finns i flera större svenska kommuner, bland andra Stockholm, Göteborg, Malmö, Örebro, Kalmar och Västerås. Programmet representerar en för svenska förhållanden ovanlig form av samverkan mellan kriminalvård, socialtjänst och arbetsförmedlingen. Personal från de tre samhällssektorerna arbetar tillsammans med utgångspunkt från en och samma programidé och med ett gemensamt arbetssätt, KrAmi-metoden.

KrAmi-metoden har hämtat inspiration från de danska träningsskolornas konsekvenspedagogik. Den bygger på tanken att unga kriminella behöver lära sig att förstå sambandet mellan sina handlingar och deras konsekvenser. Därför bygger programmet på tydliga valsituationer i vilka klienterna tillsammans med ledarna får möjlighet att reflektera över olika handlingars negativa och positiva konsekvenser. Det kontrakt med enkla regler, som upprättas mellan program och klienter tydliggör valsituationen.

Siv Nyström är filosofie doktor i pedagogik och arbetar som forskare vid Centrum för utvärdering av socialt arbete.
Kari Jess är doktorand på Institutionen för socialt arbete vid Stockholms universitet och verksam vid Mälardalens högskola.

Kontraktet, ledarnas och gruppens stöd och uppmuntran hjälper klienterna att lyckas i ovana sociala situationer. Centrala roller fyller den sociala träningen och praktikplatserna. En KrAmi-deltagare tillbringar större delen av programtiden i arbetsträning på privata företag med rekryteringsbehov och skrivs inte ut förrän han fått en anställning. För att underlätta rekrytering erbjuds arbetsgivarna lönebidrag[1]. På samma sätt som praktikplatserna är viktiga för integrationen i arbetslivet, är de fritidsaktiviteter som utövas viktiga för integrationen i det sociala livet.

Knuff är ett annat socialt program vars mål är att förbereda arbete eller studier. Till skillnad från KrAmi arbetar de med arbetsträning inom programmens ram, d.v.s. klienterna tränar arbete i programmets egna verkstäder. Både KrAmi och Knuff finns i klienternas närmiljö och bedrivs på dagtid.

Frivårdsgruppen har tagit del av frivårdens brottsförebyggande arbete i form av samtal med frivårdsinspektörer och övervakare och dessutom tagit del av samhällets ordinarie serviceutbud.

De frågor som vi ställer är vad dessa insatser betyder för klienter och samhälle? Förbättrar de klienternas sociala situation? Är de samhällsekonomiskt lönsamma? Leder några av insatserna till bättre resultat än andra? Dessa frågor bör också vara av intresse för praktiskt yrkesverksamma, liksom för beslutsfattare, finansiärer och sist men inte minst för klienterna själva.

1 Lönebidraget är ett av de bidrag som Arbetsmarknadsstyrelsen (AMS) har till sitt förfogande för att hjälpa arbetssökande med arbetshandikapp till ett arbete.

Syfte

Syftet var således att utvärdera KrAmi-programmen, dels vad gäller förbättring för klienterna, dels vad gäller samhällsekonomisk lönsamhet, och använda Knuff och frivården som jämförelsegrupp.

Vi försökte ta reda på

- i vilken grad klienterna och klienternas sociala situation förändrats i fråga om arbete och försörjning, familj och umgänge, kriminalitet samt alkoholbruk och narkotikamissbruk,
- om eventuella förändringar kan relateras till programmen och
- om programmen är samhällsekonomiskt lönsamma.

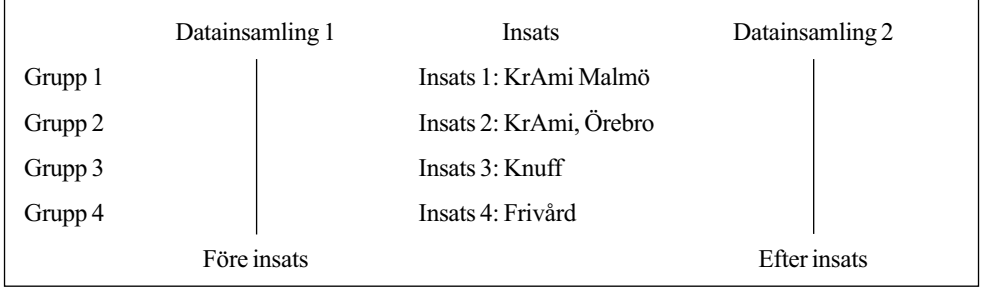
Vi genomförde därför en klienteffektstudie som visar vad insatserna leder till för klienternas del och kompletterade den med en samhällsekonomisk utvärdering. Båda finns redovisade i rapporten Med arbete som insats (Nyström, Jess, Soydan, 2002). En kvalitativ studie som visar hur samtliga klienter själva uppfattar förändring inom ramen för de studerade insatserna används som bakgrundsmaterial (Nyström, 1999).

I den här artikeln kommer vi att redovisa några av resultaten av klienteffektstudien och den samhällsekonomiska utvärderingen och diskutera hur klienteffekter och samhällsekonomisk lönsamhet hänger ihop. Först följer några ord om metoden.

Metod

Det är inte någon lätt uppgift att fånga in

Figur 1.
Forskningens design



resultaten av processer som är så komplexa som sociala insatser och som leder till så varierande resultat. Vi försökte därför ringa in resultatet med olika typer av studier som belyser förändringen från olika perspektiv och med flera olika utfallsvariabler och utfallsmått.

Vi ville jämföra de tre grupperna av klienter som deltagit i de olika sociala insatserna, i deras naturliga miljö. Därför valdes en kvasi-experimentell design som modellen i figur 1 illustrerar. Som framgår ingår två KrAmi-program, Malmö och Örebro i studien, ett annat program Knuff och två grupper av frivårdsklienter dels från Malmö/Helsingborg- och Örebro/ Karlstadsområdet.

Ett av problemen med studier av kvasi-experimentell design är risken för selektionseffekter, d.v.s. skillnader som finns redan före insatsen och som kan förklara de skillnader som visar sig efter. När gruppernas problembild före insats jämfördes fann vi att KrAmi- och frivårdsgrupperna är relativt lika med avseende på brott och kriminalitet men i frivårdsgruppen tenderar de sociala resurserna att vara större. Det visar sig i att färre har föräldrar av annan etniska bakgrund och fler bor på mindre orter. Därutöver är utbildningsnivån i genomsnitt något högre, arbetserfarenheten något längre och fler har körkort. I fråga om ålder skiljer sig KrAmi Malmö-gruppen signifikant från övriga grupper. Gruppens genomsnitt på 22 år skall jämföras med 26 år för övriga grupper. Knuff-gruppen skiljer sig från övriga med en större andel kvinnor och en mer heterogena problembild vilket gör jämförelserna mer osäkra

Klienteffektstudien genomfördes under åren 1996–98. Klienterna intervjuades första gången när programmen och frivårdsinsatserna just skulle starta och andra gången ett år senare. Vi kunde då konstatera i vilka avseenden varje klient förändrats mellan tiden före och efter och beräkna de genomsnittliga förändringen för hela grupperna. KrAmi-gruppens förändring kunde därefter jämföras med Knuff- och frivårdsklienterna.

Bortfallet i andra intervjun var ungefär lika stort i KrAmi- och frivårdsgrupperna, 7 respektive 8 personer. Även klienter som inte fullföljt programmen intervjuades en andra gång. Eftersom bortfallet är litet och avhopparna från programmen finns

Tabell 1.
Antal intervjuade vid de två intervjuerna och bortfall

	Totalt	Intervju 1	Intervju 2	Bortfall
KrAmi Örebro	32	32	27	5
KrAmi Malmö	30	30	28	5
Knuff	28	21	19	2
Frivårdsgruppen	53	53	45	8

med även vid uppföljningsintervjuerna har vi reducerat några av de validitetsproblem som ofta föreligger vid den här typen upplägggning.

Intervjuerna genomfördes i en personlig intervju med ASI, Addiction severity index. ASI är en strukturerad intervju med ungefär 180 frågor. ASI-intervjuns frågor omfattar sju livsområdena arbete och försörjning, kriminalitet och brott, familj- och umgänge, alkohol- och narkotikamissbruk samt fysisk och psykisk hälsa.

Med ASI-intervjun fick vi ett brett spektrum av utfallsvariabler och vi valde några av dem som har särskild relevans för såväl insatsernas inriktning och innehåll som för gruppernas problembild och som dessutom kunde användas för att mäta klienteffekter.

- Arbetade dagar
- Inkomst av anställnin
- Dagar i brott
- Tillfreds med familjeförhållanden
- Dagar med allvarliga konflikte
- Pengar till alkohol respektive narkotika
- Dagar av alkohol- respektive narkotikamissbruk
- Klienternas egna bedömningar av problemsituationen.
- Intervjuarnas bedömning av klienternas problem

Listan innehåller både frågor om faktiska förhållanden och subjektiva upplevelser i form av klientens egen bedömning och intervjuarens bedömningar. De flesta av dessa frågor har ett kort tidsperspektiv på 30 dagar bakåt i tiden vilket minskar risken för överlappningar när man mäter förändring över tid. Men kort tidsperspektiv för också med sig problem. Ett sådant är problemen med inkapacitet vid någon av intervjutillfällena. Om några intervjuade 30 dagar före intervjun befunnit sig t.ex. i behandling eller fängelse har möjligheterna till arbete, brott eller missbruk varit kringskurna. Man bör därför komma ihåg att problembilden för de senaste 30 dagarna i dessa fall inte motsvarar den totala problembilden t.ex. vad gäller kriminalitet eller missbruk.

Som mått på förändring användes dels skillnader i gruppernas medelvärden (faktiska förhållanden), dels skattningarnas procentuella förändringar från tiden före till efter programmet (skattningar) och effektstorlek – ES-värden – för både faktiska förhållanden och skattningar. Effektstorlek är ett statistiskt mått som används för att bestämma insatsers effekt.

Den samhällsekonomiska utvärderingen bygger på faktiska kostnader och skiljer sig i det avseendet från flera liknande studier

som bygger på uppskattningar. För varje enskild klient som ingår i studien har uppgifter samlats in från samtliga berörda myndigheters register för året före, under och efter insats för följande områden:

- Sjukvårdsutnyttjande; öppen och sluten vård, somatisk, psykiatrisk och toxikomani.
- Socialtjänstkostnader; socialbidrag, behandling extern/intern, samtal
- Försäkringskassans kostnader; sjukpenning, föräldrapenning, A-kassa, KAS, utbildningsbidrag, sjukbidrag, bostadsbidrag, bidragsförskott
- Kriminalvårdskostnader; olika typer av anstalt och frivård
- AMS kostnader i form av handläggning och samtal hos arbetsförmedlingen samt ALU, arbetslöshetsersättning, lönebidrag och OSA

Dessutom hämtades uppgifter om pensionspoäng in för att beräkna klienternas anknytning till arbetsmarknaden.

Det finns kostnader som är av intresse men svåra att identifiera och andra vars värde kan diskuteras. Här har myndigheternas kostnader för handläggning och samtal tagits med men däremot inte brotts- och missbruksrelaterade kostnader, kontakter med frivilliga och kostnader för tredje man.

Lönebidraget är ett bidrag som kan användas för att underlätta inträdet på arbetsmarknaden för klienter som ingår i studien. Arbetstagaren får full lön men staten betalar arbetsgivaren ett bidrag som motsvaras av arbetstagarens fastställda arbetshandikapp. Lönebidraget kan betraktas både som ett bidrag för produktionen och som en transferering mellan olika parter i samhället. I det första fallet anses bidraget underlätta inträdet på arbetsmarknaden och med en fullgod arbetsprestation ökar samhällets totala produktion. I det senare fallet betraktas lönebidraget som en ersättning för ett arbetshandikapp som gör att arbetstagaren producerar i motsvarande grad mindre. I det första fallet bör lönebidraget ingå i analysen men inte i det senare.

I en traditionell samhällsekonomisk investeringskalkyl tas bidrag inte med eftersom de betraktas som transfereringar. Det finns emellertid också goda skäl att ta med lönebidraget eftersom det till stor del motsvaras av en reell produktionsökning. Därför redovisas det samhällsekonomiska resultatet både med och utan lönebidrag.

Flera olika typer av analyser har således genomförts:

- Dels jämfördes kostnader för program och kostnadsutvecklingen från ett år före insatsen till ett år efter.
- Dels gjordes en kostnads-/intäktsanalys för att beräkna framtida kostnadseffekter och samhällsekonomiskt resultat.
- Dels rangordnades de olika insatserna med avseende på olika aspekter av samhällsekonomisk lönsamhet.

I en samhällsekonomisk utvärdering utgår man från samhällsekonomisk lönsamhet. Med det menas att de positiva effekterna av en insats är större än de negativa. Även om enskilda förlorar på en samhällsinsats kan nettoeffekterna totalt sett bli positiva om vinsterna överskrider förlusterna. Det är alltså inte bara monetära vinster och förluster som räknas in utan insatsens alla verkningar; positiva likaväl som negativa och kvantifierbara likaväl som andra identifierbara effekter som inte kan kvantifieras.

Samhällsekonomisk lönsamhet bygger precis som företagsekonomisk på principen om »lägsta kostnad per producerad enhet«. En samhällsekonomisk beräkning är litet förenklat en företagsekonomisk beräkning för hela »företaget« Sverige. Det som skiljer är att samhällsekonomiska kalkyler inte begränsas till kostnader och intäkter som direkt berör verksamheten. En annan skillnad är avskrivningstiden för investeringar som i samhällsekonomiska kalkyler oftast beräknas till 15-20 år. Dessa skillnader, de indirekta kostnaderna och den längre avskrivningstiden, kan göra att ett projekt som ur företagsekonomisk synvinkel inte beräknas som lönsamt mycket väl kan vara lönsamt ur samhällsekonomisk synvinkel.

Resultat

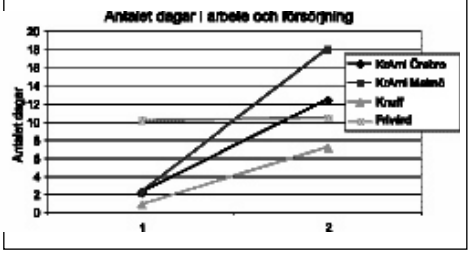
Vad blev då resultatet? Vi börjar med några av klienteffektstudiens viktigaste resultat och presenterar därefter den samhällsekonomiska utvärderingens huvudresultat.

Arbete och försörjning

Med tanke på KrAmi- och Knuff programmens mål bör arbete och försörjning betraktas som det mest relevanta livsområdet. Före insats hade fler arbete i frivårdsgruppen än i programgrupperna. Figur 2 visar förändringen i arbetssituationen.

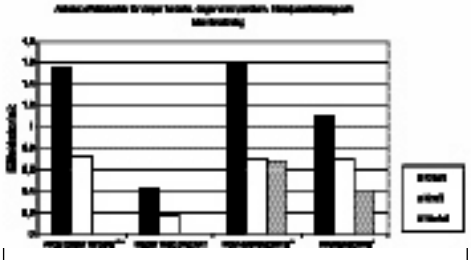
Som framgår har arbetssituationen förbättrats avsevärt för både KrAmi och Knuff, medan frivårdsgruppens situation inte har förändrats. Skillnaden över tid för de två KrAmi-grupperna är signifikant och utvecklingen pekar i samma positiva riktning för

Figur 2.
Arbete och försörjning – dagar i arbete/studier de senaste 30 dagarna



båda KrAmi-grupperna. Efter programmet har det genomsnittliga antalet dagar i arbete/studier för Malmö-gruppens del ökat med nästan 16 dagar och för Örebro-gruppen med 10 dagar. Ökningen för Knuff-programmets del är sex dagar och Frivårdsgruppen har inte förändrats. Andra utfallsvariabler pekar i samma riktning; t.ex. har inkomst av anställning ökat och bidrag från välfärdssystemen minskat i motsvarande grad för samtliga grupper. För frivårdsgruppens del visar dock såväl klienternas egna som intervjuarnas bedömningar förbättring till skillnad från de sakinriktade frågorna. I figur 3 som åskådliggörs effektstorleken för fyra variabler.

Figur 3.
Arbete och försörjning – effektstorlek



Not: Effektstorlek över 0,8 anses som en stor effekt, över 0,5 som en medelstor effekt och värden över 0,2 som en mindre effekt.

Att döma av effektstorleken har KrAmi mycket goda effekter och Knuff medelgoda effekter. För frivårdsgruppens del finner vi mindre effekter för klientens egen och intervjuarens bedömning d.v.s. de subjektiva skattningarna. Den faktiska situation har alltså inte förändrats men däremot har den subjektiva upplevelsen av problemen minskat något.

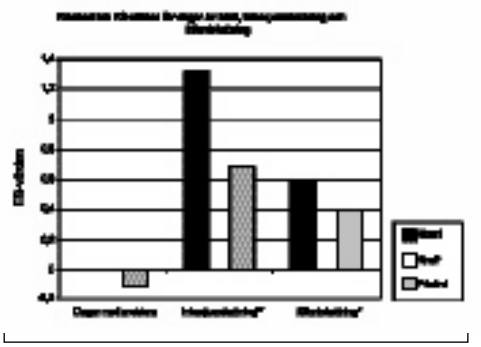
Sammanfattningsvis kan vi konstatera att vad arbete och försörjning beträffar har KrAmi-gruppens situation förbättrats avsevärt. Alla som inte hoppat av programmen har arbete och upplever sin situation som markant bättre. Knuff-gruppen har förändrats tydligt men mer måttligt. Frivårdsgruppens upplever sin situation som bättre men har inte förändrats i sak.

Brott och missbruk

Brottsaktiviteten visar sig vara låg redan vid den första intervjun före insats men något högre för frivårdsgruppens del än för KrAmi-grupperna. Med tanke på att de intervjuade i grupperna tidigare i sitt liv har lång erfarenhet av brott förefaller vår indikator– »brott i syfte att tjäna pengar de senaste 30 dagarna« – inte ge en riktig bild av problemets svårighetsgrad. De låga värdena betingas delvis av inkapacitet som beror på att en relativt stor andel satt i fängelse eller vistades på behandlingshem månaden före intervjuerna. I KrAmi-gruppen var andelen vid första intervjun ungefär en tredjedel och i frivårdsgruppen 15 procent. Vid andra intervjun var frivårdsgruppens andel högst med 17 procent.

Delvis p.g.a. de låga ingångsvärdena finner vi endast en liten minskning i

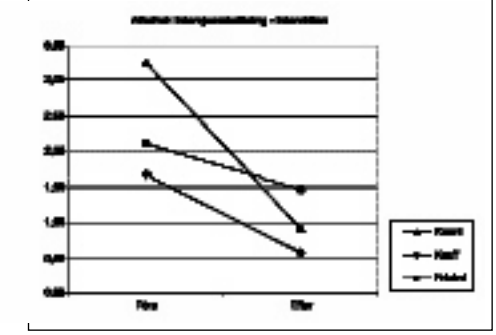
Figur 4.
Kriminalitet; effektstorlek



antalet dagar i brott men ser vi istället till effektstorleken kan vi konstatera en förbättrad situationen för både KrAmi och frivårdsgruppen. Figur 4 visar att effekterna är större för KrAmi än för frivårdsgruppen, för både klienternas egna och intervjuarnas skattningar av problem med kriminalitet och brott. Klientskattningarna förändras signifikant mer för alla programgrupper än för frivårdsgruppen. De subjektiva skattningarna av problemen med kriminalitet visar större effekter för KrAmi än frivårdinsatserna. Brottsnivån är fortsatt låg vid andra intervjun liksom också de subjektiva skattningarna av problemen.

Även för alkohol- och narkotikamissbruk finner vi låg konsumtion före insats. Lägre grad av missbruk än vad som motiveras av tidigare erfarenhet kan även det förklaras av att många har befunnit sig på institution tillsammans med det faktum att många vid programmens/insatsernas start befinner sig i en tidigare påbörjad rehabiliteringsprocess. Även om antal dagar av missbruk inte visar någon förändring tyder dock den beräknade effektstorleken för klienternas

Figur 5.
Alkoholproblem; interaktionseffekter intervjuarnas skattning



egna och intervjuarnas skattningar dock på förbättringar även på detta område.

Därutöver illustrerar figuren en intressant skillnad i utveckling mellan grupperna. Intervjuarna bedömer att KrAmi-klienterna förbättrats i signifikant högre grad än de övriga grupperna. Av intervjuarnas uppfattning att döma minskar alkoholproblem mer för KrAmi än för de övriga två grupperna.

Sammanfattningsvis har KrAmi-gruppens arbetssituation förbättras betydligt. Så gott som samtliga klienter som inte hoppat av programmen visar sig ha arbete ett år efter KrAmi-starten. Det är en klar förändring i riktning mot programmets mål, men hur vet vi att förbättringen förklaras av programmens insatser? Vi hade svårt att hitta andra förklaringar. Skillnaden till jämförelsegrupperna var avsevärd. Dessutom hade gruppen innan de blev straffade dålig prognos. Innan insatsen hade mycket få arbete, utbildningsnivån var låg, behovet av hjälp upplevdes som stort. Det fanns få positiva händelser utan samband med programmet och till detta skall läggas KrAmi-klienternas egen uppfattning som kan tolkas i samma riktning (Nyström, 1999). Knuff-gruppen har också förbättrat sin arbetssituation men i mindre omfattning. En mindre grupp klienter upplever sig hjälpta. För frivårdsgruppens del fann vi ingen förändring.

För områdena kriminalitet och missbruk fann vi före programmet för KrAmi-gruppens del låga värden som åtminstone delvis förklaras av institutionsvistelse dagarna före programmen och av goda incitament för att avstå. Dessa låga värden har förblivit stabila fram till uppföljningen. Även om den uppmätta förändringen är liten bör fortsatt små problem kunna förstås som en positiv effekt av insatserna. Programmet lyckas hjälpa klienterna att inte återgå till sitt tidigare destruktiva liv.

Vår slutsats blev att den förbättring som så tydligt visade sig för arbete och försörjning till stor del förklaras av programmen. Genom KrAmi har klienterna fått ett arbete och lyckas behålla det. Den hjälp de fått att hitta arbete och lösa konflikter, den sociala träningen och den bättre förståelse för arbets- och samhällslivet och stödet efter programmets slut är några av de komponenter som verkar i denna riktning. KrAmi- programmen har sannolikt också bidragit till att problemen med kriminalitet och missbruk fortsatt hålls på en låg nivå efter insatserna. Även Knuff lyckas hjälpa en mindre grupp klienter främst till studier.

Diskussion

Ett av problemen med den kvasi-expe-

rimentella designen är risken att blanda samman effekter av insatsen med andra faktorer. Hur påverkar t.ex. ålderskillnaden mellan KrAmi Malmö och de övriga grupperna resultaten? Yngre klienters förankring bör inte ha utvecklats lika långt och de bör befinna sig i en mer förändringsbar fas av livet. Högre ålder å andra sidan medför bättre motivation och större benägenhet att sluta av egen kraft (Blomqvist, 1999 s. 83-87). Om mognad eller självläkning följer med stigande ålder skulle insatsernas effekter kunna sammanblandas med en naturlig förändring. Sannolikt är debutålder och erfarenheternas längd en lika relevant faktor som den biologiska åldern och KrAmi Malmö- gruppen skiljer sig trots att de är yngre inte från övriga grupper i erfarenheter av kriminalitet, brott och missbruk. Det är alltså svårt att bedöma ålderns påverkan, men eftersom vi har jämfört hela KrAmi-gruppen, d.v.s. både Malmö och Örebro-gruppen, får åldersskillnaden mindre betydelse. I variansanalyserna har ålder korrigerats statistiskt och förklarar därför inte skillnaden mellan program- och frivårdsgrupperna.

I fråga om sociala resurser fann vi en annan skillnad mellan frivårdsgruppen och KrAmi-programmen. Skillnaden gjorde sig synlig i färre invandrare, högre utbildning, längre erfarenhet av arbete samt bostadsorter med mindre kriminell belastning i frivårdsgruppen. Möjligheterna att klara sig själv med stöd i sin närmaste omgivning bör vara större för personer eller grupper som har större tillgång till sociala resurser och färre möjligheter till brott och missbruk. Det finns risk att dessa sociala resurser som fanns före insats kan blandas samman med de uppmätta effekter. När vi jämför KrAmi-programmen med frivårdsgruppen är det alltså möjligt att vi underskattar KrAmi-programmens effekter och överskattar frivårdens.

Trots olikheter i sociala resurser är KrAmi- och frivårdsgrupperna relativt lika när vi ser till erfarenheten av kriminalitet och missbruk. Däremot skiljer sig Knuff med sin mer heterogena målgrupp. Gruppen består av en blandning av klienter med stora problem med kriminalitet och brott, personer vars problem kännetecknas av social oro och personer som inte har några problem med missbruk eller asocialitet. Psykologiska problem är också vanliga. I gruppen finns också fler kvinnor än vad som är fallet i de två andra grupperna. På grund av olikheterna blir jämförelser svåra och slutsatserna därför också osäkra.

Både klienter och insatser påverkas av externa faktorer som har med omgivningen och samhällsutvecklingen att göra. Några omgivningsfaktorer som kan ha påverkat resultaten är arbetsmarknadens utveckling, skillnader i tillgången på arbete och omfattningen av problem som har att göra med brott och missbruk.

Både skillnader i arbetsmarknad och problem med missbruk och kriminalitet har vi försökt eliminera genom att slå ihop de två KrAmi-programmen och frivårdsgrupperna i analysen.

Hur skall vi förstå det för KrAmi-programmens del mycket fördelaktiga resultaten? Vi jämförde resultaten med vad internationell forskning säger om vilken typ av program som ger positiva resultat. Flera forskningsöversikter visar att arbetsrelaterade program, inriktade på att kor-

rigera beteende och färdigheter tenderar att minska återfall i brott och att behandlingens längd och intensitet och forskarmedverkan påverkar utfallet. (Lipsey ,1992, 1995) Bäst lyckas program som inriktar sig på personer med hög risk för återfall och som baseras på en bedömning av klienternas brottsrelaterade behov. (Palmer, 1995 och Gendreau ,1996)

När vi jämför programmens innehåll med den internationella forskningens framgångskomponenter kan vi se att främst KrAmi, men även Knuff innehåller många av dessa. KrAmi är:

- ett heltidsprogram
- inriktat på arbete och social träning
- innehåll och arbetsformer stämmer för en väl definierad målgrupp
- förmågan att stödja klienterna praktiskt och hjälpa dem att förstå hur ett accepterat socialt liv fungerar
- stödet till klienterna efter programmets slut
- den speciella formen av samverkan mellan myndigheter.

Bristen på jämförbara svenska studier gör det dock svårt att dra slutsatser om hur generella våra resultat är men den internationella forskningen tycks peka på resultat i samma riktning.

Kostnadsjämförelse

Vi låter redovisningen av den samhällsekonomiska utvärderingens resultat börja med en jämförelse av programkostnader. När både programkostnader och andra samhällskostnader för tiden i behandling beräknats visar det sig att frivården i Örebro/Karlstad är den dyraste insatsen och Knuff den billigaste. Som framgår av tabellen är insatserna olika långa; KrAmi Malmö har den kortaste genomsnittliga behandlingstiden och frivård Örebro/ Karlstad den längsta.

När det gäller kostnader före respektive efter visar det sig att dessa kostnadsförändringar endast är signifikanta för KrAmi Örebro-gruppen.

KrAmi Örebro har till skillnad från KrAmi Malmö höga kostnader för kriminalvård året före programmet. De som hoppar av KrAmi Malmö begår brott och hamnar redan året efter programmet i fängelse.

Jämförelse av pensionspoäng

Gruppernas anknytning till arbetsmarknaden är en viktig grund för de samhällsekonomiska beräkningarna och de bedömningar som kommer att göras. Pensionspoängen som visar klienternas anknytning till arbetsmarknaden har använts vid beräkningen av värdet av klienternas rehabilitering till arbetslivet.[2] Faktiskt utbetald lön och andra skattepliktiga förmåner som skattepliktiga ersättningar från socialförsäkringssystemet som t.ex. sjukpenning, arbetslöshetskassa och utbildningsbidrag är pensionsgrundande men däremot inte socialbidrag. Högsta möjliga poäng, dvs 650 pensionspoäng uppnår en person som under ett år har en inkomst på 25 000 kr per månad. En genomsnittlig industriarbetarlön av 15 000 kr per månad ger då 385 pensionspoäng.

2 Pensionspoängen räknas ut enligt följande formel: (Den pensionsgrundande inkomsten – 37 200)/ 37 200 = pensionspoängen. Inkomst av tjänst och sjukbidrag är andra indikatorer på anknytning till arbetsmarknaden.

Tabell 2.

Kostnader för hela insatsperioden

Behandling	Medelkostnad/ dag	Insatsperiod	Kostnad för hela insats-perioden
KrAmi Malmö	537 kr	169 dagar	90 753 kr
FrivårdMalmö/Helsingborg	442 kr	425 dagar	187 850 kr
KrAmi Örebro	664 kr	299 dagar	198 536 kr
Frivård Örebro/Karlstad	909 kr	444 dagar	403 596 kr
Knuff	170 kr	309 dagar	52 530 kr

Tabell 3.

Kostnader per dygn före/efter

Behandling	Före	Efter
KrAmi Malmö (30) Fullföljda (21) Ej fullföljda (9)	281 kr 300 kr 234 kr	295 kr 129 kr 732 kr
KrAmi Örebro Fullföljda (17) Ej fullföljda (15)	610 kr 466 kr 775 kr	294 kr* 282 kr 307 kr*
Knuff Fullföljda (9) Ej fullföljda (17)	180 143 kr 200 kr	126 148 kr 116 kr
Frivård Malmö/Helsingborg (24)	322 kr	296 kr
Frivård Örebro/Karlstad (28)	732 kr	544 kr

Tabell 4.

Pensionspoäng före, under och efter insats

Pensionspoäng	Ett år före insats			Ett år efter insats			År 1999		
Behandling	Samt-liga	Fullf	Ej fullf	Samtliga	Fullf	Ej fullf	Samt-liga	Fullf	Ej fullf
KrAmi Malmö (30)	29	41	0	101	145	0	239*	342*	0
KrAmi Örebro (32)	36	45	25	38	71	0	122*	189*	46
Knuff (26)	71	55	80	84	105*	74	50	79	35
FrivårdMalmö/H(24)	29			7*			83		
Frivård Örebro/K (28)	41			53			182*		

Förklaring:*–signifikantskillnad.IkategorinFullföljtfinnsdeklientersomfullföljtprogrammetifråga,IEjfullföljtklienter klienter som hoppat av innan programmet slutförts.

Båda KrAmi-gruppernas anknytning till arbetsmarknaden har förbättrats men för KrAmi i Malmö går utvecklingen i en snabbare takt. Denna skillnad kan förklaras av skillnader i programmen, skillnader i klientgrupperna eller skillnader i arbetsmarknadssituation mellan Örebroregionen och Malmö-regionen. Knuffklienternas anknytning till arbetsmarknaden har försämrats något över tid, en utveckling som emellertid inte är statistiskt signifikant. För frivårdsgruppernas del ser vi en gynnsam ej signifikant tendens i Örebro/ Karlstadsgruppen och en signifikant försämring för Malmö/ Helsingborggruppen det första året efter insatsen.

Samhällsekonomisk lönsamhet

En KrAmi Malmöklient ger en samhällsekonomisk vinst på 1,6 mkr exkl lönebidrag. Räknas lönebidraget in blir det samhällsekonomiska resultatet totalt 1,5 mkr. Kostnaderna för programmet skrivs av efter 1,5 år. För en klient som fullföljt programmet blir kostnadsförändringarna ca tio gånger högre samtidigt som också produktionsintäkterna ökar vilket ger en samhällsekonomisk vinst på ca 2,5 mkr. Behandlingen har lönat sig ett halvår efter avslutat program. En avhoppare ger däremot en samhällsekonomisk förlust på ca 4,0 mkr vilket beror på både att kostnaderna förändras i negativ riktning och att pensionspoäng blir kvar på noll d.v.s. ökningen i pensionspoäng uteblir. Programkostnaden betalas inte tillbaka under perioden.

En klient i frivården Malmö/Helsingborg ger en samhällsekonomisk vinst på knappt 1,0 mkr. Kostnadsförändringarna är lika stora som för en deltagare i KrAmi Malmö, men produktionsintäkterna är bara hälften så stora. Programkostnaden betalas tillbaka på drygt fyra år.

En KrAmi Örebroklient ger en total samhällsekonomisk vinst på 2,5 mkr varav kostnadsbesparingarna utgör ca 1,4 mkr. Behandlingskostnaden är avskriven efter drygt ett år. Om lönebidraget inkluderas i beräkningen försämras det samhällsekonomiska resultatet med ca 70 000 kronor. De som fullföljer programmet ger en lika stor samhällsekonomisk vinst som gruppen totalt men ger mindre kostnadsbesparing och högre produktionsintäkter. Avhopparnas kostnadsbild blir tvärtom högre kostnadsbesparingar och mindre ökning i pensionspoäng.

En klient i frivården i Örebro/Karlstad ger en samhällsekonomisk vinst på totalt 2,4 mkr. Programkostnaden betalas tillbaka på två och ett halvt år. Om lönebidraget inkluderas i kalkylen försämras resultatet med cirka 16 000 kr.

En Knuff-klient ger en samhällsekonomisk vinst på knappt 700 000 kr, varav kostnadsbesparingarna från perioden före till perioden efter är ca 250 000 kr. Behandlingskostnaden är avskriven efter knappt 3 år. Resultatet blir ca 13 000 kr sämre om lönebidraget räknas in i kalkylen. För de som fullföljer programmet minskar kostnaderna i mycket mindre utsträckning och produktionsintäkterna ökar mer än för gruppen totalt. För avhopparna minskar kostnaderna mer samtidigt som produktionsintäkterna blir lägre.

Tabell 5.

Jämförelse av kostnader och resultat

	KrAmi Malmö	KrAmi Örebro	Knuff	Frivård Örebro/ Karlstad	FrivårdMalmö/ Helsingborg
Pensionspoäng	239	122	50	182	83
Pensionspoäng/ programkostnad	45	18	29	20	19
Tillbakabetalt kr per investerad kr	17,8	12,7	13,2	5,8	5,1
Tillbakabetalt* Programkostnad	9 559 kr	8 433 kr	2 244 kr	5 272 kr	2 254 kr
Avskrivningstid	1,5 år	1 år	3 år	2,5 år	4 år
Programkostnad/dag	537 kr	664 kr	170 kr	909 kr	442 kr
Besparing före minus efter	-14 kr	316 kr	54 kr	188 kr	26 kr
Besparing/ programkostnad	-2,6 %	47,6 %	31,8 %	20,7 %	5,9 %

Tabellförklaring till tabell 5:
Programkostnad = Programkostnad per dag per deltagare
Besparing = Besparing efter minus före per dag per deltagare
Pensionspoäng = Pensionspoäng per deltagare år 1999
Avskrivning = Hur lång tid det tar att skriva av programkostnaden
Tillbakabetalning = Den samhällsekonomiska besparingen per deltagare/programkostnaden per deltagare

Rangordning

När programkostnader, pensionspoäng och samhällsekonomiskt resultat sammanställs kan en rangordning göras mellan de olika insatserna. Tabellerna nedan visar resultatet i sammanfattning vilket ger underlag för en rangordning av de olika behandlingsalternativens resultat med avseende på det eller de mål som man prioriterar, dvs. om målet är:

a) att göra den största besparingen,
b) att öka sysselsättningsgraden,
c) att pengarna avskrivs snarast möjligt eller
d) att pengarna ger mesta möjliga avkastning på insatt kapital

Tabellerna ovan visar att programgrupperna – KrAmi och Knuff - är mer kostnadseffektiva än frivårdsgrupperna men de visar också att frivården Örebro/ Karlstad är lika kostnadseffektiv som Knuff. Anmärkningsvärt är att kostnaderna för de tre programmen betalas tillbaka på så kort tid som 1–3 år. Kostnaden för programmen har i stort sett betalats tillbaka under uppföljningstiden. All produktionsökning och alla kostnadsminskningar efter att programkostnaden skrivits av ger samhället vinst.

Med hjälp av kostnads-/intäktsanalysen kan de olika programmen rangordnas med avseende på målen. KrAmi Örebro ger den

största samhällsekonomiska besparingen per individ och den kortaste avskrivningstiden. KrAmi Malmö ger den största ökningen av sysselsättningsgraden och den högsta avkastningen per investerad krona. Man skulle också kunna rangordna programmen efter deras genomsnittliga resultat utan hänsyn till resultatens inbördes betydelse. Då ser vi att de två KrAmi-programmen har den högsta rangordningen i de flesta avseenden.

Diskussion

Hur skall vi nu bedöma dessa resultat? Vi skall kort diskutera hur lönebidraget, den beräknade rehabiliteringsgraden, den svarta ekonomin, arbetsmarkandsutvecklingen, ränteläget och vissa olikheter i uppföljningsperioden kan påverka resultaten.

Socialbidrag, utbildningsbidrag och andra liknande bidrag utgör ibland ersättning för någon typ av produktion men i enlighet med samhällsekonomisk teori inkluderas de inte och deras påverkan på det samhällsekonomiska resultatet är marginell. För lönebidraget som i viss grad är en ersättning för produktion har däremot beräkningar gjorts men lönebidraget påverkar inte det samhällsekonomiska resultatet nämnvärt eftersom bidragen betalas ut under en begränsad period.

När rehabiliteringsgraden har beräknats med hjälp av pensionspoängen har hänsyn tagits till att några i grupperna sannolikt kommer att återfalla i missbruk och/eller brott under den 15-åriga tidsperioden. De som fullföljt KrAmi Malmö har emellertid redan under uppföljningsperioden uppnått 90 procent av full pensionspoäng och rehabiliteringsgraden beräknas därför till 100-procentig pensionspoäng. Om gruppens poäng kvarstår på 90 procent hela 15-årsperioden skall det redovisade resultatet minskas med ca 150 000 kr per klient. Skillnaden i antagen rehabiliteringsgrad mellan undersökningsgrupper och jämförelsegrupper är 10 procent vilket är realistiskt med tanke på den 10 procentiga skillnad mellan undersökning- och jämförelsegrupper som t.ex. Lipsey´s metaanalys visar (Lipsey 1995). För alla grupper innehåller den beräknade kostnadsbesparingen mellan perioden före och perioden efter en möjlighet till återfall. Kostnaderna för perioden efter är nämligen betydligt högre än för normalbefolkningen och ligger i storleksordningen 100 000 kr –200 000 kr per år per individ. Detta kan jämföras med att en individ som mottar socialbidrag »kostar« ca 60 000 kr per år.

Ingen uppskattning har heller gjort av värdet av stulet gods eller värdet av försäljning och konsumtion av droger eftersom dessa är en del av den svarta marknaden som inte ingår i bruttonationalprodukten (BNP). Värdet av stulet gods för att finansiera narkotikahandeln kan emellertid anses vara betydande och kommer på sikt in i samhällets ekonomi via försäkringsbolagen. Den narkotikarelaterade brottsligheten beräknas dock till 500 miljoner kr per år, vilket inkluderar kostnaderna för polis, tull, åklagare, domstol, försäkringsbolag, detaljhandel och privatpersoner. Dessa icke samhällsekonomiska kostnader såsom värdet av stöldgods och självrisker uppskattas till lika mycket. (ESO-rapport Ds 1999: 46) Kostnaderna för narkotikainköp för 8 000 aktiva missbrukare uppgår till 5 mil-

joner kr per dag, ca 600 kr per dag. (SOU 2000:126)

Förändringar i ränteläge och inflation påverkar det förväntade framtida resultatet. En 8-procentig diskonteringsränta användes för att försöka fånga in eventuella framtida ränteökningar och minska effekterna av för höga förväntningar på ett positivt framtida resultatet av rehabiliteringen. Om istället en 5 procentig diskonteringsränta använts hade det redovisade resultatet blivit ca 20 procent bättre.

Arbete på öppna arbetsmarknaden är beroende av arbetsmarknadsutvecklingen och den allmänna konjunkturen. Arbetsmarknaden har under hela den studerade perioden utvecklats positivt och fler människor är i arbete. Om konjunkturen försämras skulle detta kunna påverka de som har fått arbete på ett negativt sätt; de kan förväntas få en sämre anknytning till arbetsmarknaden än sina jämnåriga.

I vår undersökning är resultatet bäst för dem som fullföljt KrAmi Malmö, de har 1,5 år efter avslutat program nått 90 procent av full rehabilitering. Under undersökningsperioden hade Malmö-regionen en besvärlig arbetsmarknad med högre arbetslöshet än riksgenomsnittet och som sannolikt medförde en ännu besvärligare arbetsmarknad för den aktuella gruppen. Det mycket goda resultatet för KrAmi Malmö blir då ännu mer anmärkningsvärt. Även resultatet för frivårdsgrupp Örebro/Karlstad är anmärkningsvärt positivt när vi sätter det i relation till KrAmi Örebro. Klienterna i frivården Örebro/Karlstad kommer dock i något större utsträckning från mindre orter och gruppen har större sociala resurser vilket kan ha påverkats resultatet i positiv riktning.

Resultaten av den samhällsekonomiska studien visar att varje investerad krona i KrAmi Malmö ger 12,9, i KrAmi Örebro 17,9, i Knuff 13,2 i minskade kostnader och/eller ökade intäkter. För frivården Malmö/Helsingborg är siffrorna 5,2 och frivård Örebro/Karlstad 6,0. Kostnaderna mellan perioden före och efter har som vi tidigare visat minskat med 25 procent för jämförelsegrupperna och 50 för programgrupperna. Resultaten pekar i samma riktning som resultaten av andra liknande studier men är större och mer positiva än vissa studier. (jmf Jones/Vischi 1979, Deschenes et al 1991, Holder et al 1997, Rundell et al 1981)

Troligen förklaras skillnaderna av att dessa studier till skillnad från vår omfattar färre kostnader och baseras på självrapporterade uppgifter. Det kan också förklaras av att dessa studier riktats mot klientgrupper som tar mindre kostnader i anspråk.

Det finns dock studier med liknande resultat. Vi kan t.ex. jämföra med återbetalning på investering motsvarande 23,5 respektive 17,8. (Jess 1998, 1999), en samhällsekonomisk analys av narkomaner vid ett rehabiliteringscenter i Oslo visar att varje investerad krona ger 25,0 tillbaka (Berg, Andersen 1992) och Riksrevisionsverket (F 1993:2) som genom kostnadsberäkningar av fem fallstudier kommit fram till att samhället kunnat spara mellan 1,8 – 3,7 mkr per missbrukare om man kunnat avstyra missbruket. Den samhällsekonomiska utvärderingens resultat, som bygger på myndigheternas egna uppgifter om varje klient verkar inte orimliga i ljuset av dessa studier.

De jämförelsevis goda resultaten kan

vara en effekt av att KrAmi-programmen medvetet satsar på utbildningsbidrag och skuldsanering under programperioden kombinerad med utskrivning till lönebidrag efter programmets slut. Klienterna försäkras på så sätt både skäliga inkomster och att kraven om återbetalning för skulder hålls nere genom skuldsanering. Just stora skulder, »vita« likaväl som »svarta«, kan ses som ett hinder för anpassning till ett »normalt« liv och ett skäl till att många väljer att fortsätta sin kriminella bana eftersom inkomster som drivs in blir meningslösa när man tvingas att leva på existensminimum för lång tid framåt.

Värdet för klienter och samhälle

Vi skall avsluta med att diskutera vad en jämförelse mellan de två studierna kan ge. Visar de två studierna på samma tendenserna? Bekräftar de två studiernas resultat varandras i några avseende? Vad säger de två studiernas sammanvägda resultat om insatsernas värde för klienter och samhälle?

Som nämndes inledningsvis har de två utvärderingarna såväl de studerade samhällsinsatserna som klienterna gemensamt: klienter som tagit del av KrAmi-programmen, i Knuff-programmet och frivårdens insatser. Gemensamt är också att de båda följer upp klienterna från en tid före insatsen till en tid efter. Samma klienter, samma insatser och samma typ av före/efter studie ger en gemensam grund som gör jämförelser möjliga.

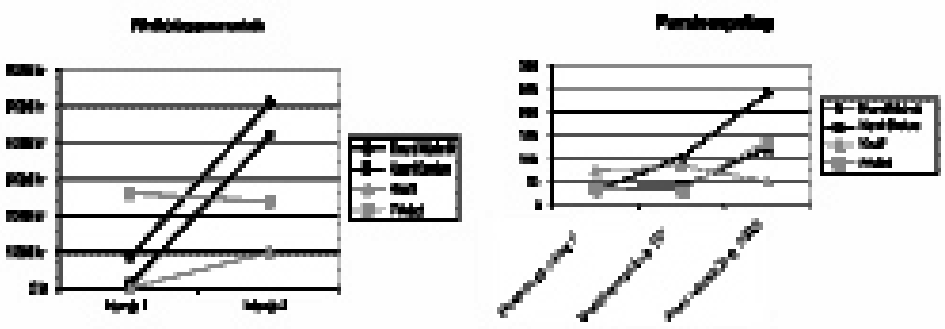
Men även om klienteffekter och kostnadseffekter kastar ljus över samma problem belyser de olika aspekter. Den mest kostnadseffektiva insatsen är inte nödvändigtvis den bästa ur klientens synvinkel. En insats som inte är samhällsekonomiskt lönsam kan ju leda till förbättring för klienterna och en samhällsekonomiska lönsam insats medför inte automatiskt goda klienteffekter. Ett bättre utfall för en klient kan ha sin orsak i att mer resurser satsas men kan också leda till att samhällets kostnader minskar. Därtill kommer att klienteffektstudien baseras på klientuppgifter och den samhällsekonomiska utvärderingen på uppgifter hämtade från myndighetsregister och att det finns skillnader i uppföljningstid – klienteffektstudien, 30 dagar och den samhällsekonomiska studien 1 år. Det här betyder att de utfallsindikatorer som används i de två utvärderingarna väl kompletterar varandra men de bör jämföras med dessa skillnader i åtanke.

Liknande tendenser

Visar de två studierna på samma tendenser? I figur 6 åskådliggörs förändringen med hjälp av effektstudiens klientuppgifter om »inkomst av anställning« och pensionspoängens utveckling. Förutom möjligheten att jämföra får vi här en tillförlitlig prognos över hur grupperna utvecklas efter uppföljningstidens slut. Jämförelsen visar på liknande tendenser i båda studierna.

Pensionspoängens utveckling visar att förändringen för KrAmi-programmens del inte är tillfällig utan den fortsätter lång tid efter att programmen är slut. För KrAmi Örebro tar det längre tid att förbättra pensionspoängen och klienterna har ett år senare inte nått samma grad av integrering på arbetsmarknaden som KrAmi Malmö.

Figur 6.
Jämförelse försörjning av anställning och pensionspoäng



Not. Tabellen till vänster visar inkomst av anställning 30 dagar före respektive intervju. Effektstudiernas första intervju gjordes mellan 96-97, den andra i slutet av 97-98. Diagrammet till höger visar pensionspoängens utveckling 1995-1999. Observera olikheterna i tid.

Pensionspoängen förbättras däremot avsevärt mellan 1998 och 1999. Förändringen för frivårdsgruppen i Örebro/Karlstad är parallell med den för KrAmi Örebro. För Knuff-klienternas del avtar förbättringen en tid efter programmet vilket kan beror på att flera Knuff-klienter söker sig till utbildning.

Pensionspoängens utveckling visar också att integrationen på arbetsmarknaden för frivårdsgruppens del förbättras efter en tid, men dock endast för frivårdsgruppen Örebro/Karlstad, vars anknytning till arbetsmarknaden förbättras. En av förklaringarna kan vara att socialtjänsten som framgick av den samhällsekonomiska utvärderingen satsar betydande resurser. Frivårdsklienterna i Örebro/Karlstad kostar samhället mer än de övriga grupperna. En annan möjlig förklaring är att gruppen kommer från mindre orter och har större socialt nätverk eller mer av social kontroll än t.ex. Malmö/Helsingborg-gruppen. Jämförelsen visar också att frivårdsgruppens pensionspoäng före insats inte motsvarar klienternas egna uppgifter om inkomst av anställning. De inkomster som klienterna uppger sig ha före insats är med andra ord inte pensionsgrundande och sannolikt till stor del »svarta« inkomster.

Klienteffektstudiens uppgifter om försörjning av trygghetssystem har jämförts med myndigheternas uppgifter om utbetalning av bidrag. Vi fann störst samstämmighet för KrAmi Malmös del. Jämförelsen visar också att klienteffektstudiens uppgifter om inkapacitet, d.v.s. tid i fängelse respektive behandling före insatserna, stämmer väl med registeruppgifterna. KrAmi Örebro har haft betydligt större andel klienter i fängelse före programmet och KrAmi Malmö ett betydligt större andel i behandling.

Avhopparna från KrAmi Malmös kommer mycket snart efter avhoppet i fängelse och kostnaderna för krimi-

nalvård ökar då dramatiskt. Även för frivårdsgrupperna skiljer sig kriminalvårdskostnader. Frivårdsgruppen i Örebro/Karlstad har höga kostnader före medan frivårdsgruppen Malmö/ Helsingborg har lägre kostnader före och kostnaderna minskar mindre.

Jämförelsen av de två utvärderingarnas har visat att klienternas egna uppgifter inte motsägs utan de facto stöds av den samhällsekonomiska utvärderingens registeruppgifter. Eftersom tendenserna pekar i samma riktning stärks våra slutsatser.

Jämförelsen synliggör dessutom några utvärderingstekniska problem. Om flertalet av klienterna före insats avtjänar fängelsestraff blir den aktuella problemsituationen lågt värderad på grund av inkapacitet medan kostnaderna av samma anledning blir höga. Om fängelse förekommer i mindre omfattning efter en insats kan klienternas problem öka – såväl om de mäts med faktiska antal dagar eller skattningar - medan kostnaderna minskar. Om man inte tar hänsyn till sådana förhållanden kan slutsatserna bli missvisande

Bäst för klienter och samhälle

Klienteffektstudiens resultat visar att KrAmi-grupperna förändrats mest och att KrAmi-programmets har mycket stora effekter. Knuff-programmet visar måttliga effekter och frivårdens insatser har måttliga effekter i vissa avseenden.

Jämför vi med den rangordning som blev resultatet av den samhällsekonomiska utvärderingen stämmer resultaten väl överens. KrAmi Örebro gav den största samhällsekonomiska besparingen per individ och den kortaste avskrivningstiden. KrAmi Malmö ger den största ökningen av sysselsättningsgraden och den högsta avkastningen per investerad krona. När programmen rangordnas efter deras genomsnittliga resultat har de två KrAmiprogrammen den högsta rangordningen i de flesta avseenden. Vi får en relativt entydig bild av programmens värde för klienter och samhälle. Det som är bäst för klienterna är också bäst för samhället.

Genom att studera samma insatser och samma klientgrupper har vi försökt belysa program och frivårdsinsatser utifrån både ett klient- och ett samhällsperspektiv. Tillsammans visar de två utvärderingarna att de insatser som är bäst för klienterna också är de mest lönsamma för samhället. Med insatser som liksom KrAmi är väl avpassade för sin målgrupp och genomförs på ett konsekvent och väl integrerat sätt vinner både klienter och samhälle.

Med flera perspektiv – klienteffektstudiens klientperspektiv, den samhällsekonomiska utvärderingens samhällsperspektiv och den kompletterande kvalitativa studie av klienternas förändring inom programmets ram (Nyström 1999) har slutsatser och tendenser styrkts och kompletterande analyser gjorts. Med flera olika perspektiv har vi tagit oss över några hinder som ligger i utvärderingstekniska problem och den kunskap som vi har nått blivit säkrare.

Denna artikeln bygger på CUS-rapporten: Nyström, Jess, Soydan (2002) »Med arbete som insats- klienteffekter och samhällsekonomisk lönsamhet i socialt arbete«, ett arbete som många lämnat ovärderliga bidrag till. Särskilt vill vi nämna professor Bengt-Åke Armelius och docent Kerstin Armelius, vid Psykologiska institutionen Umeå

universitet, docent Åke Bergmark, Stockholms universitet, fil och med dr John Berg vid Ullevål sykehus i Oslo, ekonomen Tom Nilsson vid Socialstyrelsen, medförfattaren till rapporten forskningsledaren Haluk Soydan, som tagit initiativet och var projektledare för båda utvärderingarna under projektens första år samt professor Karin Tengvald, direktör vid CUS.

Litteratur

Andréasson, S., Lindström, U., Armelius, B-Å., Larsson, H., Berglund, M., Frank, A., Bergman, H., Rydberg, U., Zingmark, D. (1996) ASI - ett sätt att intervjua klienter i missbrukarvården. Stockholm: CUS-skrift 1996:1.

Berg J. E. and Andersen S. (1992) Drug addict rehabilitation, a burden on society? International Journal of Rehabilitation Research.

Berglund, M., Andréasson, S., Bergmark, A., Oscarsson, L., Tengvald, K., Öjehagen, A. (1996) Dokumentation inom missbrukarvården. Behandling, metodutveckling, utvärdering, Stockholm: Centrum för utvärdering av socialt arbete / Liber förlag.

Blomqvist, J. (1999) Inte bara behandling – vägar ur alkoholmissbruket. Stockholm. Burner och Bruno. FoU-rapport 1999:16

Bohm, P. (1988) Samhällsekonomisk effektivitet. Stockholm: SNS förlag 1988 (4:e reviderad upplaga).

Deschenes, E. P., Anglin, M. D., Speckart, G. (1991) Narcotics addiction, Related criminal careers, social and economic costs. Journal of drug issues. 21:2.

Ds 1992:46 (1992) Statliga bidrag – motiv, kostnader, effekter? Finansdepartementet, ESO-rapport.

Ds 1999:46 (1999) Bostad sökes. Finansdepartementet, ESO-rapport.

Gendreau, P. (1996) The principles of effective intervention with offenders, in A. Hartland (ed.) Choosing Correctional Options That Work. Thousand Oaks, SAGE.

Gendreau, P. (1996) The principles of effective intervention with offenders. I A. Hartland (ed.) Choosing Correctional Options That Work. Thousand Oaks, SAGE,.

Gramlich, E. M., (1992) A Guide to Benefit-Cost Analysis. New Jersey, USA, Prentice Hall.

Jess, K., Scheffel Birath, C., Jertfelt-Gustafsson T (1999) Samordnare och samverkan. Rehabilitering för psykiskt handikappade missbrukare. Beroendecentrum Nord. Stockholm.

Jess, K., Westerlund, S. (1998) Kostnads-/effektanalys som metod för utvärdering av behandlingsresultat. Beroendecentrum Nord. Stockholm.

Jones, K., Vischi, T. (1979) Impact of Alcohol, Drug Abuse, and Mental Health Treatment on Medical Care Utilization. Medical Care 17(12): Supplement.

Levin, H. M. (1983) Cost-Effectiveness analysis: A Primer. Beverly Hills, California, Sage Publications.

Lipsey, M. W. (1992). Juvenile delinquency treatment: A meta-analytic inquiry into the variability of effects. In Cook T. D., Cooper, H., Cordray, D. S., Hartmann, H., Hedges, L. V., Light, R. J., Louis, T. A. & Mosteller, F. (eds.) Meta-analysis for explanation. A casebook, New York: Russell Sage Foundation.

Lipsey, M. W. (1995) What do we learn from 400 research studies on the effectiveness of treatment with juvenile delinquents? In McGuire J. ed. (1995), What works: Reducing reoffending. Guidelines from research and practice. Chichester: John Wiley & Sons.

Luckey, J. W. (1987) Justifying alcohol treatment on the basis of cost savings. Alcohol Health and Research World 12:1.

Nyström, S. (1999) Socialt förändringsarbete- en fråga om att omförhandla mening, Stockholm: HLS- förlag.

Nyström, S., Soydan, H., Jess, K. (2002) Med

arbetet som insats – klienteffekter och samhällsekonomisk lönsamhet i socialt arbete. CUS-rapport 2001

Palmer, T. (1996) Programmatic and Nonprogrammatic Aspects of Successful Intervention, in Hartland, A. (ed.) Choosing Correctional Options That Work, Thousand Oaks: SAGE.

Plotnick R D (1994) Applying Benefit-Cost Analysis to Substance Use Prevention Programs. The International Journal of Addiction 29/1994.

Riksrevisionsverket (1993) Narkomanvården – om kostnader, resursutnyttjande, samordning och statlig styrning. Förvaltningsrevisionen utreder F 1993:2.

Rundell, O. H., Jones, R. K., Gregory, D. (1981): Practical benefit-cost analysis for alcoholism programs. Alcoholism. 5:4.

SOU 2000:126 (2000) Vägvalet. Den narkotikapolitiska utmaningen. Slutbetänkande av Narkotikakommissionen.

Tengvald, K. & Andréasson, S. (1996) Perspektiv på uppföljning, utvärdering och kvalitetssäkring inom missbruksvården i Berglund, M., Andréasson, S., Bergmark, A., Oscarsson, L., Tengvald, K., Öjehagen, A. (1996) Dokumentation inom missbrukarvården. Behandling, metodutveckling, utvärdering, Stockholm: Centrum för utvärdering av socialt arbete / Liber förlag.

Summary
An investment in social work

The overall aim was to evaluate the KrAmi programme with regard to its effects on the client and on the cost of public services used . The main questions were whether, to what degree and in what respects the interventions improved life for clients and how they affected public expenditure.

Methods

A quasi-experimental study with before-and-after measurements based on 136 participants in the KrAmi (55) or the Knuff programme (19) or on traditional probation (45) was undertaken. The clients in the comparison samples of probation-service clients were exposed to »ordinary« probation circumstances. The dropouts were included in the follow-up. The follow-up time was also roughly six months. The Addiction Severity Index (ASI) was used for interviews and outcome variables.

Results – the study of clients´ effects

The employment situation of the KrAmi group changed dramatically during the follow-up period in terms of employment and income. The Knuff group was also better-off at follow-up We found no change in the probation sample. »Income from employment«and»socialallowances«changed accordingly. The KrAmi programme showed extraordinary effect size values for all outcomes included. (1.6 days of work 1.6 for interviewers assessment and 1.1 for clients' assessment)

Crime problems decreased more in the KrAmi group than in the probation group and use of narcotics decreased. For interviewers' assessment of alcohol problems we found high effect size for the KrAmi group, medium-high for the Knuff group and low for the probation group. The ten-

dencies were the same for drug abuse, but with some higher effects for the probation group.

Efforts to find other explanations than programme factors failed, and qualitative interviews with clients confirm the conclusion that the KrAmi programme contributed strongly to these results. The probation group did not change as far as work was concerned. However, crime prevention – which is the main task of the probation services – was enhanced.

Study of public costs

The evaluation of public costs is based on the same individuals, for one year before intervention, during intervention, and one year after.

Method

We analysed cost cuts and benefit revenues in the rehabilitation of the 140 clients of the KrAmi and Knuff programmes and in the ordinary probation service in a longitudinal perspective between 1995 and 1999. The results were used both in a 15-year investment analysis (CBA) and in a cost-effectiveness analysis (CEA).

Data for the CBA and CEA collected from official records for each client included social service costs, criminal justice system costs, health care costs and social security system costs. The follow-up rate was 100% thanks to the use of official data. Pension points were used in the human investment analysis to measure the degree of labour rehabilitation and future benefit and costs. To measure the degree of labour rehabilitation, the benefits in the analyses, pension points were used and extrapolated to future benefits in the human investment analysis.

Results

In the CEA we found a halving of expenditure for the KrAmi and Knuff groups and a 25% cut for the two comparison groups (non-custodial treatment) compared to the cost one year before rehabilitation started.

The 15-year investment analysis (CBA) showed expenditure decreases and/or benefit increases of about 2.5 million SEK per individual for the two KrAmi programmes and one non-custodial programme, and of 0.5 – 1.0 million SEK for the Knuff programme and the other non-custodial programme. For the KrAmi investment in rehabilitation, programme costs pay off in 1–1.5 years; for probation in 2.5 – 4 years and for Knuff in 4 years after the intervention. Each crown spent pays back 6.0 – 19.7 crowns, more for KrAmi and Knuff and about half as much for ordinary probation.

Diskussionen om evidensbaserad socialtjänst – en deja-vu upplevelse.

bengt-åke armelius

Går det att utvärdera en sådan verksamhet som psykoterapi och hur skall man kunna göra det? I artikeln reflekterar författaren över hur psykoterapin successivt kommit att baseras på empiriskt prövande kunskaper.

Som medlem av styrelsen för Centrum för Utvärdering av Socialt arbete har jag haft anledning att följa debatten om en evidensbaserad socialtjänst – om än på lite distans. Debatten kan sägas handla om möjligheterna att införa systematiska utvärderingar och empiriskt dokumenterade kunskaper som grund för åtgärder och beslut inom socialtjänsten. Det som slagit mig är att jag känner igen mycket av de argument och de känslor som uttrycks i den allmänna debatten från den tid då vi inom den kliniska psykologin började med psykoterapiforskning. Syftet med denna lilla artikel är att förmedla något av hur utvecklingen av psykoterapiforskning i landet varit under de drygt 20 år som förflutit sedan dess.

Den kliniska psykologin var fram till 1970-talet en hjälpdisciplin till psykiatrin och lutade sig starkt mot den kunskapsmassa som fanns kring psykologisk testning. Man gjorde testningar i syfte att klarlägga vilka förmågor och färdigheter patienter hade. I en del fall försökte man också beskriva och behandla psykopatologi, men det var vanligtvis förbehållet personer med en psykoanalytisk utbildning. Under 1970-talet ändrade sig emellertid bilden så att psykologiska behandlingsmetoder baserade på inlärningspsykologi blev spridda och tillgängliga. Därmed kunde man hävda att psykologisk behandling kunde ske på vetenskaplig grund och inte bara med stöd av praxis eller kliniska teorier. Man kunde faktiskt visa att en persons fobier försvann efter fobiträning, men hur var det med

BÅA är legitimerad psykoterapeut och professor i klinisk psykologi vid Psykologiska institutionen, Umeå universitet. Han leder flera forskningsprojekt som utvecklar metoder och utvärderar effekterna av psykosociala vårdinsatser inom psykiatri, psykosomatik, missbruksvård och ungdomsvård. Han var en av initiativtagarna till svensk psykoterapiforskning under början av 1980-talet. Han är sedan några år ledamot i den lokala fältforskningsenheten UFFE i Umeå och i styrelsen för CUS vid socialstyrelsen.

andra metoders evidens för sin effektivitet? Fanns några bevis för att fobier kunde botas med psykoanalys eller andra former av psykoterapi? Internationellt var frågan inte ny. Redan 1952 publicerade Hans Eysenck en berömd studie där han med hjälp av patientmaterialet vid ett stort sjukhus i England kunde visa att psykoterapi inte hade effekt. Resultaten var inte bättre än vad som kunde förväntas om man bara lät tiden gå och gav självläkningsprocessen tillräckligt med utrymme. Naturligtvis ledde detta till en lång och intressant debatt inom psykoterapiforskningen och det dröjde ända fram till 1980-talet innan man kunde visa att Eysenck bara hade delvis rätt och att psykoterapi faktiskt kan ha en specifik effekt utöver självläkning eller social placebo.

Det intressanta är emellertid att frågan om interventioners vetenskapliga evidens hade ställts och att svar krävdes från utövarna av praxis. Man kan kanske säga att det fanns en förhoppningsfull förväntan från myndigheternas sida. I Sverige startade 1975 en utredning om införande av psykoterapiutbildning vid landets universitet. Kanske skulle det vara ett sätt att styra utvecklingen så att staten tog kontroll över vilka som kunde kalla sig psykoterapeuter och därmed undvika den flora av praktiker och varianter av terapi som fanns i många andra länder. Då uppstod emellertid frågan om vilka terapimetoder som skulle sanktioneras och erbjudas till landets psykoterapeuter. Eftersom universiteten har frihet att utforma sina kurser hur man vill kan man i princip utbilda sig i vilken terapimetod som helst, men Socialstyrelsen har en lista av krav som måste uppfyllas för att man skall få legitimation. Där ingår att terapimetoden skall kunna uppvisa tillräcklig vetenskaplig evidens i internationell forskning. Detta är naturligtvis en maktfråga av stor dignitet och företrädare för olika skolor har all anledning att försöka bevisa sin skolas vetenskapliga underlag. Man kan förmodligen hävda att införandet av en statlig psykoterapiutbildning var det första exemplet på att man inte nöjde sig med beprövad erfarenhet utan också krävde vetenskap som grund för en utbildning som syftar till att ge kunskaper och färdigheter för interventioner inom samhällsmedicin och socialtjänst.

Låt mig nu komma tillbaks till min dejavu upplevelse. Svaren på evidenskravet lät inte vänta på sig. Terapeuterna kände sig hotade och värnade om den terapeutiska ramen precis som man lärt sig i sin utbildning. Många kände också ett stort ansvar för sina patienter och ville inte utsätta dem för den påfrestning som en utvärdering eller forskningsansats skulle innebära. Framför allt var många terapeuter tveksamma till värdet av någon form av forskning eller utvärdering eftersom man ändå inte skulle kunna fånga det väsentliga i terapin. Man såg ingen nytta med forskning utan det var enbart ett kontrollkrav från myndigheter. Några exempel på argument som förekom i debatten:

- En bandspelare i rummet innebär att införa parametrar/avvikelser som förstör relationen
- Patienterna kommer att vara negativa till systematiska utvärderingar och känna sig kränkta
- Man kan aldrig fånga den subtila relationen mellan terapeut och patient, inte ens med en videoinspelning

- Det går inte att mäta effekter av känslomässig natur
- Psykoterapin har ett intentionellt språk medan forskning har ett extentionellt. De kan aldrig mötas.
- Empirisk forskning är positivistisk och influerad av medicin och naturvetenskap och passar därför inte in i psykoterapin. Där krävs en hermeneutisk eller humanvetenskaplig ansats.
- Psykoterapi går ut på att förstå – inte förklara
- Forskning i psykoterapi måste styras utifrån de psykoterapeutiska teorierna och forskningsmetoderna skall vara konsistenta med de kliniska teorierna och metoderna.
- Den forskning som behövs bedrivs av psykoterapeuterna själva under varje timme. Själva terapiprocessen är en form av forskningsprocess som ständigt valideras mot interna kriterier.
- Varje fall är unikt och det går inte att generalisera från en patient till en annan.
- Psykoterapi är inte en rationell process utan styrs av det unika mötet mellan två subjekt och är därför inte möjlig att forska på.
- Man kommer inte att kunna mäta effekter av psykoterapi på ett för patienterna rättvisande sätt. Det är alldeles för komplext.

Som framgår av listan gäller argumenten såväl det vi idag kallar processforskning som effektforskning och man kan tydligt spåra en rädsla för att skada det kliniska arbetet, både för sin egen skull och för patientens skull. Naturligtvis var rädslan befogad. På den tiden fanns ingen terapimetod som förespråkade användande av video, frågeformulär och systematiska intervjuer genomförda av utanförstående. Idag finns dessa inslag med som hjälpmedel i många terapimetoder. Jag minns själv hur rädd jag var när jag genomförde min första inspelning av terapisamtal och hur lättad jag var när patienten kände sig stimulerad av att vara en så viktig person att samtalen spelades in. Vi har fått många anledningar att förvånas över vad patienter tycker om både sina psykoterapier och om forskning i form av systematiska utvärderingar när de väl får chansen att göra sin röst hörd.

Hur gick det sedan? På några orter i landet startades forskningsprojekt med inriktning mot psykoterapi redan under 1970-talet. Ofta involverades psykoterapeuter i arbetet, både i sin roll som psykoterapeuter och som forskare eller doktorander. År 1978 började vi ha årliga möten i det som kallas »Områdesgruppen för psykoterapiforskning«, med stöd från dåvarande HSFR (Humanistisk-Samhällsvetenskapliga ForskningsRådet). Där träffas forskare och intresserade kliniker under ett par dagar för att diskutera psykoterapiforskning och lära sig av varandras arbete. I början av 1980-talet satsade ett par forskningsråd i Sverige på att utbilda en grupp psykoterapeuter till psykoterapiforskare genom att invitera internationellt ledande forskare att berätta om sin forskning och den forskningsmetodik de använde. Det var en stor stimulans som öppnade dörren för internationella kontakter, bl.a. med SPR, Society for Psychotherapy Research, som är ett mycket viktigt forum för psykoterapiforskningens utveckling.

Min bild av utvecklingen är att vi gått igenom tre faser i utvecklingen av svensk psykoterapiforskning. Den första fasen kan beskrivas som en »sökande« fas där vetenskapsteoretiska frågor och diskussioner dominerade. Många av de frågor som listats ovan var aktuella och diskussionerna kunde ibland upplevas som oändliga och förlamande, kanske därför att de ofta handlade om vad som var rätt och fel. Det som drev utvecklingen framåt var enligt min mening att några försökte ta tag i en frågeställning och faktiskt göra något av den. Det kunde vara att försöka mäta känslor eller att intervjua terapeuter om hur de tänkte, och det bidrog till att utvecklingens andra fas, »pröva och lär« kom till stånd. Den egna erfarenheten av att försöka genomföra ett empiriskt forskningsprojekt där man samlar in relevanta data var inkörsporten till att lära sig något om vilka möjligheter och svårigheter som finns inom området. Även andra kunde lära sig av detta och på så sätt skapades en allt större gemensam kunskapsbas, speciellt om man inkluderar de internationella kollegornas insatser. Den tredje fasen skulle jag vilja beskriva som en fas av »integrering«. Idag tror jag man kan säga att psykoterapiforskningen kommit till en platå där vi kan blicka ut över fältet och säga att en hel del av de farhågor som fanns i den första fasen visade sig vara överdrivna. Naturligtvis kan vi aldrig fullständigt beskriva vad som försiggår i en psykoterapi eller exakt vilka effekter som åstadkoms genom psykoterapi. Men vi kan med stor säkerhet säga t.ex. att psykoterapi i allmänhet har en bättre effekt än kontrollbetingelser eller ingen terapi alls. Vi kan också säga att relationen mellan terapeut och patient är av stor betydelse för utfallet av terapin, speciellt patientens känsla av att vara respekterad och av att samarbeta med terapeuten. Vi vet också att psykoterapi inte botar svårare sjukdomar, men att den ändå kan åstadkomma effekter som är av stor betydelse för såväl patienten som för anhöriga. Vi vet också att en del människor inte blir bättre av psykoterapi och det går relativt bra att förutsäga för vilka psykoterapi inte är lämpligt. Resultaten från meta-analyser, eller systematiska kunskapsöversikter, av psykoterapeutiska insatsers effekter pekar också mot att alla väl utvecklade terapimetoder ger ungefär samma effekt. Det finns alltså inte någon metod som är generellt sett överlägsen andra, vilket naturligtvis inte utesluter att vissa metoder är mer effektiva eller lämpliga för vissa typer av patienter eller problem. Utmärkande för den nuvarande integreringsfasen är också att många praktiserande psykoterapeuter upplever psykoterapiforskningen som värdefull för sitt kliniska arbete. Man kan dra nytta av den kumulerade kunskapsmassan och kanske även få tips om vilka effekter man kan förvänta sig om man ger en viss sorts psykoterapi till en patient, eller hur lång tid man behöver räkna med för att nå förväntade effekter. I de två första faserna kan man säga att psykoterapiforskarna gick i kölvattnet på klinikerna och försökte översätta vad de gjorde till mer systematiska studier. I bästa fall kunde man korrigera eller bekräfta den kliniska kunskapsbilden. I den tredje fasen har man kommit ikapp och i vissa avseenden t.o.m. gått förbi klinikerna så att de kan tjäna på att hålla sig informerade om forskningsläget. Ett gott exempel är att vi idag vet hur betydelsefull

den terapeutiska alliansen i form av patientens upplevelse av att känna värme, stöd och respekt från terapeuten är för utfallet av terapin. En hel del missuppfattningar och falska påståenden styrda av ideologiska intressen kan avfärdas därför att de saknar empiriskt stöd. Ett annat sätt att beskriva de tre faserna är att säga att teorier om psykoterapi dominerade i första fasen, forskningsmetoder och mätmetoder den andra fasen och kunskapsintegrering i form av översikter och försök till integrering dominerar den tredje fasen.

Min deja-vu upplevelse går tillbaks till den första fasen som jag beskrivit ovan. Jag vet naturligtvis inte om samma utvecklingsprocess väntar inom socialtjänstens arbete med att evidensbasera sina interventioner, men det finns anledning att tro att all praktikbaserad kunskapsutveckling följer en likartad process. Den utveckling som jag beskrivit ovan kan man återfinna inom många områden där man arbetar med empirisk kunskap. Att bota sjuka har alltid varit viktigt för människor och kunskapsutvecklingen inom både medicin och psykoterapi har gått från tro till empiriskt grundad kunskap. Vi lever i en tid då stora mängder information kan hanteras på ett rationellt sätt och där sofistikerade analyser kan göras med datorers hjälp. Ur mitt perspektiv finns all anledning att se positivt på den utveckling som nu förstärks, både för socialarbetarna själva och deras klienter.

NYA AVHANDLINGAR

Öppna rum - om ungdomarna, staden och det offentliga livet

Björn Anderssons doktorsavhandling vid institutionen för socialt arbete, Göteborgs universitet 2002

Ett hav av människor mötte Björn Andersson när han trädde in i den stora hörsalen på Handelshögskolan i Göteborg för att försvara sin avhandling Öppna rum - om ungdomarna, staden och det offentliga livet. Många var nyfikna och de blev förmodligen inte besvikna eftersom avhandlingen faktiskt hjälper oss att se in i den slutna värld som ungdomars offentliga liv utgör.

Samspel och spänningar mellan offentlighetens form och innehåll

Ämnet för avhandlingen är »ungdomars kollektiva ianspråktagande av offentliga rum samt de strukturer, relationer och aktiviteter som utvecklas i samband med detta«. Avhandlingen kretsar kring offentlighetens form respektive innehåll, vilket redan de olika innebörder begreppet offentlig har givits inom forskningen antyder: å ena sidan »öppen, tillgänglig«, vilket jag ser som dess form, och å andra sidan »gemensam, kollektiv« som mer kan uppfattas som en innehållslig bestämning. Avhandlingen handlar om samspelet och spänningarna mellan offentlighetens form och innehåll. Den behandlar t.ex. den offentliga platsens tillgänglighet som just en form som ungdomar kan använda för att undgå familjens och skolans kontroll. Vidare har mycket av senare års samtal om offentliga platser handlat om innehåll, såsom våld, gäng och rädsla vilka tenderar att göra offentligheten mindre gemensam. I denna avhandling definieras »ungdomsoffentligheten« som ungdomars användning av de öppna och tillgängliga platserna och både om vad det gemensamma livet där betyder för ungdomarna och vad slags gemensamt liv ungdomarna faktiskt skapar där.

Avhandlingen inleds med en serie frågor inom fyra områden: ungdomsoffentlighetens utbredning och mönster; det sociala samspelet inom ungdomsoffentligheten; ungdomsoffentligheten som socialt problem; offentligt liv som läroprocess. Avhandlingens syfte formuleras inte rakt på sak men som jag uppfattar det är syftet att undersöka dessa fyra områden.

Teoretiska perspektiv

Avhandlingens fyra teoretiska kapitel rör sig från mer allmänna aspekter på det offentliga till den mer specifika ungdomsoffentligheten. Framställningen har karaktär av en ibland ganska collageliknande

forskningsgenomgång. Jag kan här endast antyda innehållet och rationaliserar framställningen genom att peka på det som har en avgränsande betydelse i avhandlingen.

I kapitlet Perspektiv på offentligt liv presenteras Weintraubs fyra perspektiv i diskussionen om offentligt och privat. Det första perspektivet har att göra med stat-marknad, det andra handlar om medborgarens deltagande i det offentliga livet genom kollektivt beslutsfattande, det tredje handlar om socialt samspel i det offentliga och i det fjärde, feministiska perspektivet är distinktionen mellan familj och ekonomisk-politisk ordning central. Denna studie faller inom ramen för det tredje perspektivet: socialt samspel i det offentliga. I kapitlet Det offentliga rummet förs ett par intressanta diskussioner och syftet är såvitt jag förstår just att karakterisera det sociala samspelet i det offentliga. Två positioner etableras: Simmels »avskärmningsperspektiv« där individen i storstaden av psykiska överlevnadsskäl kyler ned sin socialbilitet till ett slags saklig nollpunkt (exakt det som Crocodile Dundee inte gör när han kommer till New York, utan han skickar ett »Good day mate« till alla han möter); respektive Loflands »uppmärksamhetsperspektiv« där offentliga platser ses som miljöer för lärande. Lofland är en viktig inspiratör i avhandlingen, inte minst hennes idé att storstadens offentlighet har förutsättningar att skapa människor som tänker och handlar som kosmopoliter. Författaren presenterar emellertid också en del invändningar som gjorts mot detta, nämligen att det i stora städer i och för sig finns en mångfald av människor men att det sällan sker möten över sociala och kulturella gränser.

I kapitlet Offentligheten som social sfär redogörs för olika arbeten om offentlighetens sociala struktur. Central är Loflands karakteristisk av offentligheten som inklusiv, vilken har att göra med att individerna där möts som främlingar - de har helt enkelt inte tillräckligt mycket kunskap om varandra för att differentiera. Detta ställs mot Habermas »borgerliga offentlighet«, vilken snarare har exkluderande karaktär eftersom den förutsätter att aktörerna i offentligheten är utrustade på visst sätt för att kunna agera där. Successivt förs nya röster in i samtalet om offentlighetens karaktär, röster som talar om motoffentlighet, om genusaspekter på offentligheten, folklig offentlighet. Kapitlet avslutas med en redogörelse för Goffmans studier av socialt samspel på offentliga platser.

Jag tror att avhandlingen skulle ha blivit bättre om Goffmans analyser av socialt samspel på offentliga platser använts mer. Exempelvis hade dikotomin avskärmning - uppmärksamhet kunnat brytas ned med hjälp av Goffman, som just beskriver individen på offentlig plats som pendlande mellan vad som fritt översatt kan kallas sysselsättning och beredskap (Relations in Public 1971). Han menar att individen reagerar med en förhöjd uppmärksamhet på sådant som i någon mening bryter den offentliga samspelsordningen, medan hon återgår till sysselsättning när förhållandena framstår som normala. Goffman kombinerar alltså uppmärksamhet och avskärmning genom att se dem som delar av en handlingssekvens.

I avhandlingens centrala teoretiska kapitel, Ungdomsoffentlighet, återges Liebergs definition av ungdomsoffentlighet: »en

social sfär eller gemenskap, som fungerar som ett livsrum för självständig bearbetning och utbyte av upplevelser och erfarenheter tillsammans med andra ungdomar«. Den speciella mening som just ungdomar ger offentligheten understryks, vilket skiljer deras användning av den från vuxna, nämligen att det rör sig om speciella fysiska och sociala rum där ungdomar kan uppleva frihet från föräldrarna och skolan. Just vuxenkontrollen och frigörelsen från den är en central aspekt av avhandlingsförfattarens förståelse av ungdomsoffentligheten. Mot den bakgrunden delar han in de rum som ungdomar använder med avseende på deras olika grad av yttre kontroll. De rum som Björn valt att studera är de som kännetecknas av låg grad av yttre kontroll och där aktiviteterna har kollektiv karaktär, närmare bestämt stadens öppna rum, fritidsgårdar och caféer.

Metod

I avhandlingens metoddel tecknas en bild av Göteborg och de fyra områden där undersökningar gjorts: tre bostadsområden som valts ut därför att de representerar förorten, innerstadsdelen och tätorten samt centrala Göteborg eftersom det är en viktig arena i det som kallas ungdomsoffentligheten. Olika metoder har använts för att fånga de material som avhandlingen bygger på, såväl kvantitativa som kvalitativa. En enkät samlades in 1996 på sju olika skolor i Göteborg och besvarades av 988 ungdomar i åldern 13-18 år. Syftet med enkätundersökningen var att beskriva omfattningen av och karaktären hos ungdomars »uteliv« (som det offentliga livet kallas i enkäten). Vidare innehåller enkäten en del frågor om ungdomarnas egna upplevelser av det offentliga livet. Vidare har en besökarundersökning gjorts på Angeredsfestivalen 1997 liksom ett antal intervjuer, varav två var gruppintervjuer samt en rad mer informella samtal. Slutligen har Björn gjort observationer av s.k. ungdomshelger och andra ungdomsevenemang.

Metoden sägs innebära en triangulering, vilket jag tror förutsätter att man har ett studieobjekt. Frågan är dock om avhandlingen har det? Som framgått används offentlighet - avhandlingens centrala begrepp - i två olika betydelser: tillgänglig respektive gemensam. I enkätundersökningen översätts vidare offentligt liv till »uteliv« och inkluderar det mesta utanför familj, skola och arbetsliv. Relativt olika företeelser sammanförs alltså under beteckningen uteliv: kvarteret, gatan, bostadsområdets centrum, fritidsgården, caféet, Liseberg, ungdomsevenemang. Det talas också om »ungdomsoffentlighet« som en särskild kategori. Detta behöver inte alls vara problematiskt i sig. Problem kan emellertid uppstå när man säger att man triangulerar, som väl inte endast innebär att man arbetar med flera olika metoder, utan rimligen innebär att flera olika metoder används för att studera ett och samma studieobjektet. En annan och större fråga reser sig då: Kan olika metoder ha samma studieobjekt eller består deras olikhet just i att de konstruerar olika objekt? Jag saknar en diskussion om dessa frågor i avhandlingen.

Resultat

Resultatredovisningen har delats upp i fyra kapitel, vilka behandlar olika aspekter av

ungdomsoffentligheten. Två kapitel handlar om ungdomarna i lokal respektive central offentlighet och bygger på resultat från enkätundersökningen, vilka sedan kompletteras med kvalitativt material. I de två följande kapitlen behandlas ungdomshelger och våld på offentliga platser. Dessa kapitel bygger mer på kvalitativt material.

Kvantitativt finns en del skillnader mellan bostadsområden i förort, tätort och innerstad. I förorten är det betydligt mer aktivitet i den lokala offentligheten och på fritidsgården, medan tätorten i detta avseende är mer sovande. Innerstaden hamnar någonstans mellan för- och tätort. Det finns också könsskillnader: pojkar är mer på offentliga platser än flickor. En viktig fråga är vad ungdomarna gör ute? Kollektiva och idrottsliga aktiviteter är Björns första svar, att testa det otillåtna ett annat, att vara med andra är också en viktig aktivitet: »Just detta att umgås och fungera i situationer som präglas av en sorts 'icke-aktivitet' är en av ungdomsoffentlighetens centrala kvalifikationer.«, skriver han. Icke-aktivitet tillsammans med andra, ska kanske tilläggas, och intressant i det sammanhanget är att ungdomarna »intimiserar offentligheten« och gör saker offentligt som kan framstå som opassande, saker som en del vuxna anser att man borde göra i sin privata sfär. Denna intimisering kan förstås mot bakgrund av att ungdomar knappast har någon privatsfär, till och med det egna rummet riskerar alltid att invaderas av föräldrar och syskon.

I avhandlingen diskuteras inte gruppens betydelse för ungdomarna när det just gäller att skapa en egen plats: gruppen blir ju något att gömma sig i, en sluten värld som exkluderar vuxna. Gruppens betydelse lyfts dock fram i två andra differentierande sammanhang: dels när det gäller tillhörighet till livsstil och överhuvudtaget stilar bland ungdomar; dels när det gäller den lokala identifikationen och gruppens ianspråktagande av ett territorium som skiljer den från andra grupper.

I kapitlet Stadens centrala offentlighet undersöks ungdomars rörelser i centrala Göteborg, i synnerhet deras användning av caféer, biografer, nöjesfältet Liseberg etc. dit tillgängligheten i hög grad är reglerad av pengar. Det finns alltså två nivåer i offentligheten: den öppna och den begränsade. Med hjälp av enkätmaterialet visas vidare att det finns både ålders- och könsskillnader och skillnader beroende på bostadsort och studieinriktning i gymnasieskolan mellan ungdomar som rör sig i central och lokal offentlighet.

I kapitlet om ungdomshelger redogör författaren för en rad observationer han gjort vid det slags ungdomskarnevaler som har en särskild rituell ordning, bl.a. när det gäller att pröva och passera gränser. Av bl.a. det skälet är ungdomarna för sig själva och utanför vuxnas kontroll, i frirum, samtidigt som olika intervenerare också vid dessa tidpunkter finns ute och skapar en viss vuxenkontroll.

I kapitlet Offentlighetens hårda kant behandlas våldsproblematiken i det offentliga. Det handlar huvudsakligen om killar som är socialt problematiska, ofta från svåra hemförhållanden och likgiltiga inför skolarbete, och Björn baserar sin undersökning på gruppintervjuer med några killar som ingår i ett gäng i Mölnlycke. Detta kapitel är på flera sätt avhandlingens höjdpunkt,

inte minst beroende på att dessa killar har skapat en nyfikenhet kring sig som här blir stillad. Det är två scener som är viktiga när det gäller gängets rörelser, centrum i Mölnlycke och skolan, och gemensamt för båda är att de är mötesplatser för många olika människor. För gänget blir dessa platser i hög grad en scen för våld, medan många andra, i synnerhet flickor, bär på en rädsla för våldet på dessa platser och därför lär sig att hantera dem genom bl.a. undvikande. Samtidigt visar Björn att upplevelsen av våld varierar starkt beroende på var man bor, ungdomarna från förorten rapporterar flera våldsincidenter än ungdomarna i tätort.

Bra avhandling som fyller ett tomrum

I det avslutande kapitlet besvaras avhandlingens inledande frågor. När det gäller det offentliga livets utbredning och mönster finner författaren fyra olika kategorier bland de ungdomar han studerat. De uteorienterade är i hög grad högstadieelever, pojkar och lägenhetsboende. De är »offensiva utövare av offentligt liv«: syns och hörs på ett maskulint sätt. De urbant orienterade går företrädesvis i gymnasieskolan och utgörs i hög grad av flickor som i sitt identitetsarbete inkorporerar platser och verksamheter i den centrala offentligheten. De lokalt orienterade är högstadieelever, främst killar som är aktiva i det egna bostadsområdet. De privat orienterade besöker inte offentliga platser och många av dem känner obehag inför våld eller hot om våld på offentliga platser. Trots att de här fyra kategorierna endast representerar 1/4 av dem som besvarat enkäten, anser jag att de ökar vår förståelse av ungdomar i offentligheten.

Avhandlingen kretsar också kring ungdomars läroprocesser i det offentliga. Det handlar såvitt jag förstår i hög grad om att lära sig att läsa staden och mot den bakgrunden tillämpa olika undvikandestrategier. Men är inte det faktum att vissa ungdomar lär sig hantera det offentliga på andra sätt, t.ex. genom våld, också en läroprocess? Här tycker jag att författaren borde ha utnyttjat den amerikanske sociologen Elijah Andersons omfattande undersökningar av »gatans lag« och utvecklingen av »street wisdom«. Intressant är också att Elijah Anderson skiljer mellan två kategorier bland de ungdomar och familjer han studerar: »street families« och »decent families«. Street är sådana som lär sig hantera gatan på dess villkor, decent är sådana som försöker undvika gatan. Med iakttagande av skillnaden mellan Göteborg och amerikansk ghettomiljö, skulle Andersons, alltså Elijah, analys ha kunnat kopplats till och fördjupat de fyra orienteringar bland ungdomar i offentligheten som Björn upptäckt.

När det gäller frågan om det sociala samspelet inom ungdomsoffentligheten menar författaren att ungdomarnas betoning av relationer är slående. Han noterar en utveckling i ungdomsgruppen, som ju befinner sig i puberteten, från att göra till att vara (detta är ju för övrigt vad barn börjar säga i puberteten: inte längre leka, utan vara med kompisar). Relationerna går emellertid i två riktningar: en »solidarisk« som handlar om att uppleva sig som del av en sammansvetsad ungdomsgrupp; en differentierande som handlar om att markera tillhörighet till grupper, stilar etc. och på det viset erövra en identitet.

När det gäller det problematiska offentliga livet konstaterar Björn att upplevelsen av våld är relativt utbredd och att de som känner rädsla främst är flickor och ungdomar som inte vistas så mycket på offentliga platser. Det finns en läroprocess förknippad med dessa upplevelser nämligen att lära sig undvika våld. Men det finns också de coola eller tuffa som med hjälp av våld vänder upp och ned på den samhälleliga ordning där många av dem är offer och i stället blir herrar. För dem fyller offentligheten ett kompensatoriskt behov, menar författaren.

Redan i inledningen konstateras att det finns skilda uppfattningar vad gäller omfattningen av våld på offentliga platser. Media förmedlar bilden av ganska farliga offentliga platser och Björn menar att även intervenerare, socialarbetare och nattvandrande föräldrar exempelvis, har denna syn. Forskningsresultat som återges visar emellertid att någon ökning av våldet inte skett. Författarens egna undersökningar visar att en stor del av ungdomarna bevittnat våldsamheter och att en relativt stor del av de något äldre ungdomarna varit inblandade i våldsamheter. Vad ska vi egentligen säga om våldet? Är det så enkelt att de olika buden har att göra med skillnader mellan metoder och vetenskapssyn: en skillnad mellan det kvantitativt mätbara våldet och det socialt definierade våldet? Avhandlingen borde ha innehållit en ordentlig diskussion om våldet på offentliga platser.

Offentligt liv, slutligen, är på flera olika sätt en läroprocess. Som redan nämnts lär sig en del ungdomar att läsa det offentliga livet för att kunna undvika våld. Men samspelet i det offentliga handlar också om att lära sig fungera socialt. Och här, allra sist, menar Björn att vuxna som intervenerar i ungdomsoffentligheten måste göra det mer på ungdomarnas villkor: de måste respektera ungdomsoffentligheten som en social sfär där relationer mellan ungdomar står i centrum och därför är ett viktigt inslag i socialisationsprocessen.

På det hela taget är det lärorikt att läsa Björn Anderssons avhandling. Dess svaga sida är att den greppar över litet väl mycket, vilket minskar möjligheten till djupare analys. Dess starkaste sida är undersökningen av ungdomsgänget i Mölnlycke. På det hela taget en bra, spännande och »nyttig« avhandling som fyller ett tomrum och ger ett bidrag till förståelsen av ungdomar i offentligheten. För egen del hade jag önskat att avhandlingen hade skrivits för några år sedan när jag skrev min bok om social kompetens. Jag sökte då efter just en svensk studie av socialt samspel på offentliga platser och tyckte att det var ett fattigdomsbevis att någon sådan inte fanns. Björn har verkligen fyllt ett tomrum och gjort det på ett bra och tankeväckande sätt.

Anders Persson, fakultetsopponent.
Forskare vid Arbetslivsinstitutet i Malmö

Medvetenhet om bemötande. En studie om sjuksköterskans funktion och kompetens i närståendeundervisning

MadeleineBerghsdoktorsavhandling vid Göteborgs universitet 2002

Madeleine Bergh, sjuksköterska och vårdlärare, ger oss i sin nyligen utgivna avhandling en inblick i sjuksköterskeyrkets pedagogiska funktion. Avhandlingen är utgiven vid Göteborgs universitet i serien; Göteborgs Studies in Educational Sciences 171. Avhandlingen som har titeln Medvetenhet om bemötande. En studie om sjuksköterskans funktion och kompetens i närståendeundervisning omfattar 250 sidor. Denna avhandling distribueras av ACTA Universitatis Gothoburgensis, ISBN 91-7346-423-6 [1].

Inledning och syfte

Studien som är kvalitativ beskriver närståendes/anhörigas ökade omsorgsbörda i samband med att en nära anförvant drabbats av sjukdom eller skada. Flera studier av senare dato fokuserar just närståendes utsatthet vid anhörigas sjukdom. Ett av skälen till detta är att allt fler behöver hjälp av närstående i vård- och omsorgssituationer p.g.a. att samhället allt mer förlitar sig till frivilliga insatser från närstående. I takt med de förändringar som skett inom vården med dagens avancerade medicinska behandlingar, kortare vårdtider och stigande hemvård ökar också behovet av information och undervisning till närstående. Kraven på sjuksköterskans pedagogiska kompetens ökar till följd av att patienter och närstående behöver både allmänna och mer specifika kunskaper om de omständigheter som råder i varje enskilt fall. Sjuksköterskan har en historisk tradition att undervisa, trots detta har inte dagens situation tidigare studerats i Sverige. Behovet av forskning uppstår därför inom ett individ- och samhällsperspektiv med fokus på såväl närståendes erfarenheter av information och undervisning, men också med fokus på sjuksköterskans pedagogiska kompetens och undervisningsfunktion. Syftet med Berghs studie var därför, med utgångspunkt från ovanstående, att begreppsliggöra mötet mellan sjuk-sköterska och närstående genom att:

- Beskriva närståendes erfarenheter av information och undervisning.
- Beskriva sjuksköterskans pedagogiska funktion och kompetens i mötet med närstående.

Begrepp som rehabilitering, anhörig, närstående, undervisning, information, kompetens och förhållningssätt är mer eller mindre centrala i avhandlingen. I avhandlingens bakgrund finns en redogörelse över närståendes situation i dagens samhälle, såväl internationellt som nationellt. Stu-

1 Avhandlingen utgör en del av projektet «Sjuksköterskans pedagogiska funktion och kompetens i en förändrad vård och utbildningsorganisation«, som genomförs vid institutionen för vårdpedagogik, Göteborgs universitet.

dien anknyter till närståendefokuserad forskning samt studier med inriktning mot sjuksköterskans pedagogiska verksamhet. I en analys av litteraturen konstateras att närstående givit uttryck för att vilja ha mer information, kunskap och stöd från personalen. I studier rörande sjuksköterskans kompetens framkommer brister i bl.a. kommunikation, information och uppmuntran av närstående.

Metod och analys

De datainsamlingsmetoder som använts i avhandlingen är tre och utgörs av fokusgruppsintervjuer, observationer och intervjuer. Den klassiska Grounded Theory modellen som har sin grund i den symboliska interaktionismen har använts som utgångspunkt för datainsamling och analys. Grounded Theory innebär i korta drag att genom forskning generera begrepp och påvisa relationer dem emellan. Grounded Theory som metod presenterades i mitten av 1960-talet och har därefter utvecklats åt två håll. Den klassiska respektive den modifierade. Centralt inom Groundet Theory är dock att alla händelser som kommer i forskarens väg är data, dessa data genererar begrepp som kan framkalla mönster och teorier. Avsikten med Berghs studie var att få en allsidig bild av närståendeundervisningen. Tonvikten lades därför vid förståelsen av de sociala interaktioner som uppkommer i mötet och undervisningen.

Fokusgrupper användes för att erhålla information om närståendeperspektivet. Till första gruppen valdes närstående som var utbildade sjuksköterskor med såväl yrkeserfarenhet som erfarenhet av att vara närstående. Till andra gruppen valdes närstående utan sjuksköterskeutbildning. Urvalet till fokusgruppen var strategiskt, d.v.s. endast närstående med erfarenhet av att vårda närstående valdes. Gruppen närstående sjuksköterskor utgjordes av 13 personer och övriga närstående av 11 personer. En frågeguide med varierande temaområden var vägledande under träffarna, dessa bandades och transkriberades så ordagrant som möjligt. Exempel på frågeområden var: begreppet närstående/anhörig, information tilltro, beteende hos berörda etc. Observationerna genomfördes på två skilda rehabiliteringsavdelningar, en avdelning där yngre patienter vårdades i flera månader efter olyckor, och en avdelning där äldre patienter vårdades kortare tid för ortopediska åkommor. Observationerna varade under fem månader, perioden oktober 1999 till februari 2000, med i genomsnitt ett observationspass per vecka. Fokus för observationerna var sjuksköterskorna, särskilt i deras kontakt med närstående. Under observationerna fördes fältanteckningar. Intervjuer genomfördes med totalt 10 sjuksköterskor som strategiskt valdes utifrån kriterier som utbildning, ålder, kön och yrkesverksam tid. Intervjuerna var semistrukturerade i form av en intervjuguide, intervjuerna spelades in på band.

Analys och tolkning genomfördes i en process där de olika stegen inte var helt enkla att urskilja, eftersom insamling, bearbetning och analys i stor utsträckning skedde samtidigt. Fokusgrupperna analyserades fortlöpande för att ligga vilande då observationerna och intervjuerna analyserades. Dessa analyserades med stöd av constant comparative analysis. Analysen omfattade substantiv kodning, d.v.s. en

öppen kodning för att fånga substansen och kategorierna samt en teoretisk kodning som går ut på att identifiera så många relaterande begrepp eller begrepps-kontakter som möjligt.

Resultat och diskussion

Resultatet redovisas i två huvudteman; bemötande ur närståendeperspektiv samt bemötande ur sjuksköterskeperspektiv. Bergh använder sig av en övergripande kärnkategori där begreppen medvetenhet och bemötande vuxit fram för att beskriva sjuksköterskans pedagogiska funktion i mötet. Vidare nämns förhållningssätt som en underkategori till ovanstående begrepp. Utifrån detta påvisas mer konkret hur de närståendes förhållningssätt kan variera. Bergh gör en indelning av aktiva närstående som tar initiativ, engagerar sig och söker svar på sina frågor direkt eller via ombud. Som motsats till detta finns de passiva närstående som intar en låg profil, håller sig undan och på så sätt förbigås. Detta förhållningssätt är det samma i relation till såväl sjuksköterskan som andra samhällsrepresentanter. I mötet med den närstående agerar den aktivt närstående i syfte att förbättra situationen för patienten. De passivt närstående är de som inte aktivt söker information och kunskap, de agerar inte utifrån den närstående patientens behov utan styrs av andra behov. Det pedagogiska förhållningssättet, informationsöverföringen och kunskapsutbytet avspeglas genom att: Vårderfarna närstående samt yngre närstående är mest aktiva i sitt informationssökande, de kräver information, de går ofta vidare och söker information på annat håll och på egen hand. Icke vårderfarna närstående har en stor tilltro till auktoriteter, de ifrågasätter inte informationen utan tar den som sanning.

Ytterligare uppdelning som Bergh gör är en kategorisering av närstående i första ledet och närstående i andra ledet. Om närstående i första ledet är aktiva sätter de fokus på praktiska och administrativa behov, är de passiva så agerar de ej. De aktiva närstående i andra ledet sätter fokus på socioemotionella och existentiella behov.

När perspektivet skiftar till att gälla den yrkesverksamma sjuksköterskans förhållningssätt och pedagogiska ställningstagande, nämns begreppen, handlande och relation. Handlande, belyser sjuksköterskans aktivitetsgrad i bemötandet av närstående. Relation, speglar sjuksköterskans reflekterande eller distanserande i mötet. Liksom de närstående kategoriseras även sjuksköterskorna i grupper, de som agerar aktivt respektive de som agerar passivt. De aktivt agerande sjuksköterskorna fungerar på olika sätt, man kan dock säga att de är initiativtagande, lyssnande, informerande och empatiska. Att vara passivt agerande innebär att de inte tar initiativ, undviker kontakt eller distanserar sig och tar ett intellektuellt och känslomässigt avstånd i relationen till närstående. Pedagogiskt innebär detta att den aktivt agerande sjuksköterskan har strategier för information, undervisning och kunskapsutbyte. Denna sköterska inhämtar information om motpartens kunskapsbehov genom att bedöma vilken kunskap som efterfrågas. Strategier som används är att etablera kontakt, att utgå från patientens önskemål och bekräfta motparten. De vägleder och stämmer av

genom att följa upp.

Även den passiva sjuksköterskетypen varierar i sitt förhållningssätt och pedagogiska tillvägagångssätt. Generellt kan sägas att denne inte själv tar kontakt med de närstående utan de närstående får söka upp dem. Sköterskan är inte passiv i den meningen att hon inte utför något arbete, utan passiv i den bemärkelsen att hon vet och har reflekterat över närståendes behov av information och undervisning men inte kunnat handla. Bland annat beroende på att hon styrs av rutiner och regler. Bergh talar i sammanhanget om en oprefessionell hållning, där sjuksköterskorna utvecklat icke pedagogiska strategier i mötet med närstående. Sådana strategier skulle kunna vara att ej bekräfta närståendes initiativ, att de är uppgiftsstyrda och inte reflekterar över det egna agerandet.

Ur ett pedagogiskt perspektiv beskrivs etablerandet av kontakt och mötet utifrån följande didaktiska frågor[2]:

- Var, befinner sig den närstående kunskapsmässigt?
- Vad, med avseende på kunskapsinnehåll?
- Vem, fokuserar närstående och patient?
- Hur, utformas sjuksköterskans förhållningssättochpedagogiskatillvägagångssätt?

Slutord

I den avslutande diskussionen redogör Bergh för några begrepp som tagit form ur studien. Bemötande är ett sådant centralt begrepp som utgör grunden för att beskriva sjuksköterskans pedagogiska funktion och närståendes behov av kunskap. Förhållningssätt är ytterligare ett begrepp som tydliggörs i texten och har att göra med såväl närståendes som sjuksköterskans förhållningssätt. Detta begrepp som hänger samman med erfarenhet, medvetenhet och kompetens återfinns i mötet. Strategier som tillvägagångssätt är också viktiga i informations- och kunskapsutbytet. Vårderfarenhet, ålder, relation och i viss mån kön har visat sig ha betydelse för informations- och kunskapsutbytet

I resultatet framkom också att förhållningssättet till närståendes skulle kunna förbättras, dels för gruppen aktiva närstående som ibland söker andra vägar till information, men också för gruppen passiva närstående som uppfattas vara i stort behov av vägledning. Jag tror att denna insikt har betydelse inte bara för sjuksköterskan utan också för andra vård- och omsorgsprofessioners insikt i mötet med närstående.

Syftet med studien var att begreppsliggöra mötet mellan sjuksköterska och närstående inom rehabiliteringsvården samt att beskriva sjuksköterskans pedagogiska funktion och kompetens. Detta har Bergh lyckats med i en avhandling som är både välstrukturerad och välskriven, intressant att läsa för oss inom vård- och omsorgsprofessionerna. Avhandlingen fokuserar mötet mellan närstående och sjuksköterskor på ett i mitt tycke övergripande sätt. Jag saknar dock explicit en pedagogisk koppling till såväl teori som data. Personligen tror jag att avhandlingen skulle ha vunnit på en tydligare sådan pedagogisk koppling

2 Didaktik kan betraktas som läran om undervisning och kan sammanfattas i tre frågor vad? varför ? och hur? (Arfwedson 1992, Didaktik för lärare. Stockholm: HLS förlag).

där framförallt närståendeundervisningen, avgränsats. I föreliggande studie ger närståendeundervisningens vida definition en oklar bild av sjuksköterskeyrket som omvårdnadsprofession. Skulle det inte lika väl kunna vara en social- eller pedagogiskprofession blir min undran. Frågan blir därför om sjuksköterskans pedagogiska kompetens tydligare skulle kunna studeras inom ett mer avgränsat område, t.ex. någon situation/moment i sjuksköterskans arbetsområde som innehåller undervisning mer explicit.

Jag ser denna avhandling som ett viktigt bidrag till den vårdpedagogiska forskningen, som ett inslag i den närståendedebatt som förs i samhället samt som ett bidrag till sjuksköterskeyrkets pedagogiskt/didaktiska utformning.

Thomas Strandberg
Fil.mag. Socialt arbete. Vårdlärare.
Doktorand i handikappvetenskap.
Institutet för handikappvetenskap - IHV.
Linköping och Örebro universitet.
E-post:thomas.strandberg@ivo.oru.se

Möten med marknaden. Tre svenska fackförbunds ageranden under perioden 1945-1976

Carina Gråbackes doktorsavhandling vid Göteborgs universitet 2002

Har fackförbunden bromsat tillväxten?
I en nyutkommen doktorsavhandling av ekonomisk-historikern Carina Gråbacke ställs flera frågor om fackföreningsrörelsens drivkrafter och samhällsekonomiska betydelse på sin spets: Ska fackförbund uppfattas som väloljade särintressen som påverkar samhällsekonomin negativt eller som organisationer som tvärt om gynnar ekonomisk tillväxt och bidrar positivt till välfärden? Avhandling utgör ett värdefullt bidrag till forskningen om arbetsmarknadsorganisationer i allmänhet och fackliga organisationers beteende i synnerhet.

Forskningsintresset för partsrelationer på arbetsmarknaden kan bland annat ses mot bakgrund av ett trendbrott i tillväxt och arbetsmarknadsrelationer från slutet av sextiotalet och början av sjuttiotalet. På ekonomiska svårigheter och tilltagande konflikter på arbetsmarknaden följde successivt en ny och mer marknadsorienterad syn på ekonomisk politik. Höga skatter, marknadsregleringar och starka intresseorganisationer gavs skulden för ekonomins kräftgång, lägre tillväxt och påtagligt högre inflation.

Det faktum att de ekonomiska svårigheterna i flera västerländska länder förknippades med starka intresseorganisationer bidrog till ett ökat intresse för frågor som anknyter till klassisk politisk ekonomi. Den grundläggande utgångspunkten är här att marknadsekonomins inte kan studeras iso-

lerat från de lagar, regler och normer som påverkar ekonomiska aktörers beteenden.

Studier om institutioner och regelverk på arbetsmarknaden omfattar något förenklat två skilda utgångspunkter. För det första hävdas att marknaden i sig själv skapar optimala villkor för produktion och konsumtion. Externa faktorer – organisationer och politiska ingrepp – ger negativa återverkningar. Den oreglerade marknaden präglas av harmoniska förhållanden och jämlika förutsättningar. För det andra hävdas att marknaden i sig själv är instabil och präglas av ojämlika maktförhållanden. Kollektivavtal och regleringar är ett sätt att lösa konflikter. Avtalen minskar kostnader som följer av osäkerhet och informationsbriser och underlättar samarbete mellan marknadsaktörer. Arbetsmarknadens institutioner minimerar därmed de effektivitetsförluster som skulle ha uppstått på en helt oreglerad marknad.

Särintressen?

I centrum för Carina Gråbackes avhandling står det fackliga agerandet på förbundsnivå under perioden 1945–76; hur detta agerande påverkats av förändrade branschvillkor och spänningar i förbundens relationer till den centrala nivån (det vill säga LO). Den teoretiska utgångspunkten bottnar i public choice och den amerikanske nationalekonomen Mancur Olsons teorier om intresseorganisationers betydelse och beteendemönster. Det betyder att konventionell ekonomisk teori utnyttjas för att förklara organisationers handlingsmönster. Kollektiva aktörers beteenden förklaras med hjälp av den ekonomiska teorins rationalitets- och maximeringsantaganden. Intresseorganisationer och dess ledning maximerar inte vinster, men väl storlek, det vill säga medlemsantal. Precis som i traditionell ekonomisk teori uppfattas fackliga organisationer som ett slags karteller som syftar till att kontrollera arbetsutbudet för att upprätthålla medlemmarnas status och löner. Inflytandet kanaliseras både i förhandlingar med arbetsgivare och via påverkan på statsmakten. Resultatet blir att löner och priser pressas upp ovanför de nivåer som skulle ha etablerats på en fri konkurrensmarknad. Effektivitetsförluster i form av lägre produktionsnivåer och högre varupriser får bäras av skattebetalare och konsumenter.

Mancur Olson skiljer mellan »smala« och »heltäckande« organisationer. En smal organisation – som organiserar en liten del av arbetsmarknaden – har rationella motiv att påverka löner och produktionsvillkor på ett sätt som inte är effektivt ur samhällsekonomisk synpunkt eftersom de negativa effekterna mest drabbar personer utanför organisationen. En bredare organisation vars medlemmar representerar en större del av samhällsekonomin har däremot anledning att ta hänsyn till vad som är samhällsekonomiskt effektivt. Detta helt enkelt därför att medlemmarna i större utsträckning får känning av de effektivitetsförluster som följer av en snävt inriktad facklig strategi.

Tre fallstudier

I fokus för Carina Gråbackes avhandling står tre fackförbund: textilarbetarförbundet, byggnadsarbetarförbundet samt handelsanställdas förbund. Förbundsurvalet återspeglar olika marknadsförhållanden: textilindustrin var en konkurrensutsatt

bransch på tillbakagång, handeln en växande bransch inom den privata tjänstesektorn och byggnadssektorn en politiskt genomreglerad och skyddad bransch. I avhandlingen analyseras hur dessa förbund påverkades och agerade i relation till ändrade marknadsförhållanden. I relation till det övergripande syftet formuleras två delfrågor: 1. Hur förhöll sig förbunden till den fortgående branschrationaliseringen? 2. Hur förhöll sig förbunden till LO som i linje med den Rehn-Meidnerska modellen strävade efter att forcera strukturrationaliseringen inom näringslivet? LO:s strategi gick ut på att arbetskraft och kapital skulle frigöras från lågproduktiv verksamhet till företag och branscher med ett högre förädlingsvärde per arbetad timme. Det skulle möjliggöra högre tillväxt och indirekt ökade reallöner. Men en sådan strategi förutsatte också statlig medverkan i form av arbetsmarknadspolitik och socialpolitiska insatser som underlättade överföringen av arbetskraft från lågproduktiv till högproduktiv verksamhet. De risker som enskilda individer mötte i form av utslagning och arbetslöshet skulle bäras gemensamt via skattefinansierade välfärdspolitiska insatser.

Avhandlingen domineras av tre fallstudier där förändringarna i de tre branscherna beskrivs och bildar utgångspunkt för en diskussion om fackförbundens agerande. Det fanns ett gemensamt mönster i utvecklingen. Samtliga tre branscher kännetecknades av tilltagande produktivitetsproblem, något som bland annat yttrade sig i akuta lönsamhetsproblem under andra hälften av sextiotalet. Branscherna drabbades hårt av den solidariska lönepolitikens effektivitetspress. För textilindustrins del blev resultatet omfattande företagsnedläggningar i spåren av hård internationell konkurrens, växande lågprisimport och lönekostnadsökningar som inte alls var anpassade till branschens betalningsförmåga. Byggnadsbranschen hölls under armarna via politiska regleringar och växande offentliga subventioner, men tvingades samtidigt till hårda rationaliseringar, tilltagande storskalighet och industriella produktionsmetoder. Fackets ställningstaganden till branschfrågor påverkades givetvis i hög grad av de stora fackliga byggföretagen:SvenskaRiksbyggenochBPA.Dagligvaruhandeln bars upp av en inkomstelastisk efterfrågan och antalet förvärvsarbetande i parti- och detaljhandeln var förhållandevis stabilt från början av femtiotalet fram till mitten av sjuttiotalet. Men även handeln genomgick kraftiga rationaliserings- och koncentrationsrörelser. På nya självbetjäningsbutiker, kedjevaruhus och stormarknader följde tilltagande storskalighet och ökad kapitalintensitet.

Fackförbundens reaktioner på de förändrade villkoren inom respektive bransch följer också ett gemensamt mönster. I början av femtiotalet var flera förbund skeptiska till den branschrationaliseringslinje som LO förordade. Textilarbetarförbundet var till exempel starkt pådrivande för ökat tullskydd. Handelsanställdas förbund (handelsarbetarförbundet före 1957) var inte med på LO:s kritik mot bristande konkurrens och höga priser inom handeln. Man betonade istället olika fördelar med en småskalig handelsstruktur, framför allt ur servicesynpunkt. Men vid övergången från femtiotalet till sextiotalet intog

samtliga tre fackförbund en positiv syn på branschrationalisering. Organisationerna slöt alltså upp bakom LO och den Rehn-Meidnerska modellen. För textilarbetarförbundets del innebar det att man ställde sig bakom en frihandelspolitik, trots att det skulle innebära svårigheter för förbundets medlemmar. Villkoret för detta stöd var låglönesatsningar i kombination med aktiva insatser på det arbetsmarknadspolitiska och socialpolitiska området. Det var först mot slutet av sextiotalet och början av sjuttiotalet som förbunden återigen började ställa krav på branschskydd och offentliga ingrepp för att begränsa konkurrensen och garantera sysselsättningen. Denna omsvängning skedde parallellt med att LO började distansera sig från den hårdföra rationaliseringslinjen.

Spänningen mellan LO och de enskilda förbunden i synen på branschrationalisering utgör det centrala temat i avhandlingen. Det var inte självklart att förbunden skulle ställa sig bakom en utveckling som i praktiken innebar att den bransch förbundet representerade skulle utsättas för ett allt hårdare rationaliseringstryck.

Ett rikt empiriskt arbete

Carina Gråbacke har presterat ett rikt empiriskt arbete och avhandlingen ger ny och viktig kunskap om motiven för de behandlade fackförbundens ställningstaganden i flera strategiska branschfrågor. Dessutom tydliggörs intressekonflikter mellan förbundsnivån och den centrala huvudorganisationen på områden som inte fått så stor uppmärksamhet tidigare.

Men även en bra bok har sina brister. Frågan är om inte den teoretiska ansatsen hade kunnat utvecklats mer. Det finns en hel del forskning som går ut på att förklara fackliga organisationers beteende och betydelse i ekonomin. Enligt läroboken i ekonomi bidrar fackliga organisationers inflytande på arbetskraftsutbud och lönebildning till att försämra arbetsmarknadens funktionssätt, vilket har ett pris i form av minskad sysselsättning, lägre tillväxt och sämre välfärd.

Detta perspektiv bortser emellertid helt från frågor om organiseringens drivkrafter i ojämna maktförhållanden och de »imperfektioner« och kostnader som är förknippade med marknadsutbyte. Givet ojämna maktförhållanden och stora rörlighetskostnader på arbetsmarknaden blir bilden av fackliga organisationers välfärdsekonomiska och samhällsekonomiska betydelse en annan. Fackföreningarna bidrar bland annat till stabilare anställningskontrakt och underlättar därmed teknisk förändring. Via stabilare och långvarigare anställningar dämpas motståndet mot förändring, den funktionella flexibiliteten ökar. Företagens kostnader för arbetskraftsomsättning och personalrekrytering reduceras och investeringar i utbildning underlättas. Detta ger en helt annan betydelse för de fackliga organisationernas handlingsmönster än den som förutsägs i klassisk ekonomisk teori. Teorier om socialt kapital och betydelsen av samarbete för att skapa förtroende och ömsesidigt utbyte skulle kunna ges en liknande innebörd.

En annan fråga gäller de fackliga organisationernas motivbild. Traditionellt har det handlat om olika kombinationer av löneutveckling och sysselsättning som ger maximal utdelning i form av medlemstill-

strömning och organisatorisk stabilitet. Sociologen Claus Offe har emellertid betonat svårigheterna med rationalitetsantaganden vad gäller fackliga organisationers handlingsmönster.[1] Det går inte att urskilja något tydligt mål. För den enskilde löntagaren handlar det hela tiden om parallella mål: konsumtionsnivå, tryggad sysselsättning och arbetsförhållanden/miljö. Vad som är utslagsgivande är inte alls givet. Och det är inte alls givet hur dessa mål ska förstås på aggregerad nivå i den fackliga organisationen. Fackliga organisationer formulerar mål som inte alls behöver överensstämma med enskilda medlemmars omedelbara intressen. Solidaritet och disciplinering under något överordnat socialt och politiskt mål blir därmed centralt. Peter Swensson har också betonat svårigheterna att förena de mål som påverkar fackliga organisationers verksamhet på olika nivåer: löneutjämning, sysselsättningstillväxt och en ökad löneandel.[2] Teoretiskt är det svårt att avgöra vilket mål som väger tyngst och alltså väljs på de andra målens bekostnad. Det val som sker beror ofta på strukturella förhållanden: ekonomiska villkor, utrikeshandelsberoende och politiskt inflytande (relation till regeringsmakten).

EFO-modellen?

De fackliga organisationerna har knappast varit omedvetna om de negativa följderna av ohämmade krav på löner och branschskydd. Ett av uttrycken för detta var EFO-modellen som lanserades i växlingen mellan sextio- och sjuttiotalen.[3] Carina Gråbacke diskuterar över huvud taget inte detta dokument. I rapporten sammanfattade några ekonomer på fackförenings- och arbetsgivarsidan förutsättningarna för en samhällsekonomiskt hållbar lönebildning. Den grundläggande utgångspunkten var uppdelningen mellan den konkurrensutsatta (k-sektorn) respektive skyddade sektorn (s-sektorn). För att garantera tillväxt och samhällsekonomisk balans måste lönebildningen avvägas efter det utrymme som fanns i privat konkurrensutsatt sektor. I den konkurrensutsatta sektorn var varupriserna givna av de internationella konkurrensförhållandena medan den skyddade sektorn hade större möjligheter att skjuta över ökade lönekostnader på konsumentpriserna. Huvudprincipen blev att det var produktivitetsökningen i kombination med inflationen i omvärlden som satte ett tak för löneökningarna i k-sektorn. Detta tak skulle sedan vara normerande för lönebildningen i s-sektorn.

Uppdelningen mellan k- och s-sektorn hade också varit ett viktigt inslag i de tidiga diskussionerna om den solidariska lönepolitiken. Löneökningsutrymmet i konkurrensutsatt verksamhet uppfattades som väldigt begränsat. Det ansågs inte rättvist att förbunden inom den skyddade sektorn tvingade fram höga löneavtal som i slutän-

1 Claus Offe & Helmut Wiesenthal, »Two Logics of Collective Action«, i Claus Offe (ed), Disorganized Capitalism. Contemporary Transformations of Work and Politics, Cambridge 1985.

2 Peter Swenson, Fair shares. Unions, pay, and politics in Sweden and West Germany, London 1989.

3 Gösta Edgren, Karl-Olof Faxén & Clas-Erik Odhner, Lönebildning och samhällsekonomi. Rapport från expertgrupp tillsatt av LO, SAF och TCO, Stockholm 1970.

dan fick bäras av samtliga löntagare i form av ökade varupriser. Under mellankrigstiden hade också lönerna ökat snabbare inom skyddade sektorer än inom konkurrensutsatt industri. Detta var bakgrunden till att metallarbetarförbundet började ställa krav på att LO skulle få stärkta lönepolitiska funktioner för att därmed skapa förutsättningar för en »solidarisk« löneutveckling och förhindra att skyddade sektorer gick sina egna vägar på konsumenternas bekostnad.[4]

Makroekonomisk ram

En fråga som naturligtvis aktualiseras i anslutning till avhandlingen är vilka kriterier som ligger till grund för urvalet av branscher. Carina Gråbackes framställning saknar en makroekonomisk ram. Framställningen och läsarens förståelse hade vunnit på om den period som undersöks och de branschområden som behandlas satts in i ett bredare ekonomiskt sammanhang. Perioden täcker »guldåren« respektive den tilltagande strukturkrisen från slutet av sextiotalet. Vad kännetecknade denna period i termer av tillväxt, inflation och förutsättningar på arbetsmarknaden? Hur påverkade de skiftande ekonomiska förutsättningarna de fackliga organisationernas respektive LO:s agerande?

Branschernas förutsättningar och problem under tidsperioden skulle med fördel kunna beskrivas i kvantitativa termer: sysselsättningsutveckling, produktivitet och löneandel. Det hade också gett en klarare bild av de förhållanden som de aktuella fackförbunden hade att ta ställning till. I samtliga tre branscher avvek produktivitetsutvecklingen från det generella mönstret, framför allt från senare delen av sextiotalet. I vilket utsträckning påverkade detta fackförbundens agerande? En rimlig hypotes är att det – mer än på andra förbundsområden – bidrog till att sysselsättningsfrågorna kom i förgrunden framför löne- och branschrationaliseringsfrågor. Den solidariska lönepolitiken skapade särskilda svårigheter för dessa branscher, om än från olika utgångspunkter. En bredare belysning av branschernas ekonomisk utveckling hade också gett perspektiv på den samtida diskussionen om branschproblem. Byggbranschen brottades med akuta vinstproblem under slutet av sextiotalet och tidigt sjuttiotal: hur påverkade detta utformningen av Bygg 70-programmet? Var det sant att produktivitetsökningen var svag och vinstandelen låg inom handeln under femtiotalet, något som ofta framhålls i debatten? Ökade antalet förvärvsarbetande i handelsbranschen under hela perioden? Många av de samtida debattörernas föreställningar om branschvillkoren kan ifrågasättas.

I detta sammanhang kan det tilläggas att analysen vunnit på ett mer genomarbetat genusperspektiv. Byggnadsarbetarförbundet var ett utpräglat manligt dominerat förbund, medan både textilarbetarförbundet och handelsanställdas förbund dominerades av kvinnliga medlemmar. Kan detta ha påverkat förutsättningen för de fackliga arbetet – både vad gäller prioriterade frågor och relationerna till ett påtagligt manligt dominerat LO? En hypotes är till exempel att rörlighetsrestriktionerna var

4 Jörgen Ullenhag, Den solidariska lönepolitiken i Sverige, Uppsala 1971.

skarpare på kvinnodominerade områden, något som indirekt kan ha bidragit till att kvinnodominerade förbund gav relativt sett större utrymme för frågor om anställningstrygghet, socialpolitiska insatser och regionalpolitiska stöd.

Slutsatser

Går det att beskriva de tre fackförbunden i termer av särintressen som drevs av en individuell rationalitet som stod i strid med bredare samhällsintressen och kollektiv rationalitet? Jag tror knappast det. De fallstudier som presenteras i Carina Gråbackes avhandling kan lika gärna tolkas i motsatt riktning. Samtliga tre fackförbund anslöt sig från slutet av femtiotalet till den solidariska lönepolitiken och accepterade omfattande omstruktureringar inom respektive bransch. Från slutet av sextiotalet skedde en omsvängning. Parallellt med önskemål om låglönesatsningar tilltog kraven på samhällsingripanden och branschskydd. Men omsvängningen från slutet av sextiotalet skulle lika gärna kunna förklaras i relation till förändrade strukturella förhållanden som med teorier om fackliga organisationers»skråmässighet«.DenRehn-Meidnerska medicinen slutade att verka när arbetskraftsefterfrågan i näringslivet inte längre var tillräckligt stark. Det handlade också om en ideologisk radikalisering som inte självklart behöver tolkas i termer av skråmässiga särintressen. Uppmärksamhet ökade kring strukturrationaliseringens och tillväxtens kostnadssida och tog sig uttryck i tilltagande motstånd mot centralisering och storstadskoncentration. Nya krav på inflytande och utvecklingsmöjligheter i arbetslivet lanserades. Anställningstrygghet och en bättre arbetsmiljö framhölls framför ett ständigt ökat konsumtionsutrymme.

I praktiken var fackförbundens inflytande över bransch- och politikutvecklingen marginell. Varken textilarbetarförbundet eller byggnadsarbetarförbundet kan sägas ha uppnått några påtagliga resultat. Handelsanställdas förbund fick inte heller gehör för krav på ökat branschskydd. Textilindustrin fortsatte att bantas i snabb takt, trots tullskydd och industristöd på sjuttiotalet, samtidigt som byggandet reducerades. Konsekvensen av den solidariska lönepolitiken blev hårdhänta rationaliseringar. Överhuvud taget skedde anpassningen av krisbranscherna snabbt under sjuttiotalet. Omstruktureringarna saknar internationellt motstycke vad gäller hastighet och omfattning.[5] Överskottskapaciteten reducerades snabbt i jämförelse med andra länder i Västeuropa. En förklaring till detta kan vara det starka utrikeshandelsberoendet som hela tiden haft ett disciplinerande inflytande på politik och arbetsmarknadsorganisationer. Krispolitiken under sjuttiotalet förhindrade inte omställningen till nya villkor.

Avslutningsvis har Carina Gråbacke skrivit en bra avhandling som behandlar många centrala frågor kring marknadsvillkor, partsorganisationers beteende och politiskt dikterade regleringssträvanden. Visserligen bekräftas inte tesen om fackförbunden som smala »särintressen«, men den teoretiska svagheten förmörkar ändå inte bilden av en i grunden stark och uppslagsrik avhandling. Det går att dra två centrala

5 Lennart Schön, Omvandling och obalans. Bilaga 3 till LU 94, Stockholm 1994.

slutsatser:

1. Fackförbundens ställningstagande rymmer både det »skråmässiga« branschperspektivet och det »breda« samhällsperspektivet. Vilken tendens som dominerar beror mycket på den ekonomiska utvecklingen och de sociala villkoren – på den egna delarbetsmarknaden och i samhället i stort. Avhandlingen illustrerar att låg tillväxt ökar intressemotsättningar och ger fördelningskampen karaktären av ett nollsummespel.
2. Det finns alltid en spänning i relationen mellan förbundet och huvudorganisationen. Huvudorganisationens möjligheter att skapa konsensus bland medlemsförbunden kring en tillväxtinriktad politik beror inte bara på täckningsgraden på arbetsmarknaden utan också på tillväxt, arbetskraftsefterfrågan och välfärdspolitiska insatser (arbetsmarknads- och socialpolitik).

Jonas Olofsson

NYA BÖCKER

Perspektivpåsocialaproblem
Anna Meeuwisse & Hans Swärd (red), Natur och Kultur, 2002

Anna Meeuwisse & Hans Swärd (red)
Perspektiv på sociala problem
Natur och Kultur
2002

Låt mig först som sist säga att Perspektiv på sociala problem är en angelägen bok som fyller en lucka i den skandinaviska facklitteraturen. Trots att sociala problem är ett centralt begrepp i en rad utbildningar i människobehandlande yrken har det tidigare saknats en bok som utifrån ett brett spektrum behandlar olika teorier om sociala problem. Nu finns det en.

Ett konfliktfyllt ämne

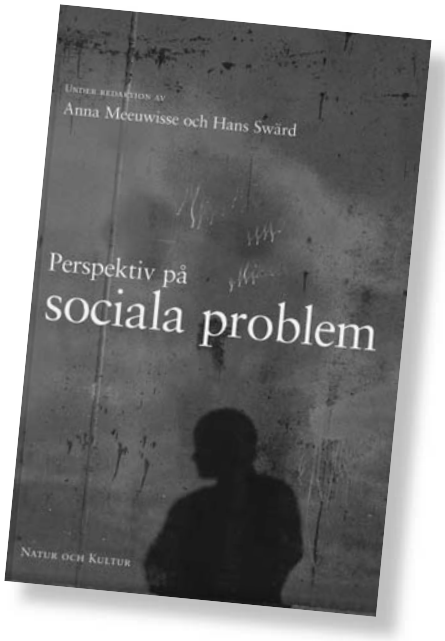
Vad menar vi egentligen när vi talar om ett socialt problem och hur förstår och förklarar vi sådana problem? Studier av sociala problem är ett konfliktfyllt och svårt ämne. Det finns en rad motsättningar och kontroverser, inte bara inom olika vetenskapliga discipliner utan också mellan olika vetenskapsgrenar och mellan vetenskap, politik och sunt förnuft. Hur skall vi t. ex. förklara missbruk bland ungdomar? Är det miljonprogrammets betongförorter med segregation och alienation som skapar problemen? Beror missbruket snarare på familjens sönderfall och frånvarande föräldrar? Skall vi söka orsakerna i gängmentalitet eller i populärkulturen? Är det en besvärlig uppväxt som är grundorsaken? Eller kan orsaken till missbruket sökas i neuropsykologiska och biologiska faktorer?

Det finns en uppsjö teorier som betonar antingen biologiska, psykologiska, socialpsykologiska eller sociologiska orsaksfaktorer. En del teorier kompletterar varandra, andra är motstridiga. Det är en djärv satsning av redaktörerna och författarna att ta sig an denna djungel av teorier och försöka att bringa ordning i dem – att sammanföra allt från biologiska och psykologiska till hårdkokta strukturella teorier i en och samma bok.

En flervetenskaplig antologi

Boken är en antologi där 17 olika författare bidrar medsammanlagt15kapitel. Författarna representerar olika vetenskapliga discipliner och lärosäten i Norden och är alla väletablerade ”problemforskare”. Härutöver deltar den legendariske amerikanske mikrosoci-

ologen Thomas J Scheff som tillsammans med professor Bengt Starrin skrivit ett kapitel om sociala band och skam. I boken får vi veta vad ämnen som sociologi, socialpsykologi, psykologi och socialt arbete kan bidra med i studiet av sociala problem. Även flervetenskapliga perspektiv som etnicitet och genus belyses.



Innehållet

Boken består av tre delar. Den första delen behandlar frågan om vad ett socialt problem är utifrån olika definitioner, historiska utvecklingslinjer och en nordisk kontext. Sune Sunesson redogör för hur man sett på sociala problem i historien – från 1500-talstraditionen till våra dagar. De danska forskarna Morten Ejrnæs och Søren Kristiansen jämför skandinaviska och amerikanska perspektiv på sociala problem. De menar att amerikanska resonemang allt för lätt överförts till en nordisk kontext utan att ta hänsyn till att det handlar om helt olika samhällssystem.

I del II presenteras olika teorier om sociala problem. Sociologen Ingrid Sahlin behandlar sociala problem som verklighetskonstruktioner och diskuterar frågan om konstruktivism och objektivism. Därefter behandlas strukturella förklaringar till sociala problem (Lennart Nygren), interaktionistiska perspektiv (Sven-Axel Månsson), känslors och emotioners betydelse för sociala problem (Thomas Scheff och Bengt Starrin), utvecklingsekologiska perspektiv på sociala problem (Gunvor Andersson), psykodynamiska perspektiv (Gunilla Lindén) och slutligen psykologiska, neuropsykologiska och biologiska faktorers betydelse för utvecklingen av sociala anpassningsproblem (Håkan Stattin och Henrik Andershed).

Del III bär titeln ”Fokus på makt, kön, etnicitet och ojämlikhet”. Margaretha Järvinen från Danmark anläg-