Meta-analytisk utvärdering av programresultat

Mark W. Lipsey

Vanderbilt University

Utvärdering ger en bedömning av hur ett speciellt socialt program fungerar i den miljö det skall verka och i förhållande till ställda förväntningar. Vid utformningen av ett nytt program, eller vid förbättring av ett redan existerande, är det emellertid utvärderingsforskningen som ansvarar för att ta fram bevis för vilka program som visat sig mest effektiva i tidigare utvärderingsstudier. För att detta skall kunna uppnås måste utvärderarna dra lärdom från tidigare studier om vilka interventioner som fungerar bäst, för vilka ändamål och under vilka förhållanden. Detta kräver i sin tur att åtminstone några forskare inom utvärderingsområdet, på bred front, systematiskt samlar och integrerar utvärderingsresultat avseende program och programvariationer. Som en extra förmån ger ett sådant bemödande också möjlighet att undersöka de metoder som utvärderarna använder sig av och hur de relaterar till resultaten som genereras av metoderna, så att utvärderingsforskarna därigenom kan lära sig förbättra sin metodologi.

Den centrala fråga som ställs här gäller *generalisering* , d.v.s. hur man går vidare från detaljer i ett individuellt program till en bredare förståelse av differentierad effektivitet hos olika program för olika sociala problem (Cook, 2000). Välgrundad generalisering är det verktyg vi använder för att få fram bevisbaserade principer om vad som karakteriserar mer eller mindre effektiva program. En väl utvecklad uppsättning av sådana principer i ett givet programområde är ett nödvändigt verktyg för att utforma, förbättra och förstå effektiva interventioner.

Till skillnad från mer akademiska socialvetenskapliga områden, där granskning av

forskningslitteratur och andra former av kunskapsframställning är vanligt förekommande, har relativt liten uppmärksamhet ägnats åt systematiska sammanställningar inom utvärderingsområdet. Detta beror huvudsakligen på utvärderingsforskningens egen karaktär och inte på att sådana sammanställningar skulle vara oanvändbara. Till sin karaktär tenderar utvärderingar att fokusera på det program som skall granskas, och utvecklas därmed så att de blir skräddarsydda för att passa detaljerna i det programmet, de förväntningar som gjorts och specifika avsikter med utvärderingen. När resultaten analyseras och rapporteras ägnas liten eller ingen ansträngning åt generaliserbarheten av resultaten, hur de kan tillämpas på andra programsituationer, vad som kommit fram och som kan vara av intresse för de som ännu inte startat ett program av motsvarande typ o.s.v. Som tillämpad forskning är utvärdering organiserad kring applicering på ett visst programsammanhang. När ett sådant projekt avslutas flyttar utvärderarna vidare till nästa utan större tanke på att utvinna, rapportera och vidareförmedla de lärdomar som dragits av projektet.

Ett speciellt intressant och viktigt område, där utvärderingsfältet skulle kunna dra stora fördelar av systematisk sammanställning av vad som framkommit i tidigare utvärderingsstudier, är det som gäller resultatutvärdering. Av största vikt för de flesta program, både politiskt och praktiskt, är den avsedda förbättrande effekten på det målproblem man vänder sig till. För programplanering och förbättring är det emellertid lika viktigt att veta vilken sorts program som har betydelsefulla effekter på sådana problem och vilka bland dem som är de mest effektiva. Mer bestämt, vi kanske vill veta vilka programkarakteristika, målpopulationer och utvärderingsmetoder som är förknippade med slutsatser från större och mindre programeffekter på betydande utfall variabler.

Vad är meta-analys?

Utfallsutvärdering utförs vanligtvis med experimentell eller quasi-experimentell forskningsdesign med kvantitativa utfallsmätningar och resultat som rapporteras i statistiska termer. I denna typ av forskning är den teknik som är känd som meta-analys speciellt lämpad för systematiska sammanställningar av resultat från olika studier (Cooper, 1998; Cooper & Hedges, 1994; Lipsey & Wilson, 2001). Meta-analys använder sig av ett statistiskt mått kallat effektstorlek (*effect size)* som representerar vad som kommit fram om den uppskattade programeffekten på en utfallsvariabel, till exempel från en jämförelse mellan utfallet för ett urval av programdeltagare och de i en kontrollgrupp som inte varit föremål för åtgärd. Det vanligast förekommande måttet på effektstorlek är standardiserad skillnad i medelvärde, definierat som skillnaden mellan medelvärdet på en utfallsvariabel för den behandlade gruppen och den för kontrollgruppen, delat med den kombinerade standardavvikelsen. Delning genom standardavvikelsen standardiserar effektstorleken på så sätt att det statistiska måttet i originalmätningen presenteras av effektstorleken i standardavvikelseenheter. En effektstorlek av .50, till exempel, indikerar att utfallet för programgruppen vid en speciell mätning var hälften av en standardavvikelse bättre än för kontrollgruppen, oberoende av den mätskala som användes. Antag att en utvärderingsstudie mäter utfallet av depression i Beck Depression Inventory och finner att genomsnittet för den behandlade gruppen är .40 standardavvikelser lägre (bättre) än för kontrollgruppen. En annan studie av liknande behandling kan visa utfallet av depression på Hamilton Depression Scale och komma fram till en skillnad likvärdig med .25 standardavvikelse mellan behandlings- och kontrollgruppsmedelvärdet. Vi kan då jämföra dessa och notera att den första studien visade en större behandlingseffekt på depression. Likaså kan vi, om vi så önskar, kombinera dessa effektstorlekar med liknande utfall på depression från flera andra utvärderingar av behandlingseffekter. Vi får då en

uppsättning data som kan ge oss tillgång till fördelningen av utfall, deras medelvärde, vilka typer av interventioner som ger de största effekterna på depression och vilka som ger de minsta, o.s.v. När vi nått dit kan vi säga att vi utför en meta-analys.

Andra typer av effektstorlek används också i meta-analys för att visa utfallet av olika studier i ett vanligt decimalsystem. När utfallet variabler är binärt, d.v.s. består av två som till exempel sjuk eller frisk, död eller levande, icke hemlös eller hemlös, o.s.v., är oddskvoten ett användbart mått, alltså sannolikheten för att någon i programgruppen får ett gynnsamt utfall delat med sannolikheten för att någon i kontrollgruppen skulle få det utfallet (Haddock, Rindskopf, & Shadish, 1998). Således betyder en oddskvot på 1.5 att sannolikheten för ett gott utfall i behandlingsgruppen är en och en halv gång större än vad som är fallet för kontrollgruppen. Oddskvoter används i stor omfattning som effektstorleksmått för att visa utfallet av biomedicinska interventioner och förekommer ofta i utvärderingar inom medicinsk behandling.

En systematisk sammanfattning av utvärderingsresultat med användning av meta-analytisk teknik innebär framräkning av en effektstorlek för varje utfallsvariabel av intresse för de utvärderingar som involverar samma eller liknande interventioner. Dessa effektstorlekar kan bäst beskrivas som observerade effekter av interventioner, effekter som observerats med användning av mätinstrument och metoder som tillämpas i utvärderingsforskning. Annan information beträffande beskaffenhet och förhållanden gällande interventionen, personkarakteristika för de personer som är föremål för interventionen, studiemetod och tillvägagångssätt, och liknande kodas också vanligtvis för en meta-analys. All denna information för de studier som inkluderas i en meta-analys organiseras sedan i en databas som tillåter statistiska analyser av fördelningen av de observerade effekterna i dessa studier.

Typiska statistiska analyser av en meta-analytisk databas sorterar först

effektstorlekar enligt den typ av utfallsvariabler de representerar. Om det till exempel är så att de utvärderingsstudier som inkluderades i meta-analysen bedömde effekterna av familjeterapi på sådant utfall som hur pass nöjd man är med sitt äktenskap, kvaliteten på kommunikationen och barnets/barnens uppförande, skulle effektstorlekarna för varje utfallsvariabelkategori bli separat analyserade. Effektstorlekens medelvärde över alla studierna skulle då kalkyleras för varje utfallsvariabel; sedan skulle variationer av effektstorlekarna runt det medelvärdet bedömas. Om skillnaden av effektstorlekarna inte var större än väntat i förhållande till de samplingsfel som är förenade med urvalet av studerade personer, skulle effektstorlekens medelvärde ge en bra summering av interventionseffekten. Eftersom detta medelvärde gäller för studier som är inkluderade i meta-analyser ger det en mer representativ uppskattning av effekten av en speciell typ av intervention jämfört med en uppskattning som härrör från vilken som helst annan utfallsstudie.

Ofta visar emellertid effektstorlekarna från olika studier mer variation än vad som sannolikt hade resulterat från samplingsfel. I den situationen är det meta-analysens uppgift att fastställa om det föreligger systematiska samband mellan de olika studiernas karakteristika och den effektstorlek de producerar. De observerade effekterna av en samling interventionsstudier kan generellt ses som en funktion av behandlingens verkan, egenskaper hos dem som är föremål för åtgärder, de metoder som används för att studera effekterna samt till viss del av statistisk störning. Ett användbart sätt att summera den information som genereras av meta-analys är att ange proportionen av variation i de observerade effekter som är förknippade med var och en av dessa olika aspekter på utvärderingssituationen. Ytterligare undersökning kan göras av interventionens specifika karakteristika, behandlingsmottagare, och de metoder som är närmast förknippade med större och mindre observerade effekter.

Resultaten från denna process tillhandahåller bevis som ger oss stöd för användbara generaliseringar om vilka behandlingar som är mest effektiva, på vilka utfall och för vilken typ av mottagare.

Lärdomar från meta-analys

Meta-analys används flitigt i utvärderingsresultat alltsedan det pionjärarbete som utfördes av Smith och Glass (1977). Även om den inte är fullt utvecklad till alla delar, har den redan genererat viktig lärdom om sociala program och de metoder som utvärderare använder för att studera dem. För att illustrera beskaffenhet och resultat av meta-analyser, och dess potential för ytterligare förbättringar inom området för programevaluering, skall vi beskriva sex lärdomar som vi dragit från meta-analyser. Vi kommer emellertid inte att här försöka granska relevant litteratur inom området. Istället skall vi helt enkelt sammanfatta vad vi ser som signifikanta slutsatser som kan dras genom att använda exempel som finns till hands från vårt eget arbete under det sista årtiondet.

1. Många sociala program är mer effektiva än vad vi hade tänkt oss

Ett bekymmer med utfallsutvärdering är att utvärderaren ofta inte hittar någon signifikant effekt som producerats av det aktuella sociala programmet. Det är inte ovanligt att det som framkommer från utfallsutvärderingen är så svagt att vi inte kan vara säkra på att programmet har haft någon meningsfylld påverkan. Det som Rossi och Wright (1984) en gång kallade "the parade of null results in evaluation" kan leda till den pessimistiska slutsatsen att "ingenting fungerar" i de sociala programmens värld. Den vanliga grunden för sådana slutsatser är en samling utfallsutvärderingar som använder experimentell och quasi-experimentell design som visar relativt få statistiskt signifikanta positiva effekter på relevanta utfallsvariabler.

En distinkt karakteristik hos meta-analyser är att de fokuserar på magnituden av

de effekter som observeras i varje studie, inte deras statistiska signifikans. Därutöver kan man genom att kombinera dessa uppskattningar av storlek från ett flertal utfallsstudier, avslöja den faktiska fördelningen av effektstorlekar som karakteriserar en viss typ av intervention. När detta gjorts, blir det ofta uppenbart att många av de programeffekter som observerats i de ursprungliga utvärderingsstudierna, är större och mer genomgående positiva än de verkade vara när endast de som hade statistisk signifikans räknades. Anledningen till detta är, kort sagt, att statistisk signifikans är påverkad både av magnituden av en interventionseffekt och av samplingsstorleken på vilken den mäts (Cohen, 1988). Därför är det så, att effekter som är tillräckligt stora för att vara av praktisk signifikans kan, vilket ofta är fallet i utvärderingar, tappa statistisk signifikans i en individuell utvärderingsstudie därför att forskningen utförs på små urval och med motsvarande låg statistisk kraft.

Det är relativt lätt att visa den annorlunda och mer positiva bild av programeffekter som kommer fram genom meta-analyser. Lipsey och Wilson (1993), till exempel, samlade alla meta-analyser av effekter av psykologiska, utbildningsmässiga, och beteendemässiga interventioner som gick att få fram vid den tidpunkten vilket blev fler än 300. Många av dessa hade utförts i programområden präglade av en historia av kontroverser över huruvida interventionerna hade åstadkommit några positiva effekter. Vid undersökningen framkom emellertid att fördelningen av medeleffektstorlekar över dessa omfattande interventioner, och de hundratals studier och tusentals deltagare som inkluderats i de studier som meta-analyserats, avslöjade att den stora majoriteten av resultatet var positiva och hade en icke obetydlig magnitud.

Fig 1 visar den summerade fördelningen av medeleffektstorlekar från alla dessa meta-analyser. Den övervägande majoriteten av meta-analyserna visade positiva effekter på

utfallet (medeleffektstorlekar större än noll) och genomsnittet för dessa medelvärden var omkring .50. Det innebär att på ett genomsnitt av alla representerade interventioner blev utfallet för de individer som omfattats av programmen ungefär en halv standardavvikelse bättre jämfört med kontrollgruppen, oavsett vilken skala som användes vid mätningen. För att få perspektiv på detta kan vi anta att 50 % av individerna i kontrollgruppen själva skulle erhålla acceptabla utfall. En effektstorlek på .50 betyder att jämförelsevis skulle 70 % av alla i programgruppen erhålla acceptabla utfall. Inom många programområden är även mindre effekter än denna av stor praktisk betydelse.

Dessa meta-analytiska resultat betyder naturligtvis inte att alla sociala interventioner har positiva effekter. Icke desto mindre indikerar de att för att uppnå någon generalisering om programeffektivitet måste vi analysera den verkliga kvantitativa effektstorleken som genererats av de utfallsstudier som finns tillgängliga. Den uppenbara visheten i detta tillvägagångssätt som materialiseras i meta-analytisk teknik, visar att utfallsstudier ofta ger ett bredare och mer positivt utfall än vad som annars är uppenbart.

2. Individuella utfallsstudier kan lätt producera felaktiga resultat

Den situation som beskrivs ovan, där många utvärderingsresultat visar positiva effekter, och ibland relativt stora sådana, men som ändå brister när det gäller konventionella nivåer på statistisk signifikans, har klargörande implikationer för utformningen av designen för individuella utvärderingsstudier. Genom att undersöka effektstorlekar i ett antal utvärderingar, och därigenom i grunden kombinera alla deras individuella urval, kan meta-analyser fokusera direkt på fördelningen av observerade effektstorlekar utan större hänsyn till huruvida var och en är statistiskt signifikant. Vad vi kan se vid detta tillvägagångssätt är emellertid att många av de

individuella utvärderingsstudierna inte visar statistiskt signifikanta effekter, även när meta-analyser visar att faktisk magnitud av effekterna av den aktuella interventionen generellt är positiva. Med andra ord, uppskattningarna av effektstorlekarna för viktiga utfall i individuella studier producerar positiva resultat men otillräcklig statistisk signifikans och kan därför inte med visshet identifieras som fördelaktig programpåverkan inom en individuell utvärderingsstudies kontext.

Som tidigare noterats, kan det lätt hända att studiens urvalsstorlek är för liten för att tillhandahålla tillräcklig statistisk kraft för att uppnå statistisk signifikans, även när den uppskattade effekten är av meningsfull storlek. Meta-analyser har avslöjat att otillräcklig statistisk kraft är ganska vanligt i utvärderingsforskning (Lipsey, 2000). En underdimensionerad utvärderingsdesign som appliceras på ett effektivt program kommer vanligtvis att ge resultat av otillräcklig statistisk signifikans som därmed åstadkommer det som kallas Typ II fel, d.v.s. misslyckande att avvisa noll-hypotesen (av ingen effekt) när det i själva verket är felaktigt. Från ett vetenskapligt perspektiv är det så att effekter som har otillräcklig statistisk signifikans i en individuell studie, oavsett av vilken orsak, åtnjuter liten trovärdighet. De har, per definition, en oacceptabelt hög sannolikhet för att vara falska, dvs de representerar statistiska fel istället för faktiska interventionseffekter.

Avsaknaden av statistisk signifikans i en underdimensionerad utfallsstudie innebär tekniskt sett endast att forskningen har misslyckats att avvisa noll-hypotesen av ingen effekt, inte att den har bekräftat avsaknaden av effekter. Detta är emellertid en subtil distinktion som lätt går förlorad när det gäller beslutsfattare, programintressenter och många forskare. Statistiskt icke signifikanta resultat tolkas i stor omfattning som indikationer på att programmet inte är effektivt, med de implikationer det politiskt och praktiskt har. Avseende detta anklagas programmet för att ha varit ineffektivt när det

istället är utvärderingsforskningen som har misslyckats med att använda en design med tillräcklig statistisk kraft för att finna meningsfulla effekter när sådana står att finna.

Relationen mellan observerade effektstorlekar, som framräknats i meta-analyser, och den statistiska signifikans som framkommit i de individuella utvärderingsstudierna är illustrerad i fig 2. Den figuren visar fördelningen av effektstorlekarna på alla utfallsvariabler som rapporterats i över 500 utvärderingsstudier av interventionsprogram för ungdomsbrottslingar. För att underlätta tolkningen framställs effektstorlekarna i termer av procentuell förbättring av behandlingsgruppen i relation till kontrollgruppens median. +30 betyder att oavsett utfallsvariabel som uppmätts, visade de behandlade ungdomarna en 30 %-ig förbättring jämfört med kontrollgruppen. Som synes är över hälften av de observerade effektstorlekarna positiva (större än noll) och många är relativt stora (de representerar t ex 20 %-ig eller högre förbättring med behandling). Som helhet råder det lite tvivel om att interventionerna som utvärderats i dessa studier hade positiva effekter på en majoritet av de utfall som mättes (notera dock att 17% av utfallen var noll och omkring en fjärdedel var negativa, vilket betyder att kontrollgrupperna klarade sig bättre).

I figur 2 visas proportionen av effekter som befunnits vara statistiskt signifikanta i de ursprungliga utvärderingsstudierna. Eftersom urvalsstorlekarna i dessa utvärderingsstudier tenderade att vara blygsamma (en median på omkring 60 av varje i interventions- och kontrollgrupper), besitter de inte någon större statistisk kraft. Figur 2 visar att majoriteten av positiva effekter inte funnits vara statistiskt signifikanta i de individuella studierna förrän de befunnit sig i en verkningskrets där behandlade ungdomar uppvisade förbättringar på 40 % eller mer jämfört med de i kontrollgrupperna. I praktiska termer kan sägas att naturligtvis uppstår meningsfulla

effekter under denna nivå. Många program skulle känna sig nöjda med en förbättring på 10-20% bland de aktuella ungdomarna. Därutöver är de många positiva effekterna inom området ganska uppenbara i meta-analyserna. Men, som också kan ses, de individuella utvärderingsstudierna har en minskande sannolikhet för att upptäcka programeffekterna på en statistiskt signifikant nivå allteftersom de blir mindre.

Det är intressant att notera att ett liknande mönster framkommer på den negativa sidan. Effekter för behandlade ungdomar måste vara 50 % värre än för kontrollungdomarna, eller mer, innan majoriteten var statistiskt signifikant. De krympande proportionerna av statistiskt signifikanta resultat i den negativa riktningen jämfört med den positiva riktningen, som framkommer i figur 2, representerar också ett problem med små urvalsstorlekar. Urval i vilka man funnit negativa effekter har visat sig vara speciellt små och det är därför möjligt att även bland dem som befunnits signifikanta, representerar många inget annat än ett samplingsfel.

De praktiska begränsningar som är förenade med resultatredovisningar i fältmiljöer är av sådan art att det ofta är ganska svårt att göra tillräckligt stora urval för att garantera en hög grad av statistisk kraft. Den väsentliga roll som "statistical noise" spelar i sådan forskning, som har demonstrerats av meta-analyser, gör att resultatredovisningar av individuella program lätt kan misslyckas med att uppnå statistisk signifikans för något som ändå är meningsfulla programeffekter. Följaktligen kan resultatet av sådana utvärderingar i sig vara missledande. Ett viktigt bidrag som meta-analyser kan ge i denna situation är att tillhandahålla jämförbara resultat från andra liknande interventioner inom vilka tolkning av de eventuella oklara utfallen av individuella utfallsstudier kan göras. Till exempel: effektstorlekar från en utfallstudie kunde jämföras med fördelningen av effekter funna i relevant meta-analys. Deras magnitud i relation till vad som framkommit i liknande program kunde sedan

bedömas som ett supplement till den statistiska signifikanstesten.

3. Metoder spelar roll!

Idealiskt skulle de experimentella och quasi-experimentella forskningsmodeller och tillvägagångssätt, som vanligtvis används i utfallstudier generera uppskattningar av faktiska programeffekter som var relativt oberoende av själva metoderna. Vi förväntar oss till exempel att slumpmässigt tilldelade försök skall producera opartiska uppskattningar av interventionseffekter, men det är inte alltid möjligt att använda sådana modeller i utfallstudier. Det skulle vara betryggande att veta att mer hanterliga icke-slumpmässiga modeller kunde tillhandahålla resultat av rimlig överensstämmelse med de från en slumpmässig design. Likaledes skulle det i fall där det finns flera rimliga sätt att mäta utfallet av variabler, vara önskvärt att dessa skulle ge jämförbara resultat när det tillämpas på samma intervention.

En av fördelarna med meta-analyser är att de kan undersöka i vilken grad variation i metoder och tillvägagångssätt är relaterad till de effekter som framkommer i studierna. Ett enkelt tillvägagångssätt är att bedöma proportionerliga varianser i observerade effekter som har samband med den metodologiska karakteristiken av studien i motsats till den som har samband med sådana faktiska aspekter av programmet som deltagarnas karakteristik, typ av intervention, och mängden behandling. Om huvuddelen av varianserna i effektstorlekarna har samband med olikheter i programrelaterad karakteristik, är det en god indikation på att de observerade utfallen huvudsakligen avslöjar information om faktiska programeffekter. Om det däremot istället är så att en stor del av varianserna i effektstorlekarna har samband med metodologiska skillnader mellan studierna, säger det oss att utfallet från dessa studier kan vara kraftigt påverkat av det sätt på vilket programmet studerats, istället för det utfall som faktiskt producerats.

När vi har analyserat varianser i effektstorlek på detta sätt med stora meta-analytiska datauppsättningar, har vi med bestörtning funnit att ungefär lika många varianser i effektstorlek har samband med metodologiska skillnader som med programkarakteristika. Av de 300 meta-analyser av psykologiska, utbildnings- och beteendemässiga interventioner som tidigare nämnts, har vi till exempel funnit det mönster av samband som visas i figur 3 (från Lipsey & Wilson, under publicering). Vi har redan kommenterat samplings-felens betydelse, som är speciellt utmärkande för små samplingsmängder. Vid jämförelse mellan programrelaterad och metodrelaterad påverkan på effektstorlekarna, visar emellertid figur 3 att variansen i effektstorlekarna som har samband med de metoder som använts av utvärderarna är större än den som har samband med interventionskarakteristika (21% vs 25%).

När olika kategorier metodologiska karakteristika tas ut väntar ytterligare överraskningar. I forskningsmodeller, som huvudsakligen representerar slumpmässig versus icke slumpmässig tilldelning av interventionsförhållanden, och som är närbesläktade, är, som väntat, typen av kontrollgrupp (exempelvis "ingen behandling" versus placebo) betydelsefull. Det finns en omfattande metodologisk litteratur om den potentiella snedhet som är förknippad med designfaktorerna. (t ex Shadish, Cook, & Campbell, 2002). Aspekterna på utfallsmått, som emellertid har erhållit betydligt mindre uppmärksamhet i litteraturen om utvärderingsmetoder, verkar också ha påtaglig inverkan på de observerade effektstorlekarna. Ett kännetecken för mätningar representerade i denna kategori inkluderar det sätt på vilket utfallskonstruktionen är operationaliserad (t ex självrapporterade mätningar, standardiserade tester, officiella register) och val av tidpunkt för mätningen (t ex omedelbart efter interventionen eller med eftersläpning någon tid senare).

Ytterligare utforskning av utvärderingsstudier med meta-analytiska tekniker skulle hjälpa till att fastställa vilka metoder och tillvägagångssätt som gav mest giltigt resultat och vilka som skapade så mycket förvanskning att de är olämpliga för användning i resultatutvärderingar. Vad meta-analys redan har demonstrerat är att neutraliteten hos mer typiska metoder för resultatutvärderingar inte kan tas för given. Det vi observerar som programeffekter kan reflektera lika mycket påverkan från de metoder, med vars hjälp programmet har studerats, som de faktiska effekter programmet har haft på sina tilltänkta mottagare.

4. Programeffektivitet är en funktion av identifierbara programkarakteristika

Varje socialt program är i vissa avseenden unikt och bedömningen av dess påverkan måste skräddarsys till dess speciella karakteristika och situation. Icke desto mindre finns det likheter mellan program i ett givet interventionsområde som tillåter generaliseringar över program. Det är viktigt för utvärderaren att veta vilka karakteristika i en intervention som tenderar att associeras med de mest positiva utfallen. Sådan information gör det lättare att utforma en effektiv utvärdering genom att belysa de aspekter på programmet som utvärderingen skall fokuseras på. Med tanke på programutformning och förbättringar är dessutom identifikation av karakteristika hos effektiva program en hjälp för att definiera de "bästa tillämpningarna" i ett speciellt interventionsområde som skall försöka efterliknas.

Meta-analys gör det möjligt att närmare analysera karakteristika hos interventionsprogram som urskiljer dem som producerar stora utfallseffekter från dem som producerar små. På grund av relationerna mellan metoderna som används i utvärderingsstudier och de ovan beskrivna observerade utfallen, kan det emellertid vara missvisande att enbart jämföra effektstorlekarna för program med olika karakteristika. En potentiellt klarare bild ges genom att använda meta-analytiska

tekniker för att statistiskt kontrollera metodologiska skillnader mellan studier så att programkarakteristika som närmast är associerad med stora och små effekter kan lösgöras från metodologiska artefakter.

Med sådana statistiska kontroller kan analyser liknande dem som visas i figur 4 genomföras för studier av våra meta-analyser avseende brottslighet. Detaljer från dessa analyser beskrivs på annan plats (Lipsey, 1992a,b, 1995), men resultaten visar att det finns samband mellan typ av program, hur väl programåtgärder genomförts och implementerats och utfallet. Figur 4 visar till exempel att olika grupper av interventionsprogram mot brottslighet har helt olika genomsnittliga effekter på återfall i brott hos ungdomar.

Det är emellertid inte så enkelt att de största effekterna följer genom användning av en av de mer effektiva programmodellerna. Figur 4 visar också att behandlingsintegriteten har minst lika stor påverkan på utfallet. Behandlingsimplementation omfattar i dessa analyser summan av behandling som givits och omfattningen av programmets ansträngningar att kontrollera eventuella försämringar eller ofullständigheter i behandlingen. Även programmen i den generellt mest effektiva gruppen saknar effekt i någon större utsträckning i de fall de inte är väl implementerade. Omvänt är det så att program av en generellt mindre effektiv typ ändå kan ha relativt stora effekter genom en väl implementerad intervention. Programeffektivitet beror på speciella kombinationer av programmets utformning och särdrag som måste konfigureras optimalt för att uppnå bästa möjliga utfall. Därutill kommer att de avgörande kännetecknen för programmet inte nödvändigtvis är unika för något speciellt program men visar generella mönster över program

Generaliseringar gällande karakteristika hos de mest effektiva programmen, och hur dessa bäst kombineras, kan inte definieras i utvärderingen av ett enda program.

De uppenbaras endast när mönster över flera program kan undersökas. Upptäckten av sådana samband är därför ett utmärkande och viktigt bidrag av meta-analyser till utvärderingsforskningen.

5. Det finns stort utrymme för programförbättringar

De utfallstudier som generellt finns tillgängliga för meta-analyser i något programområde inkluderar vanligtvis en mix av pågående "real world" program för vilka en utvärdering har genomförts och olika slags demonstrationsprogram eller forskningsorienterade tester av programkoncept. En av de användbara jämförelser som kan göras i meta-analyser är att kontrastera storleken på effekterna för de bäst utformade och implementerade programmen med de som är av en mer vardaglig typ. Demonstrationsprogram som är utformade och implementerade av forskare för att testa senaste och bästa interventionskoncept, förväntas producera bättre utfall än rutinartade praktiska program. De använder inte endast mer effektiva interventionsangrepp utan de har generellt sett också större kontroll över sammanhanget i de insatser som görs och det klientel det gäller.

I detta avseende undersöker demonstrationsprogrammen den övre gräns för prograeffektivitet som kan uppnås med tillgängliga interventionstekniker, och visar därmed vad praktiska program kan sträva efter under optimala förhållanden. Ett stort gap mellan effekterna av praktiska program och de som uppvisas av demonstrationsprogram i ett interventionsområde, tyder på att de praktiska programmen kan förbättra sin effektivitet genom att utformas enligt huvuddragen i demonstrationsprogrammen. Olyckligtvis har meta-analytiska undersökningar av effektiviteten hos demonstrationsprogram i jämförelse med vardagliga praktiska program ännu så länge endast utförts i begränsad omfattning. Tidiga indikationer visar emellertid ganska ansenliga gap till förmån för demonstrationsprogrammen (t ex

Weisz, Weiss, & Donenberg, 1992, on childrens mental health programs).

Problematiken kan illustreras med data från de meta-analyser från programmen gällande ungdomsbrottslighet till vilka vi redan flera gånger har refererat.

Vi delade programmen i verkliga praktiska program som utvärderats av en forskare som inte var involverad i utformningen av program eller åtgärder och jämförde utfallen med demonstrationsprogram som utformats och implementerats av forskaren. Genom att helt enkelt jämföra utfallet av totala effektstorlekar i brottsåterfall framkom att medelvärdet för de praktiska programmen (.07) endast var omkring hälften av vad som uppvisades av demonstrationsprogrammen (.13), även om båda var blygsamma (men omgärdade av stora variationer).

När karakteristiken hos de praktiska programmen och demonstrationsprogrammen jämfördes, framkom ett antal specifika skillnader. De mest viktiga och intressanta följer här.

- Typ av program: minst troligt att det blir en av de mer effektiva typerna av praktiska program jämfört med demonstrationsprogram.
- Administrerade av personal som arbetar med ungdomsbrottslingar: mer troligt för praktiska program än demonstrationsprogram.
- Bevakning av integriteten i implementering av åtgärder: mindre troligt för praktiska program jämfört med demonstrationsprogram.
- Rapporterade svårigheter i implementering av behandlingen: mer troligt för praktiska program jämfört med demonstrationsprogram.
- Programmets varaktighet: omkring 25 veckor för praktiska program, omkring 38 veckor för demonstrationsprogram.
- Intensitet i behandlingen: lägre frekvens för praktiska program än för demonstrationsprogram.

Även om vissa fördelaktiga karakteristika hos demonstrationsprogrammen kan vara svåra för de praktiska programmen att efterlikna (t. ex. programtyper som fordrar högt tränad personal), är andra klart genomförbara. Resultaten från jämförelser liknande dessa kan därför användas för att leda till förbättringar av praktiska program på ett sätt som skulle öka storleken på utfallseffekterna. Validiteten hos detta perspektiv stöds av analyser av den betydande variationen inom själva domänen för praktiska program. Det är inte förvånande att praktiska program har många av de fördelaktiga programegenskaper som identifierats ovan medan andra har mindre fördelaktiga konfigurationer. Om vi undersöker medelvärdet av utfallseffekter för praktiska program som är mer gynnsamt konfigurerade i dessa termer, finner vi att de också är mer effektiva.

Figur 5 visar en sådan jämförelse för program gällande ungdomsbrottslingar som fokuserar på utfallet av återfallsbrottslighet. De praktiska programmen är kategoriserade i enlighet med hur många karakteristika de innehar från den uppsättning som funnits i meta-analyser som kan relateras till effektstorlekar. Det finns en tydlig trend för de med ett större antal gynnsamma karakteristika att producera större reduktioner i återfall bland sina ungdomsklienter jämfört med kontrollgrupper. De som saknar gynnsamma karakteristika visar på en ökning av återfall bland de ungdomar de behandlar.

Lika intressant är kanske fördelningen av de program som är representerade i meta-analyser av de olika kategorier som visas i figur 5. Mer än hälften av de utvärderade programmen hade noll eller en gynnsam karakteristika och, likaledes, minimala eller kontraproduktiva effekter. Å andra sidan hade endast 2 % av de praktiska programmen fullt antal gynnsamma karakteristika och uppnådde de högsta nivåerna för återfallspåverkan. Möjligen är de mest gynnsamt konfigurerade

programmen inte utvärderade, eller så har utvärderingarna inte rapporterats, och de kan därför vara underrepresenterade i den forskning som finns tillgänglig för meta-analyser. Det verkar dock mer troligt att de flesta praktiska program inte är konfigurerade för optimal påverkan och har betydande utrymme för förbättringar.

6. Det finns säkerhet i antal

Många faktorer påverkar det som kommer fram i en utfallstudie och även under de bästa förhållanden, är validiteten av dessa fynd osäkra. Eftersom det finns och kommer att finnas en viktig roll för resultatutvärderingar av individuella program, måste vi vara mycket uppmärksamma på tolkningen av en enkel uppsättning resultat, även från en väl utformad utvärderingsstudie. Slutligen kommer de mest tillförlitliga bevisen gällande effektiva program att komma från omsorgsfull integration av utvärderingsresultat från många studier och program. Likaledes är en av de största utmaningarna, som utvärderarna står inför, att försäkra sig om en hög kvalitet, att användbara sammanställningar av utvärderingsstudier utförs och att resultaten sprids till relevanta utvärderare, praktiker och politiker.

Ett viktigt och sentida initiativ ger stora förhoppningar om att denna utmaning skall kunna mötas. Under 1999 träffades en grupp utvärderare, politiker och forskare vid University College in London och kom överens om att starta "the Campbell Collaboration" för utveckling och spridning av systematiska sammanställningar av utfallstudier för sociala program. Denna strävan har "the Cochrane Collaboration" som förebild. Denna organiserar sammanställningar inom medicinsk hälsoforskning och har fått sitt namn efter den amerikanske psykologen och metodologen Donald Campbell, en berömd förespråkare för rigorösa programutvärderingar. "The Campbell Collaboration" (C2) har vuxit snabbt och har för närvarande medlemmar från 16 länder och koordineringsgrupper inom fält som exempelvis kriminalvård, utbildning

och social välfärd. C2 strävar efter att främja och underlätta högkvalitativa sammanställningar av utfallstudier för sociala program och göra dem tillgängliga på datanätet för alla intresserade (http://www.campbellcollaboration.org/). Även om detta arbete fortfarande är i sin linda, så har Campbell Collaboration ett flertal sammanställningar på väg. Dessa utlovar stort hopp om möjligheten att kunna få fram och dela de lärdomar som kan erhållas från tusentals studier utförda inom det kraftfulla programutvärderingsfältet.

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*.

Figure Captions

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

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

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

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

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