

**Sammendrag:**

# **Vurdering av kvaliteten på evalueringsforskning ved hjelp av meta-analyse**

Temaet for denne avhandlingen er hvordan man kan vurdere kvaliteten på evalueringsforskning ved hjelp av meta-analyse. Med evalueringsforskning menes anvendt forskning som har til hovedformål å måle virkninger av offentlige tiltak, for eksempel trafikksikkerhetstiltak. Kvaliteten på slik forskning beskrives ut fra et sett av kriterier for hva som er god forskning. Meta-analyse er en tallmessig oppsummering av resultater av en rekke undersøkelser som er gjort for å måle virkninger av bestemte offentlige tiltak.

## **Evalueringsforskning er ofte kontroversiell**

Bakgrunnen for avhandlingen er at evalueringsforskning ofte er kontroversiell. Strid om slik forskning oppstår særlig når den kommer til overraskende og kontraintuitive resultater. Eksempler på slike resultater i norsk trafikksikkerhetsforskning er funn som tyder på at oppmerking av gangfelt øker ulykkestallet og at glattkjøringskurs for bilførere øker ulykkestallet. Slike resultater blir ikke alltid trodd. Spørsmålet blir da ofte: Kan en egentlig tro på resultatene av evalueringsforskning, eller når kan en tro på resultatene av slik forskning? Hva er en god undersøkelse om virkninger av et tiltak, og hva er en dårlig undersøkelse om dette?

## **Gode og dårlige undersøkelser kan skilles fra hverandre**

Enkelte vil muligens hevde at det ikke er mulig å skille mellom gode og dårlige undersøkelser. Det hele blir til syvende og sist et spørsmål om vi liker resultatene eller ikke. I denne avhandlingen argumenteres det klart mot en slik oppfatning. Denne avhandlingens utgangspunkt er at det er fullt mulig å formulere et tilnærmet objektivt sett av kriterier for hva som er gode og dårlige undersøkelser i evalueringsforskning. Med "tilnærmet objektivt" menes at kriteriene for hva som er god forskning kan:

- 1 formuleres så klart at de ikke gir rom for sterkt divergerende tolkninger, og at
- 2 kriteriene bygger på normer for god forskningsmetode som har svært bred tilslutning blant forskere, og ikke minst at
- 3 kriteriene er uavhengige av innholdet i resultatene av en undersøkelse og dermed uavhengige av om vi "liker" eller "ikke liker" disse resultatene.

Kriterier for gode og dårlige undersøkelser i evalueringsforskning er i avhandlingen formulert med utgangspunkt i Cook og Campbells (1979) validitetssystem. Validitet defineres i denne sammenheng som graden av tilnærming til sannheten. Jo nærmere sannheten vi har grunn til å tro at resultatene av en undersøkelse, eller et sett av undersøkelser, ligger, desto høyere er validiteten. Det er i avhandlingen utformet i alt 20 kriterier for validitet i evalueringsforskning. Kriteriene er knyttet til fire hovedformer for validitet: statistisk validitet, teoretisk validitet, intern validitet og ekstern validitet.

## **Kriterier for å skille gode og dårlige undersøkelser fra hverandre**

Med statistisk validitet menes graden av tallmessig nøyaktighet, feilfrihet og representativitet i resultatene av en undersøkelse eller et sett av undersøkelser. Det er formulert ni kriterier for statistisk validitet. De fem første gjelder enkeltundersøkelser, de fire siste gjelder et sett av undersøkelser. Kriteriene gjelder:

- 1 Utvalgsmetoden som er brukt til å velge ut enhetene i en undersøkelse
- 2 Utvalgsstørrelsen, det vil si antallet enheter i en undersøkelse
- 3 Målingers reliabilitet, både for uavhengige og avhengige variabler
- 4 Forekomst av systematiske feil i datagrunnlaget i en undersøkelse
- 5 Valg av analyseteknikk for å analysere data i en undersøkelse
- 6 Sammenlignbarhet i definisjonen av de avhengige variabler i et sett av undersøkelser
- 7 Forekomst av publikasjonsskjevhet i et sett av undersøkelser
- 8 Formen på fordelingen av resultater i et sett av undersøkelser med hensyn til modalitet, skjevhet og sterkt avvikende datapunkter
- 9 Hvor robust et gjennomsnittresultat fra et sett av undersøkelser er med hensyn på måten det er beregnet på.

Teoretisk validitet betegner i hvilken grad en undersøkelse bygger på et klart formulert teorigrunnlag som forklarer resultatene av undersøkelsen. Mye evalueringsforskning er relativt ateoretisk. Kriterier for teoretisk validitet omfatter:

- 1 I hvilken grad det er formulert et eksplisitt teorigrunnlag for en undersøkelse, for eksempel i form av hypoteser som skal testes.
- 2 Om teoretiske begreper som brukes i en undersøkelse kan operasjonaliseres tilfredsstillende.
- 3 Om teorien som er formulert kan forklare hvordan det undersøkte tiltaket kan virke på det problem det er ment å løse (trafikkulykker eller personskader for trafiksikkerhetsforskning).
- 4 Om teorien som ligger til grunn for en undersøkelse støttes av resultatene av undersøkelsen eller ikke.

Intern validitet gjelder spørsmålet om i hvilken grad en undersøkelse, eller et sett av undersøkelser, gir grunnlag for å hevde at det er en årsakssammenheng mellom

det undersøkte tiltaket og de endringer som kan påvises i den eller de avhengige variablene. Det er formulert sju kriterier for kausalitet i evalueringsforskning.

- 1 Det må være en statistisk sammenheng mellom årsaksvariabelen og virkningsvariabelen.
- 2 Årsaksretningen må kunne bestemmes entydig, det vil si at det må kunne avgjøres hva som er årsak og hva som er virkning.
- 3 Den statistiske sammenhengen mellom årsak og virkning må holde ved kontroll for andre mulige forklaringer.
- 4 Det må være mulig å identifisere en årsaksmekanisme som forklarer hvordan eller hvorfor årsaken skaper virkningen.
- 5 Sammenhengen mellom årsak og virkning bør være reproduisert under varierende betingelser i flere undersøkelser.
- 6 Hvis både årsaksvariabelen og virkningsvariabelen har en stor nok variasjon, bør det være en dose-responsammenheng mellom årsak og virkning.
- 7 Hvis det er mulig å identifisere en klar målgruppe for årsaksvariabelen, bør man finne en virkning av den bare i målgruppen, ikke i andre grupper (spesifisitet i effekt).

Disse kriteriene overlapper delvis kriterier for statistisk, teoretisk og ekstern validitet. Kun kriteriene 2, 3, 6 og 7 er spesifikke for intern validitet. Ekstern validitet betegner muligheten for å generalisere resultatene av en undersøkelse utover den spesifikke konteksten den er utført i. Det dreier seg her ikke om statistisk generalisering, men om en mer skjønnsmessig vurdering av om resultater fra undersøkelser utført i, for eksempel, landene A, B og C også kan antas å gjelde i land D. Et slikt spørsmål er ofte aktuelt i evalueringsforskning, fordi ikke ethvert land kan drive egen forskning om ethvert tenkelig problem eller tiltak. Kunnskapsoverføring mellom land er det normale. Ekstern validitet kan bare bedømmes ut fra et sett av undersøkelser. Kriteriene for dette gjelder graden av sammenfall eller stabilitet i resultatene av et sett av undersøkelser:

- 1 Over tid
- 2 På tvers av landegrensler
- 3 På tvers av trekk ved konteksten undersøkelsene er utført i (relevante trekk ved konteksten må konkretiseres i hvert tilfelle).

### **Kriteriene for gode undersøkelser er anvendt i sju artikler**

De kriterier for gode undersøkelser som er formulert i del 1 av avhandlingen, er i del 2 anvendt i sju artikler publisert i fagtidsskrifter. I alle disse artiklene er meta-analyse anvendt for å oppsummere resultater av et sett av undersøkelser og sortere disse undersøkelsene etter kvalitet. Sorteringen etter kvalitet er gjort ved å kode undersøkelsene på grunnlag av de kriterier for validitet som er nevnt over. I alt er 13 av de 20 kriteriene anvendt i de sju tidsskriftartiklene. Seks artikler er publisert i Accident Analysis and Prevention i årene 1995-1998, en artikkel er publisert i Transportation Research Record i 1995.

Artiklene 1 (om vegrekkverk og støtputer), 2 (om vegbelysning) og 4 (om kjøreløys på biler) er forholdsvis like i sin oppbygging. I disse tre artiklene legges hovedvekten på å vurdere ulike sider ved statistisk validitet og intern validitet i de undersøkelsene som oppsummeres. Logoddsmetoden for meta-analyse er brukt i disse artiklene.

Artikkel 3 konsentrerer seg om ekstern validitet og viser en enkel måte for testing av stabiliteten over tid i resultatene av et sett av undersøkelser. Metoden går ut på å dele inn undersøkelsene i fraktiler og bruke resultatene av "tidlige" fraktiler, det vil si av de eldste undersøkelsene, til å predikere resultatene av "sene" fraktiler, det vil si de nyeste undersøkelsene.

Artikkel 5 (utbedring av ulykkesbelastede steder) er i sin helhet viet spørsmålet om kontroll for konkurrerende forklaringer i før-og-etterundersøkelser av utbedring av spesielt ulykkesbelastede steder. Artikkelen viser at jo bedre kontroll en undersøkelse har over en del kjente feilkilder i før-og-etterundersøkelser, desto mindre blir den virkningen som kan tillegges utbedringstiltakene. Dette mønsteret er kjent som Effektmålingenes Jernlov: Jo bedre en undersøkelse om effekten av et tiltak er, desto mindre effekt finner den av tiltaket.

Artikkel 6 handler om statistisk validitet og bruk av meta-analyse til å bedømme den statistiske validiteten i et sett av undersøkelser. I denne sammenheng drøftes kort også spørsmålet om hvordan valg av teknikk for meta-analyse kan påvirke resultatene av analysen. Dette er et spørsmål det bør arbeides grundigere med. Artikkel 6 viser for øvrig at meta-analyse kan fungere som et ypperlig diagnostisk redskap for å teste betingelsene for at det skal gi mening å beregne et veid gjennomsnittresultat fra et sett av undersøkelser. En vanlig innvending mot meta-analyser, er at slike analyser går ut på å beregne "meningsløse" gjennomsnittresultater av undersøkelser som ofte er innbyrdes svært ulike og derfor bør holdes fra hverandre. Artikkel 6 viser at det, et langt stykke på veg, er mulig å teste holdbarheten av en slik innvending innenfor rammen av meta-analyse. Det er paradoksalt nok slik at man, i alle fall et stykke på veg, må gjøre en meta-analyse for å avgjøre om en sammenveining av resultater av et sett undersøkelser i form av en meta-analyse gir mening.

Artikkel 7 har et annet fokus enn de andre seks artiklene og drøfter faktorer som påvirker kvaliteten på evalueringsforskning, herunder spesielt om forskning som publiseres i internasjonale fagtidsskrifter med peer review holder høyere kvalitet enn forskning som ikke publiseres i slike tidsskrifter. For å drøfte dette spørsmålet, anvender artikkelen et utvalg av de validitetskriterier for undersøkelser som er nevnt foran. Analysen som gjøres i artikkelen er svært enkel og må kun betraktes som eksplorerende. Den tyder likevel på at forskning som publiseres i vitenskapelige tidsskrifter ikke nødvendigvis er noe bedre enn forskning som ikke publiseres i slike tidsskrifter.

## **Meta-analyser har et stort anvendelsesområde i transportforskning**

Avhandlingen viser at en kritisk bruk av meta-analyser kan være et nyttig hjelpemiddel til å oppsummere kunnskap på områder der det foreligger et stort antall empiriske undersøkelser, og der disse undersøkelsene ikke alltid har så god kvalitet som man ideelt sett skulle ønske.

Evalueringsforskning, i det minste når det gjelder trafikksikkerhet, er ofte ikke-eksperimentell forskning, utført under stramme tidsrammer og økonomiske rammer, og ofte på grunnlag av mangelfulle data. Det er derfor ikke særlig overraskende at slik forskning ikke alltid oppfyller de kriterier for god forskning som kan stilles opp på grunnlag metodelitteraturen. Tvert om må man vente at svakheter ved datagrunnlaget og metoden er hovedregelen, snarere enn unntaket, i slik forskning.

Denne virkeligheten kan kanskje friste noen til nærmest å bli kunnskapsfornektende: Evalueringsforskningen er jevnt over så dårlig at vi ikke kan stole på noe av den. En slik innstilling er imidlertid ikke spesielt konstruktiv, fordi det er vanskelig å tenke seg at evalueringsforskningen noensinne skal kunne foregå under de ideelle betingelser som sikrer at alle kriterier for god forskning blir oppfylt i alle undersøkelser overalt og til enhver tid.

Vi må i stedet regne med at evalueringsforskningen alltid vil være av varierende kvalitet, og kun sjelden komme i nærheten av det fullkomne. Oppgaven for den som skal få fram/oppsummere de mest holdbare konklusjonene ut fra den kunnskap denne forskningen gir, blir da, enkelt sagt, å skille de gode undersøkelsene fra de dårlige. Til dette formål er meta-analyser et nyttig hjelpemiddel, men det kan aldri bli det eneste. Ikke alle kriterier for god forskning egner seg like godt for en tallmessig koding innenfor rammen av en meta-analyse.

Det synes likevel åpenbart at meta-analyser har et stort anvendelsesområde i evalueringsforskning, ikke bare i trafikksikkerhetsforskning, men også i transportforskning generelt og på andre fagområder.