TØI report
430/1999

Assessing the Validity of Evaluation Research by Means of Meta-Analysis

Case Illustrations from Road Safety Research

Dissertation for the Degree of Doctor Philosophiae
Department of Economics
The Faculty of Social Sciences
University of Oslo
1999

Rune Elvik

ISSN 0802-0175
ISBN 82-480-0091-5

Oslo, May 1999

Tittel: Assessing the Validity of Evaluation Research by Means of Meta-Analysis

Forfatter(e): *Rune Elvik*

TØI rapport 430/1999
Oslo, Mai 1999
187 sider
ISBN 82-480-0091-5
ISSN 0802-0175

Finansieringskilde:
Transportøkonomisk institutt

Prosjekt: 2526 Vurdering av kvaliteten på evaluerings-forskning ved hjelp av meta-analyse

Prosjektleder: Rune Elvik

Kvalitetsansvarlig: Marika Kolbenstvedt

Emneord:
Evaluering; validitet; meta-analyse; trafikksikkerhet

Sammendrag:
Denne avhandlingen bygger på sju vedlagte artikler, som alle er publisert i internasjonale vitenskapelige tidsskrifter. Hovedproblemstillingen i avhandlingen er om det er mulig å benytte meta-analyse som et hjelpemiddel til å bedømme den metodiske kvaliteten på evalueringsforskning. Med evalueringsforskning menes all forskning som har til hovedformål å undersøke effekter av offentlige tiltak på et bestemt område. I avhandlingen benyttes studier av effekter av trafikk-sikkerhetstiltak som eksempel.
Avhandlingen drøfter validitetsbegrepet og foreslår et sett av formelle validiletskriterier som tenkes benyttet til å bedømme den metodiske kvaliteten til evalueringsstudier. Det skilles mellom fire former for validitet: Statistisk validitet, teoretisk validitet, intern validitet og ekstern validitet. Det foreslås tjue kriterier på validitet. Ni av disse gjelder statistisk validitet, fire gjelder teoretisk validitet, fire gjelder intern validitet og tre gjelder ekstern validitet.
I de sju vedlagte artiklene brukes disse kriteriene systematisk til å bedømme validiteten til effektmålinger av trafikksikkerhetstiltak. Det konkluderes med at meta-analyse til en viss grad gjør det mulig å skille mellom gode og dårlige undersøkelser, men at man neppe kan forvente at bruk av meta-analyse vil avklare alle stridsspørsmål som omgir evalueringsforskning.

Title: Assessing the Validity of Evaluation Research by Means of Meta-Analysis

Author(s): *Rune Elvik*

TØI report 430/1999
Oslo: May 1999
187 pages
ISBN 82-480-0091-5
ISSN 0802-0175

Financed by:
Institute of Transport Economics

Project: 2526 Assessing the Validity of Evaluation Research by Means of Meta-Analysis

Project manager: Rune Elvik

Quality manager: Marika Kolbenstvedt

Key words:
Evaluation; Validity; Meta-Analysis; Road Safety

Summary:
This dissertation is based on seven appended papers, all published in scientific journals. The main research problem discussed in the dissertation is whether meta-analysis can be used to assess the methodological quality of evaluation studies. Illustrations of the use of meta-analysis for this purpose are given. The illustrations have been taken from road safety research.
The dissertation discusses the concept of validity and proposes a set of formal criteria of validity for use in assessing the quality of evaluation studies. A distinction is made between four types of validity: Statistical conclusion validity, theoretical validity, internal validity and external validity. Twenty criteria of validity are proposed, of which nine criteria concern statistical conclusion validity, four refer to theoretical validity, four refer to internal validity and three refer to external validity.
In the seven appended papers, these criteria are used systematically in meta-analyses of road safety evaluation studies. It is concluded that it is to a certain extent possible to assess the validity, and hence the methodological quality, of evaluation research within the framework of meta-analysis, but that one should not expect the use of meta-analysis to resolve all controversies surrounding this kind of research.

Language of report: English

Rapporten kan bestilles fra:
Transportøkonomisk institutt, biblioteket,
Postboks 6110 Etterstad, 0602 Oslo
Telefon 22 57 38 00 - Telefax 22 57 02 90
Pris kr 0

The report can be ordered from:
Institute of Transport Economics, the library,
PO Box 6110 Etterstad, N-0602 Oslo, Norway
Telephone +47 22 57 38 00 Telefax +47 22 57 02 90
Price NOK 0

Table of Contents

Appended Papers:

Paper 1:
The safety value of guardrails and crash cushions: A meta-analysis of evidence from evaluation studies

Paper 2:
A meta-analysis of evaluations of public lighting as an accident countermeasure

Paper 3:
Does prior knowledge help to predict how effective a measure will be?

Paper 4:
A meta-analysis of studies concerning the safety effects of daytime running lights on cars

Paper 5:
Evaluations of road accident blackspot treatment: A case of the Iron Law of evaluation studies?

Paper 6:
Evaluating the statistical conclusion validity of weighted mean results in meta-analysis by analysing funnel graph diagrams

Paper 7:
Are road safety evaluation studies published in peer reviewed journals more valid than similar studies not published in peer reviewed journals?

Preface by the Institute of Transport Economics

This report contains a dissertation for the degree of Doctor Philosophiae at the University of Oslo. It is based on seven papers published in scientific journals. These papers, in turn, were mostly written as part of a major research project to revise the Traffic Safety Handbook (Trafikksikkerhetshåndbok) that went on from 1994 to the end of 1997.

The Institute of Transport Economics would like to thank the Ministry of Transport and the Public Roads Administration of Norway for sponsoring the research that made this dissertation possible. This is the first dissertation written at the Institute of Transport Economics that employs methods of meta-analysis in order to summarise and assess the validity of empirical research. In this dissertation, meta-analysis has been applied to road safety evaluation studies. It is likely that these methods can be fruitfully applied to other areas of transport research as well.

The permission of Elsevier Science Ltd, publisher of Accident Analysis and Prevention to reprint six of the seven papers that are part of this dissertation is gratefully acknowledged. The dissertation is distributed free of charge as a public service.

The Institute of Transport Economics would like to thank the University of Oslo for evaluating this dissertation. This kind of external evaluation provides a check on the quality of the research done at the Institute, which is important both from the Institute's point of view and in relation to our sponsoring partners.

Oslo, May 1999
THE INSTITUTE OF TRANSPORT ECONOMICS

Knut Østmoe
Managing Director

Marika Kolbenstvedt
Head of Department

Authors Preface

This dissertation is a long-held dream come true. My ambition of doing meta-analyses of road safety evaluation studies started to form around 1985 strongly inspired by the pioneering contributions of Ezra Hauer to quantitative research synthesis in road safety. His contributions have continued to serve as a source of inspiration throughout the years that have since passed.

The dissertation is based on seven appended papers, all published in scientific journals. The papers were written during the years 1994-1997 and published during the years 1995-1998. The introductory synthesis was written in 1997-1998.

In finishing this work, my thanks go to many people. I wish first of all to thank professor Ezra Hauer of the University of Toronto, Canada, both for the inspiration he has provided in his own work and for being my teacher on the far too few occasions when we have met in person.

I would also like to thank economist and statistician Peter Christensen of the Institute of Transport Economics, whose early simulation studies convinced me that the logodds method of meta-analysis provided unbiased and efficient estimates of the weighted mean effect in a set of road safety evaluation studies.

Frank A. Haight, Editor-in-Chief of Accident Analysis and Prevention, is thanked for having published six of the seven appended papers and for selecting referees that contributed to improving those papers. Let me note in passing that all papers have undergone the normal reviewing process, although, as an Associate Editor of Accident Analysis and Prevention since 1997, I am authorised to accept my own papers for publication in that journal.

Inger-Anne Sætermo, formerly at the Institute of Transport Economics, now at Det norske Veritas is thanked for giving me the gentle prodding that helped me stay the course and finish this work, when giving up seemed to be a more tempting option.

Anne Borger Mysen and Truls Vaa, colleagues in preparing the Traffic Safety Handbook on which most of the papers are based, are thanked for patiently enduring my sometimes heavy handed teaching style in introducing them to meta-analysis. They have both repaid my efforts by suggesting ways in which to improve these analyses.

Finally my thanks go to the Norwegian Ministry of Transport and the Public Roads Administration, who, by sponsoring the revision of the Traffic Safety Handbook, made this dissertation possible.

Oslo,
October 1998

Rune Elvik

TØI report 430/1999
Author: Rune Elvik
Oslo 1999, 187 pages English language

Summary:

Assessing the Validity of Evaluation Research by Means of Meta-Analysis

The subject of this dissertation is how to assess the validity of evaluation research by means of meta-analysis. The term evaluation research denotes applied research designed to measure the effects of public measures taken to reduce social problems, like road accidents. The quality of this kind research is described in terms of a set of criteria of validity. Meta-analysis denotes quantitative techniques for summarising the results of a set of studies made to evaluate the effects of certain measures.

Evaluation research is often controversial

The starting point of this dissertation is the fact that evaluation research is often controversial. Controversies over evaluation research tend to start when the results of this research are unexpected or counterintuitive. Examples of counterintuitive results from road safety research in Norway include the finding that marked pedestrian crossing facilities increase the number of accidents and that skid training of car drivers increases the number of accidents. Results like these are met with disbelief. A relevant question then becomes: When can we trust evaluation studies? What characterises a good evaluation study, and what characterises a poor evaluation study?

It is possible to identify good and bad evaluation research

Some people might be inclined to say that it is impossible to identify good and bad evaluation research. In the final analysis, it all boils down to whether we like the results of a study or not. This point of view is emphatically rejected in this dissertation. It is argued that comparatively objective criteria of good evaluation research can be developed. The term "comparatively objective" implies that the criteria of good evaluation research are:

1 Stated in sufficiently clear terms to rule out highly diverging interpretations, and

2 Based on methodological principles and rules that are very widely (but perhaps not universally) supported by researchers, and not at least,

3 Independent of the results of the studies, and therefore also independent of whether we "like" or "dislike" these results.

In this dissertation, criteria of good evaluation studies have been developed within the framework of the validity system proposed by Cook and Campbell (1979). In

The report can be ordered from:
Institute of Transport Economics, PO Box 6110 Etterstad, N-0602 Oslo, Norway
Telephone: +47 22 57 38 00 Telefax: +47 22 57 02 90

this framework, the validity of a study or set of studies is defined as approximation to the truth. The more and stronger reasons we have for believing that a study or set of studies comes close to the truth, the higher is the validity of that study or set of studies. A total of 20 criteria of validity are proposed. These criteria refer to four types of validity: Statistical conclusion validity, theoretical validity, internal validity and external validity.

Criteria of validity in evaluation research

Statistical conclusion validity refers to the numerical accuracy, reliability and representativeness of the results of a study or set of studies. Nine criteria of statistical conclusion validity have been developed. The first five of these refer to a single study, the last four refer to a set of studies. The criteria are:

1 The sampling technique used in a study

2 Sample size

3 Measurement reliability, for all variables included in a study

4 The presence of systematic errors in data

5 Choice of technique of analysis

6 The commensurability of the dependent variables in a set of studies

7 Publication bias

8 The shape of the distribution of a set of results, particularly in terms of modality, skewness and outlier bias

9 The robustness of the mean result of a set of studies with respect to how it is estimated.

Theoretical validity denotes the extent to which a study has an explicit theoretical basis that provides an explanation of the findings of the study. Large parts of evaluation research are comparatively atheoretical. The following criteria of theoretical validity have been formulated:

1 The extent to which an explicit theoretical basis has been developed for a study

2 The possibility of giving adequate operational definitions of theoretical concepts used in a study

3 If the theory on which a study is based can contribute to explaining the findings of the study or not

4 If the theory on which a study is based is supported by the findings of the study or not.

Internal validity refers to the possibility of inferring a causal relationship between the measure that is being evaluated and the dependent variables this measure is intended to influence. Seven criteria of internal validity are proposed:

1 There should be a statistical relationship between the causal variable and the dependent variable.

2 The direction of causality should be clear.

3 The relationship between cause and effect should persist when confounding variables are controlled.

4 It should be possible to identify a causal mechanism that explains why the cause produces the effect.

5 The relationship between cause and effect should be reproduced in several studies, preferably made in different contexts.

6 If there is sufficient variation in both cause and effect, there should be a dose-response relationship between cause and effect.

7 If an effect is believed to exist only in certain group, it should be found only in that group and not outside it (specificity of effect).

These criteria partly overlap those of statistical, theoretical and external validity. It is only criteria number 2, 3, 6 and 7 on the above list that refer specifically to internal validity. External validity refers to the possibility of generalising the results of a set of studies to other contexts and settings than those in which each of studies in the set was made. This kind of generalisation is often desirable in evaluation research. One wants to know, for example, if the results of studies made in countries A, B and C apply to country D as well. Generalising across countries in this manner is common in evaluation research, since not every country can do its own research in every subject. Three criteria of external validity are proposed:

1 The stability of the results of a set of studies over time

2 The stability of the results of a set of studies across countries

3 The stability of the results of a set of studies across study contexts (details of the context have to be specified on a case-by-case basis).

The criteria of validity have been applied in seven journal papers

The criteria of validity proposed in part 1 of this dissertation have been applied in seven journal papers that make up part 2 of the dissertation. These papers apply meta-analysis in order to assess the validity of road safety evaluation studies. Six of the papers were published in Accident Analysis and Prevention (1995-1998), one was published in Transportation Research Record (1995). In the papers, studies have been sorted according to validity by using 13 of the 20 criteria listed above.

Papers 1 (guard rails and crash cushions), 2 (road lighting) and 4 (daytime running lights on cars) are quite similar in their general approach to analysis. All papers test various aspects of statistical conclusion validity and internal validity, with some attention paid to external validity as well. The logodds methods of meta-analysis is applied in all these papers.

Paper 3 concentrates on the external validity of studies and introduces a simple way of testing the stability of results over time. This is done by partitioning the evidence from previous studies into fractiles, and using the results from "early" fractiles, that is the first studies, to predict the results of "later" fractiles, that is the most recent studies.

Paper 5 (black spot treatment) assesses an important aspect of internal validity, which is the control of confounding variables in non-experimental before-and-after studies. Using studies of road accident black spot treatment as a case, the paper shows how different levels of control of known confounding factors can influence the results of studies. The results confirm what is known as the Iron Law of Evaluation Studies. This "law" states that the better an evaluation study is technically, the smaller are the effects it attributes to the measure that is evaluated.

Paper 6 discusses various aspects of the statistical conclusion validity of a set of results and of meta-analyses of a set of results. This paper also briefly discusses the choice of technique of meta-analysis – a subject deserving more attention. The paper shows how meta-analysis can be used as a diagnostic tool to assess if it makes sense to estimate a weighted mean result based on a sample of results. One of the most common objections to meta-analysis, is that it computes meaningless "mean effects" that paste over important differences. Paper 6 shows that, at least to some extent, it is possible to test the merits of this objection within the framework of meta-analysis. In other words, and perhaps somewhat paradoxically, one has to do at least part of a meta-analysis in order to determine if it makes sense to combine a set of results into a weighted mean by means of meta-analysis.

The focus of paper 7 is rather different from the other six papers. Paper 7 discusses factors that influence the validity of evaluation studies, in particular whether studies published in peer reviewed scientific journals score higher for validity than similar studies not published in scientific journals. In order to shed light on this issue, the paper applies the validity system developed in the other six papers and in part 1 of this dissertation. The paper shows that there is, at best, only a slight tendency for papers published in scientific journals to score higher for validity than papers not published in such journals. The analysis in this paper is, however, very simple and should be regarded as exploratory only.

Meta-analyses can be widely applied in transport research

The dissertation shows that a critical application of meta-analysis can be of help in summarising the results of studies in subjects where there is a large number of empirical studies, and some of these studies do not have the technical quality one would ideally want in evaluation studies.

Evaluation research, at least road safety evaluation research, is usually applied non-experimental research done with tough deadlines and a small budget, and usually relying on incomplete or error ridden data. It should come as no surprise that this kind of research does not always meet the strictest standards of scientific rigour as far as study design and data analysis are concerned. On the contrary, one should rather expect shortcomings in both data and methods in this kind of research to be the norm, and not the exception.

This fact may lead some people to become overly pessimistic with respect to the prospects of ever getting credible results from evaluation research: This kind of research is so flawed that we can never be in a position to trust the results of it. Such a point view is, however, not very constructive, because it is difficult to imagine that evaluation research will ever be granted terms that are maximally conducive to scientific rigour.

It is more realistic to expect the quality of evaluation research to continue to vary substantially, but only rarely come close to perfection. The task facing those who want to extract the best established knowledge from this research is, simply put, to sort out the good studies from the bad ones. Meta-analysis can help in accomplishing this task, but it can never capture all relevant considerations in assessing study quality. There are aspects of study quality that do not lend themselves to numerical coding and cannot be brought within the framework of meta-analysis.

It is nevertheless obvious that meta-analysis can be widely applied to evaluation research, not just road safety research, but transport research in general, as well as research in other subject areas.

TØI rapport 430/1999
Forfatter: Rune Elvik
Oslo 1999, 187 sider

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.

Rapporten kan bestilles fra:
Transportøkonomisk institutt, Postboks 6110 Etterstad, 0602 Oslo
Telefon: 22 57 38 00 Telefax: 22 57 02 90

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 tilnærmelse 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 gjennomsnittsresultat 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 trafikksikkerhetsforskning).

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 reprodusert under varierende betingelser i flere undersøkelser.

6 Hvis både årsaksvariabelen og virkningsvariabelen har en stor nok variasjon, bør det være en dose-responssammenheng 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 landegrenser

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 kjørelys 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 gjennomsnittsresultat fra et sett av undersøkelser. En vanlig innvending mot meta-analyser, er at slike analyser går ut på å beregne "meningsløse" gjennomsnittsresultater 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 sammenveiing 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 kunnskapsfornektere: 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.

1 Introduction

Applied research, in particular evaluation research, is generally held in low esteem in the academic world. Reasons for this are not difficult to find. Evaluation research is widely regarded as atheoretical. It rarely contributes to the development of models of general interest. The results of evaluation research are rarely published in the most prestigious academic journals. The knowledge embodied in this research therefore rarely finds its way into the material used for teaching in academic institutions. Evaluation research is often non-experimental. It is done on an ad hoc basis, often using poor data and simple techniques of analysis. Its results are therefore highly uncertain and of unknown generality. Finally, but perhaps not of least importance, evaluation research is often done on a contract basis. A sponsor with vested interests in the results pays for the research and decides what use, if any, is to be made of the results. Evaluation researchers are hence suspected of being less than perfectly objective. Cynthia Crossen (1994, 154) puts it bluntly:

> ”It is rare that a public policy study contradicts the beliefs of its sponsor. Contradictory studies suggest data so compelling that the researcher is essentially forced to shoot him- or herself in the foot by displeasing whoever is paying the bills. The sponsor usually fights back, trying to neutralize the research by disavowing it.”

Her book contains numerous examples of controversies that have arisen as a consequence of evaluation research in the United States. It is perhaps only a slight exaggeration to say that, in the United States, controversy over the results of evaluation research has become the norm. For nearly every evaluation study claiming that A is true, there is at least one study claiming that not-A is true. All findings are disputed. Policy makers are essentially free to believe whatever they like. They can almost always cite an evaluation study to support their position. It is small wonder that the status of evaluation researchers has fallen like a rock.

Is there a way out of this mess? This dissertation suggests ways of assessing evaluation research that may resolve at least some of the controversies currently surrounding it and restore some of the confidence in this kind of research. It is not suggested that every controversy can be resolved by appealing to objective criteria for assessing the quality of evaluation research. It is argued, however, that a number of methodological aspects of studies that are widely regarded as important in the scientific community, can be assessed in a fairly, if not perfectly, objective manner to help identify the best studies in a set of evaluation studies dealing with a certain subject. The basic message of this dissertation is that meta-analysis of evaluation research can be applied in order to assess its validity.

The dissertation rests on the firm belief that validity is of utmost importance in evaluation research. While some objections to this belief can be imagined and will be examined, they are in my opinion not convincing. There is indeed a profound irony in the low academic status of evaluation research, and it has to do with the role of validity in evaluation research. If a university professor fouls up an experiment, it is in most cases only his or her own academic career that suffers. Nobody else are affected. But if, say, a road safety researcher wrongly concludes that a measure he or she has evaluated is ineffective in preventing accidents, people on the road may be unnecessarily killed or injured. Evaluation researchers had better be right, otherwise lives may be unnecessarily lost or avoidable injuries may be sustained. The potential consequences of erroneous conclusions in evaluation research are of course not always this serious. But in some areas of evaluation research, particularly in subjects related to public health and safety, the potential practical consequences of erroneous conclusions in research are very serious indeed.

If status in the academic community was based on the social responsibility that researchers carry for the use of results of their research, evaluation researchers ought to be on top of the pecking order, not at its bottom. Herein lies the irony of the present low status of evaluation research.

The present dissertation is based on the appended papers, which have been published in scientific journals. The papers contain meta-analyses of road safety evaluation studies, and focus on different aspects of the validity of these studies. They illustrate the uses to which meta-analysis can be put in order to assess the validity of evaluation research in a certain subject area. The purpose of this introduction and synthesis is to summarise the appended papers and put them into a larger perspective. The introduction will be devoted to broadening the perspective and discuss some more fundamental questions that are not dealt with in the appended papers. Once the positions taken on the more fundamental questions have been clarified, a fairly detailed account of various aspects of validity and approaches to testing it is given. This account paves the way for a summary of the appended papers and a discussion of possible future developments in meta-analysis.

2 Statement of the Problem

The basic question to be discussed in this dissertation can be stated as follows:

> To what extent is it possible to assess the validity of evaluation research by conducting meta-analysis of evaluation research studies?

In order to meaningfully discuss this question, it is necessary to first deal with some fundamental issues that arise in the assessment of research. The most important of these issues include:

> Is it possible at all to establish objective criteria of validity in research? Or do the criteria accepted at any time merely reflect the dominant prejudices among researchers?
>
> Provided that criteria of validity can be established, what is the relevance of those criteria for assessing evaluation research? Should evaluation research be assessed strictly in terms of its validity, or are other bases for assessment more relevant?
>
> What forms of knowledge, and which aspects of the research process, can be incorporated into formal criteria of validity? Is any formal list of criteria of validity likely to be supported by the majority of researchers and by the public?
>
> Provided widely accepted formal criteria of validity can be established, is meta-analysis the best approach to assessing the extent to which research conforms to these criteria? Will different approaches to meta-analysis give different results?

These questions have been put in a logical sequence. The first question refers to the epistemologic basis for establishing criteria of validity in science. One school of thought within epistemology, epistemologic relativism, argues that no objective criteria can be given to separate science from pseudo-science. A leading proponent of epistemologic relativism is Paul Feyerabend (1975, 1978, 1987). His position on the status of science will be discussed in the next section. If his position is accepted, the other questions listed above become irrelevant. If it is accepted that there are no objective criteria for deciding if an activity is scientific or not, then, a fortiori, there are no criteria for deciding if it is good science or bad science.

The second question assumes that criteria of validity make sense, but raises the issue of their relevance. It has been argued, for example, that credibility is more important in evaluation research than truth. Moreover, criteria of validity generally apply strictly to the technical aspects of research, not to the issue of how topics are chosen for research. It is more important to concentrate on important social problems in evaluation research, than to study the impacts of often minor interventions that at best constitute a very limited contribution to solving the problems.

The third question concerns the possibility of developing criteria of validity that are widely accepted by researchers and fruitful in the sense that they can be applied to all forms of knowledge that are recognized as part of scientific knowledge. There is no standard definition of validity. For some common definitions, see, for example, Black and Champion (1976), Hellevik (1977), Cook and Campbell (1979) and Carmines and Zeller (1979). The lack of a standardized concept of validity entails the risk that any set of formal criteria for assessing validity will be parochial and not adequately cover all the aspects identified by the various definitions of the concept. Besides, formal criteria of validity may have greater difficulty in capturing the relevant aspects of validity of some forms of knowledge than of others. Scientific knowledge comprises not just the quantified results of empirical research, but theories, concepts and even tacit knowledge. These forms of knowledge can be difficult to assess by means of formal criteria of validity.

Finally, the fourth question raises the issue of whether meta-analysis is the best approach for assessing the validity of research, granted that criteria of validity have been formulated. Meta-analysis is quantitative. This means that it is more readily applied to those aspects of research that are quantified than to aspects that are difficult or impossible to quantify. Several techniques of meta-analysis exist. Which of these techniques, if any, is the best one to use if one wants to assess the validity of a set of studies? This question needs to be answered, otherwise the element of arbitrariness in the results of meta-analyses designed to assess the validity of a set of studies may be felt to be too large.

Before discussing these questions more carefully, it is necessary to briefly discuss and define the key concepts of this dissertation. They are: evaluation research, validity, assessment of validity and meta-analysis.

3 A Brief Discussion of Key Concepts

The basic problem to be discussed in this dissertation was formulated in section 2. The key concepts involved in the discussion of this question are: evaluation research, validity, assessment of validity and meta-analysis. The concepts will be discussed in that order.

Evaluation research denotes applied research designed to estimate the effects (impacts, consequences) of measures (interventions, programs) implemented to alleviate social problems. The terms effects, impacts and consequences are used interchangeably. They all denote the dependent variable in evaluation research, which is usually the size of the change in a quantitative variable that measures the prevalence or severity of a certain social problem. Typical examples of social problems that are the subject of evaluation research include crime, poverty, unemployment, accidents and drug abuse. Measures taken to alleviate the problems may be of a technical, economic or behavioural nature. The terms measures, interventions and programs are used interchangeably. Introductory textbooks in evaluation research include Weiss (1972), Cook and Campbell (1979), Rossi and Freeman (1985), Pollard (1986), Mohr (1992) and Stern and Kalof (1996).

Validity will be defined in this dissertation as the degree to which research approximates the truth. This definition is taken from Cook and Campbell (1979). It is preferred to the more common definition given in, for example, Hellevik (1977), which states that research is valid to the extent it measures what it purports to measure. As will become apparent in subsequent sections of this dissertation, the definition of validity given by Cook and Campbell (1979) covers more aspects of the concept than any other definition found in social science textbooks. The words ”approximates the truth” in the definition are used deliberately, since researchers can never claim to know the truth for sure. The best that can be accomplished in empirical social research, is to conduct studies in ways that are not known to lead to systematic errors, and to argue on that basis that the results are not (positively) known to deviate from the truth. This, however, is not the same as to claim that the truth has been found.

Assessment of validity denotes a systematic evaluation of the validity of research for the purpose of identifying the most valid studies in a set of studies dealing with a certain subject. In order to be included in an assessment of validity, all studies should deal with the same subject; hence, assessment of validity requires a delineation of the subject for which the validity of studies is to be assessed. The main point of conducting an assessment of validity is, of course, to get as close to

the truth as possible. It will be assumed that validity comes in degrees. It will not be assumed that an assessment of validity is, or ought to be, entirely quantitative.

Meta-analysis denotes a family of statistical techniques that have been developed for the purpose of synthesising or summarising the results of a set of evaluation studies. Meta-analysis is the quantitative analysis of literature. It will often be the case that, say, some 15-20 evaluation studies have estimated the effects of a measure. The results of these studies are likely to differ. Meta-analysis seeks to answer the question of what is the best estimate of the average effect of the measure, by using statistical techniques to summarize the results of the studies. It also investigates sources of variation in study findings, including the technical quality of the studies. Introductory textbooks in meta-analysis include Fleiss (1981), Glass, McGaw and Smith (1981), Light and Pillemer (1984), Hedges and Olkin (1985), Wolf (1986), Hunter and Schmidt (1990), Rosenthal (1991) and Cooper and Hedges (1994).

A more detailed discussion of these concepts, particularly the concept of validity, will be undertaken in subsequent sections of the dissertation.

4 The Arguments of Epistemologic Relativism

If concepts like truth and reason are as elusive as argued by epistemologic relativism, the task of trying to assess the validity of research may founder before it gets started. This section will discuss some of the arguments of epistemologic relativism as they have been presented by its most outspoken advocate, Paul Feyerabend, concentrating on those arguments that seem to be most relevant to the subject of this dissertation.

One of the main points of epistemologic relativism is that no objective criteria exist to separate science from non-science. Feyerabend (1987, 5) defines objective as ”valid irrespective of human expectations, ideas, attitudes and wishes”. He argues that (1987, 304) ”the way in which scientific problems are attacked and solved depends on the circumstances in which they arise, the means available at the time *and the wishes of those dealing with them. There are no lasting boundary conditions of scientific research*.” (emphasis added).

It follows from this that it is not possible to distinguish on an objective basis between good and bad science. Feyerabend states (1987, 75) that ”what counts as evidence, or as an important result, or as ”sound scientific procedure”, depends on attitudes and judgements that change with time, profession and occasionally even from one research group to the next.” He further claims that (1987, 36): ”There is no one ”scientific method”, but there is a great deal of opportunism; anything goes - anything, that is, that is liable to advance knowledge as understood by a particular researcher or research tradition.” The widespread belief that knowledge grows and is refined as research makes progress is dismissed as unfounded by Feyerabend (1987, 188): ”The development of knowledge is not a well planned and smoothly running process; it, too, is wasteful and full of mistakes; it, too, needs many ideas and procedures to keep it going. Laws, theories, basic patterns of thinking, facts, even the most elementary logical principles are transitory results, not defining properties of this process.”

According to Feyerabend, normative epistemology, as taught in textbooks and propagated by, for example, Popper (1979) is just a set of post hoc rationalizations of opportunistic choices made by researchers who were not always motivated by an interest in the truth exclusively, but may have taken their own future academic careers into consideration as well. He repeatedly stresses that ”science is just one tradition among many”, clearly implying that truth is just one virtue among many.

Feyerabend is known to be deliberately provocative (Siegel 1989). However, by yielding to that temptation, Feyerabend has painted himself into a corner he cannot get out of. The problem is essentially one of self contradiction. Feyerabend says that science is just one tradition among many. So indeed are Feyerabend’s

own views of science. They are just one point of view among many. Complete relativism is completely self contradictory. If, as argued by Feyerabend, certain normative theories of science cannot be rationally justified, then neither can the argument that such theories cannot be rationally justified. Principles of rational argument either exist or they do not. If they do not exist, Feyerabend cannot use them to defend his points of view. If they do, then complete relativism cannot be correct.

Feyerabend uses rational argument to argue against rationality and reason (Siegel 1989). Although insisting on the opposite, he is in fact fully committed to the objectivity of reasons and arguments. Otherwise, nobody would have any reason to take any of Feyerabend's arguments seriously. But Feyerabend clearly intends his arguments to convince other people.

Hovi and Rasch (1996, 19), in discussing Feyerabend's position, point out that the fact that science may have been less than perfectly rational at certain times cannot be invoked as an argument for rejecting an *ideal* of scientific rationality. It does not make sense, they conclude, to argue against scientific rationality altogether, only against particular interpretations of scientific rationality.

What, then, are the most convincing elements of relativism? It is certainly true that the normative standards of good science have evolved over time and are neither immutable nor independent of the social setting in which they were developed. Bertrand Russell has nicely captured the social basis of preferences in his theory of the origins of Hell (1935, 143):

> "Norway and Sicily both have ancient traditions; they had pre-Christian religions embodying men's reactions to the climate, and when Christianity came it inevitably took very different forms in the two countries. The Norwegian feared ice and snow; the Sicilian feared lava and earthquakes. Hell was invented in a southern climate; if it had been invented in Norway, it would have been cold."

It seems likely that influences of a similar nature (though not in a literal sense, of course) have shaped the development of normative standards of science. The invention of computers has made it possible to conduct vastly more complex mathematical and statistical analyses of data than before computers were invented. Studies that do not avail themselves of these opportunities are more likely to be labelled as simplistic and naive today than similar studies were 50 years ago. In this sense, there is clearly an element of relativism in how the scientific community rates the quality of studies.

This kind of relativism is, however, completely harmless as far as the prospect of developing an objective set of criteria for rating studies according to validity is concerned. It does not preclude the development of such a rating system. It only means that the rating system will be subject to changes over time as research methodology becomes more sophisticated.

In recognition of this fact, it is not claimed that the set of criteria for assessing study validity that will be proposed in this dissertation can be applied universally. It is, at best, applicable to evaluation research as it is currently done in the Western countries.

5 The Relevance of Validity in Evaluation Research

Evaluation research is applied research. The results of evaluation studies are usually intended to serve as a basis for making decisions concerning the programs or measures that have been evaluated. But the results of evaluation studies are not always taken seriously by those who are in charge of the programs subject to evaluation. In particular, if an evaluation study shows that the program is ineffective, or even counterproductive, the sponsoring agency will be tempted to argue that the evaluation study is flawed and cannot be used as a basis for policy making.

Most evaluation researchers who have been in the business for some years will at least once have experienced the frustration of not being believed or being attacked by the sponsoring agency, because the evaluation did not give the results the sponsor wanted. These frustrations are vividly expressed in the volume edited by Palumbo (1987). Palumbo himself opens by stating that (1987, 31) that ”there is no single, true set of facts; the facts one looks for are determined by the epistemological and political values that guide the inquiry.” He adds (1987, 32) that ”values are a part of any evaluation. This means that evaluations will not result in a ”correct” finding; they will take a political position about the desirability of various goals, whether *directly*, by judging that the goals are worthwhile, or *indirectly*, by concluding that the goals are being achieved efficiently.” (italics in original).

It is difficult to make much sense of these comments. It is, of course, true that a very large part of evaluation research has an explicit normative basis. The research is done for the purpose of solving or alleviating a social problem. But this does not imply that the determination of matters of fact is based on the policy objectives that evaluation research is intended to serve. To suggest so is, effectively, to say that evaluation research is nothing more than an exercise in wishful thinking. Although road safety research, to take one example, is intended to contribute to improving road safety, this policy objective is not relevant for determining whether a certain safety measure is effective in reducing the number of accidents or not. According to the philosophy of science espoused in this dissertation, matters of fact can be determined according to criteria that are entirely independent of the purposes for which the research is being conducted. This point of view will be elaborated in chapter 8, dealing with operational criteria of validity in evaluation research.

Nevertheless, the current system for carrying out evaluation research is a problem, because the sponsors of research have no institutionalised interest in finding the truth about the programs they carry out. Hauer (1991, 137) puts it like this: ”It is in the nature of road safety that it is not visible to the naked eye. Nobody can tell whether a programme was a success or failure unless trained and independent researchers are given an opportunity to devise and carry out long-term studies. By the time estimation of programme effect is possible, the public body has already developed a large stake in its success. Under these circumstances why should the stewards of public bodies wish to find out what effect their programme has had? Nobody is attracted by the possibility of political, institutional, professional or personal embarrassment. The upshot is that programmes are rarely evaluated, and if evaluated, this is done ”in-house”, with success eagerly sought and failure unpublished. In this inhospitable soil, spindly flowers of factual knowledge grow in the shadow of the weeds of misinformation.”

Hauer’s point of view are entirely consistent with the position that objective truth exists; the trouble is that no powerful interests are pushing for its discovery. Guba and Lincoln (1987, 210), on the other hand argue that what they call ”objective reality”, that is a reality that exists independent of the interest that human beings may exhibit in it, is untenable. This point of view is not supported in this dissertation. One is, in fact, tempted to invite Guba and Lincoln to the top of a high building and ask them to jump from it, in order to test if they really are convinced that gravity is not a part of objective reality, but merely a figment of the human imagination.

To summarise, it is argued in this dissertation that objective criteria of validity, in the sense that these criteria are: (1) independent of the objectives for which research is carried out and (2) widely shared by evaluation researchers, can be formulated. It is, on the other hand, not claimed that actual debates about the merits of evaluation research are conducted solely in terms of these criteria of validity. One needs only to open a newspaper to ascertain that the positions taken by participants in debates over evaluation research are very often influenced by their vested interests primarily, not by an overriding desire to discover the truth.

6 Concepts of Validity and Forms of Knowledge

6.1 The multiplicity of concepts of validity

Do widely shared criteria of validity for evaluation research exist? A quick glance at some textbooks in the methods of social research would seem to suggest otherwise. Every author seems to propose his or her own definition of validity and his or her own techniques for testing validity.

Black and Champion define validity (1976, 222) as ”the property of a measure that allows the researcher to say that the instrument measures what he says it measures.” A measure is valid, in other words, if it actually measures what it purports to measure. Black and Champion go on to distinguish between three main types of validity: content validity (or face validity), predictive and concurrent validity and construct validity. They do not formally define content validity, but from their discussion of the concept one can infer that it refers to the way in which theoretical concepts are operationalized. Predictive validity is defined as the association between what a test predicts behaviour will be and the subsequent behaviour exhibited by an individual or group. Concurrent validity differs from predictive validity in that the scores of predictive behaviour are obtained at the same time as the exhibited behaviour. Finally construct validity refers to the success in constructing external criteria to measure unobservable traits, like various mental states and predispositions.

Black and Champion distinguish between validity and reliability. Reliability is defined as the ability of measuring instrument to measure consistently the phenomenon it is intended to measure. They point out that reliability is a necessary condition for validity: a test that is unreliable is never valid, whereas a valid test is always reliable as well.

Hellevik’s discussion of validity and reliability in a standard Norwegian textbook in research methods in sociology and political science (Hellevik, 1977, 155-171) closely follows Black and Champion’s discussion of these concepts. Hellevik defines validity as the relevance of data for the research problem a study is designed to answer. He defines reliability as the accuracy with which the variables included in a study are measured. He discusses in fairly great detail various techniques for testing reliability. As far as validity is concerned, his discussion is more brief. In fact, Hellevik comes close to claiming that validity cannot be tested, by stating (1977, 167) that ”the degree of concurrence between the theoretical and the operational definition of a concept is usually not amenable to direct empirical testing.” He adds, however, that it is sometimes possible to develop several operational definitions of the same theoretical concept and study

the correlations between measurements based on the different operational definitions. He ends his discussion of validity on the following rather pessimistic note (1977, 170): ”Despite the fact that validity is a very central concept in research methodology, there seems to be widespread confusion with respect to the meaning of the various terms (like content validity, construct validity, internal validity, etc) that are used to denote the concept.”

Carmines and Zeller (1979) discuss reliability and validity assessment in social research. They define reliability (1979, 11) as ”the extent to which an experiment, test, or any measuring procedure yields the same results on repeated trials.” Validity is defined (1979, 12) as the extent to which a measuring instrument does what it is intended to do. Validity, according to Carmines and Zeller, concerns the crucial relationship between concept and indicator. They go on to distinguish between criterion-related validity, content validity and construct validity. These concepts are closely analogous to the concepts of predictive, content and contruct validity proposed by Black and Champion. Carmines and Zeller interpret all these types of validity as referring to various aspects of the relationship between a theoretical concept and its empirical referent.

Cook and Campbell (1979) present an extensive discussion of validity in which they distinguish between four types of validity and a total of 33 so called ”threats to validity”, whose presence or absence from a specific study determine how valid it is. The validity framework developed by Cook and Campbell is definitely the most elaborate currently available in social research. Its various elements will therefore be discussed in some detail.

The first type of validity defined by Cook and Campbell is denoted statistical conclusion validity and refers to how well supported inferences about a statistical relationship, or covariation, between two variables are. Cook and Campbell identify seven threats to statistical conclusion validity, of which the most relevant for evaluation research include:

1 *Lack of statistical power*: In small samples, detecting a relationship between some ”treatment” and a measure of the effects of treatment is more difficult than in larger samples.
2 *Violated assumptions of statistical tests*: It is often convenient to rely on the standard normal distribution when testing the statistical significance of findings. This assumption may, however, be seriously wrong, as not all phenomena obey the normal distribution. Counts of accidents, in particular, do not conform to the normal distribution.
3 *Fishing and the error rate problem*: Sometimes, multiple tests are made on the same data set. If not guided by prior hypotheses or theory, this is called ”fishing” or ”data mining”. By analysing the data this way, researchers will almost always happen to find a statistically significant relationship between some variables. The problem is, however, that any data set will by chance contain some significant relationships.

4 *Unreliability of measures*: Low reliability in the data set reduces the chances of detecting true effects or relationships between variables.
5 *Unreliable treatment implementation*: A special problem in evaluation research, is the extent to which the treatment whose effects are evaluated has actually been implemented. Sometimes implementation is easily monitored, on other occasions this is more difficult.

Cook and Campbell treat reliability as an aspect of statistical conclusion validity, thus obviating the need for a distinction between reliability and validity. This would seem to be a reasonable approach, granted that reliability is a necessary, but not sufficient condition for validity.

The next type of validity discussed by Cook and Campbell is denoted internal validity. By internal validity, Cook and Campbell refer to the possibility of inferring a causal relationship between two or more variables. They point out that one must first establish that two variables covary, since the presence of a statistical relationship between two variables is a necessary, but not sufficient condition for the existence of a causal relationship. Cook and Campbell identify thirteen threats to internal validity, of which the most relevant in the present context include:

1 *History*: This threat is relevant in evaluation studies relying on a before-and-after design. It denotes an event that takes place between the before and after period and whose effect may be mixed up with the treatment that is evaluated.
2 *Maturation*: This threat is also relevant in evaluation studies relying on a before-and-after design. It denotes the presence of general, long term trends in the dependent variable that can be mistaken for a treatment effect.
3 *Statistical regression*: Once again, this threat to internal validity is particularly relevant in before-and-after studies, although it may in principle be relevant to other study designs as well. It denotes the effects of random fluctuations on successive measurements of the same variable. If, for example, an abnormally high number of accidents was observed in the before period, a subsequent decline towards the long term mean number of accidents would be expected to occur even if no treatment had been introduced. This threat to internal validity is highly relevant in many road safety evaluation studies.
4 *Self selection*: This threat to internal validity is particularly relevant in cross section, case-control or other comparative study designs. It denotes bias that may arise in the comparison of those who have received a treatment and those who have not, if those who received the treatment voluntarily chose to do so, rather than being assigned to the treatment or control conditions at random.
5 *Mortality*: This threat to internal validity refers to the tendency for experimental subjects to drop out from an experiment the longer it lasts. It is therefore most relevant in long term studies involving human subjects.

6 *Ambiguity of causal direction*: It is not always possible to ascertain the direction of causal influence. This threat to internal validity is most relevant in cross section studies.

As is apparent from this list of threats to internal validity, the threats that are relevant depend on study design. In principle, an experimental study design, involving the random assignment of study subjects to one or more treatment conditions and a control condition not getting any treatment, eliminates all threats to internal validity on the list above.

The third type of validity discussed by Cook and Campbell is construct validity. They do not formulate a formal definition of construct validity. However, their discussion of it clearly indicates that construct validity denotes the adequacy of operational definitions of theoretical concepts and propositions. Ten threats to construct validity are discussed, of which the most relevant for the present study include:

1 *Lack of clarity in theoretical definition*: If the theoretical definition of a concept is vague, operationalising the concept adequately becomes difficult.
2 *Mono-operation bias*: A theoretical concept can often be given several operational definitions. If the results of empirical studies based on multiple operational definitions of the same concept agree, these studies constitute a stronger test of the validity of the concept than if just one operational definition was used.
3 *Mono-method bias*: By the same token, if the results of studies using different methods agree, more confidence can be placed in the results than if just one method had been used or the results of studies using different methods diverged.

The fourth and final type of validity discussed by Cook and Campbell is external validity. It denotes the possibility of generalising research findings to other settings or contexts than those in which the studies were made. According to Cook and Campbell, this amounts to testing whether there are statistical interactions in study findings across the variables over which one wishes to generalise findings. If, for example, studies made in different countries get different results, then generalising across countries would not be justified. If, on the other hand, results were the same in all countries, generalising across countries would be more defensible, especially if studies have been made in a broad set of countries. The three threats to external validity listed by Cook and Campbell are:

1 *Interaction of selection and treatment*: This threat to external validity refers to whether treatment effects vary depending on how treatment subjects were recruited for treatment.
2 *Interaction of setting and treatment*: This threat to external validity refers to variation in treatment effect with respect to study setting.

3 *Interaction of history and treatment*: This threat to external validity refers to variation in treatment effect with respect to when studies were conducted.

The validity framework of Cook and Campbell is very comprehensive and captures all aspects of validity discussed by other authors (Black and Champion 1976, Hellevik 1977, Carmines and Zeller 1979). While both Black and Champion (1976), Hellevik (1977) and Carmines and Zeller (1979) focus mainly on construct validity, or how to operationalize theoretical concepts, Cook and Campbell recognise that this focus is too narrow for evaluation research, whose main objective rarely is to determine if a certain theoretical concept can be adequately measured or not. In fact, much of evaluation research is more or less atheoretical. It merely tries to determine the effect of some public program or policy and rarely discusses the theoretical implications of the findings.

This dissertation does not subscribe to Hellevik's suggestion that there is widespread confusion about the meaning of validity in social science. What seems to be the case is rather that different authors emphasize differents aspects of validity. In theoretical research, whose main objective is concept formation and theory development, it is of course essential to focus on construct validity. In evaluation research, on the other hand, internal validity is more important.

It is nevertheless true that no universally accepted concept of validity exists in social research. Perhaps the diversity of topics and methods in social research is too great to be encompassed by a single, unifying and universally accepted concept of validity. Rather than trying to develop such a concept, this dissertation seeks to develop a validity framework specifically suited for evaluation research, and developed within the context of road safety evaluation research. No claims are made to the effect that this validity framework is universally applicable. The standard for judging the success or failure of the framework is whether it can be used to distinguish between good and bad evaluation studies within the specific area of knowledge for which it was developed.

6.2 The concept of objective knowledge

One reason for the lack of standardized concepts in social research may be that the standards for what counts as knowledge are subjective. If no universally accepted standards of knowledge exist, there is likely to be a profileration of parochial concepts of validity, based on the personal standards of knowledge of each researcher.

In discussing what ought to count as scientific knowledge, epistemology has traditionally relied on a *subjective conception of knowledge*, in which knowledge is regarded as justified true belief. Within this framework, knowledge cannot exist without a knowing subject. In short, a justified and true statement does not constitute knowledge unless someone is aware of the statement and believes it.

This conception of knowledge lies close to everyday usage of the term. Hauer, for example, in discussing the state of knowledge with respect to the effects of road safety measures, states (1988, 3): ”My own critical views about the amount of factual knowledge that is available in the field of road safety delivery rest on years of study. As I moved from one inquiry to another and began to notice how shallow are the foundations of what passes for knowledge, I gradually realized that ignorance about the safety repercussions of the many common measures is not the exception.” Three years later, he remarked (Hauer 1991, 135): ”How little we know about the safety consequences of our road design decisions and about the repercussions of our traffic control actions is simple to demonstrate. One needs only to ask the engineer: ”Approximately how many accidents per year do you expect to occur with design X?” While the engineer might venture an opinion, in truth, the arsenal of knowledge at the disposal of the North American engineer just does not suffice to give an answer.”

While conforming both to everyday usage and the traditions of epistemology, the subjective conception of knowledge creates a number of difficulties. Although it makes sense to say that person A knows more about a subject than person B, if person A can pass a more difficult examination about the subject than person B, it hardly makes sense to say that the amount of knowledge that is available to the general public concerning a subject is determined primarily by how much person A can remember when undergoing an examination about the subject.

Karl Popper has introduced the concept of objective knowledge (Popper 1979), which he defines (1979, 73) as ”the logical content of our theories, conjectures, guesses.” He adds that: ”Examples of objective knowledge are theories published in journals and books and stored in libraries; discussions of such theories; difficulties or problems pointed out in connection with such theories, and so on.” Knowledge in the objective sense, according to Popper (1979, 109), is knowledge without a knower; it is knowledge without a knowing subject.

In short, the *concept of objective knowledge* can be defined as all results of research, theoretical or empirical, that are available to the general public by virtue of being written or otherwise stored in a medium that is accessible to anyone who wants to learn its contents. Knowledge in this sense exists, as pointed out by Popper, in the shelves of libraries and archives. This kind of knowledge is objective in the sense that it exists irrespective of whether anyone keeps it inside his or her head. It is, however, not necessarily objective in the sense that everyone who reads a certain paper in a journal will find the results reported in the paper convincing and therefore believe them, as required according to the subjective conception of knowledge.

The framework proposed in this dissertation to assess the validity of evaluation research is intended to apply to the body of objective knowledge derived from such research. It applies to published, or at least written studies, and not to oral communications, personal beliefs, tacit knowledge or other forms of subjective knowledge.

Restricting the scope of the validity framework to objective knowledge in this sense has both advantages and drawbacks. The chief advantage is that the system for assessing the validity of evaluation research itself becomes objective, by (1) having a clearly defined empirical reference (i e the set of documented studies dealing with a subject), (2) relying on explicitly stated criteria (i e using a list of clearly defined criteria of validity and a system for scoring studies according to these criteria), and (3) becoming testable, in the sense that agreement between researchers in the use of the criteria of validity can be determined experimentally.

The drawback, on the other hand, is that a set of explicit criteria of validity, applied to a set of published (or at least documented) studies, may be regarded as an overly restrictive and highly simplistic way of assessing the validity of evaluation research. There is no doubt that scientific knowledge comprises not just objective knowledge in the Popperian sense of the term, but also subjective knowledge and even tacit, or subconscious, knowledge. Hence, it can be argued that assessing the quality of knowledge about a certain subject in terms of objective knowledge exclusively cannot adequately represent the highly complex interplay of the various forms of knowledge that, put together, constitute what most researchers and laymen would regard as ”what is known” about a subject.

This point is readily conceded. However, three points can be made in response to it. Firstly, the set of criteria for assessing the validity of evaluation research that are proposed in this dissertation are intended as *normative* criteria, not as descriptive criteria. The criteria are explicitly normative in the sense that they summarise the points that ought to be emphasized when debating the merits and demerits of a certain contribution to evaluation research. All too often, debates about evaluation research revolve around the contents of the results, rather than the methodological rigor of the research, and are heavily influenced by vested interests, rather than a disinterested search for the truth (see Crossen 1994 for some striking examples of these tendencies).

Secondly, it is recognised that a set of normative criteria is bound to be incomplete, in the sense that it does not exhaust the considerations that are regarded as relevant in assessing the validity of research. To give an example of what is meant by this, consider the following case. Two evaluation studies that are identical in terms of all formal criteria of validity have been reported. However, in one of the studies the authors carefully discuss the shortcomings of the study. In the other study, no mention is made of any shortcomings and the authors are highly confident in stating their conclusions. Which of these studies is likely to be regarded as the best one by a senior researcher in this area? There is little doubt that the study discussing its own shortcomings would be regarded as the best one, because the authors clearly show that they are aware of the limitations of the study. But it is very difficult to turn this assessment into a formal, normative criterion of validity. The nature of the assessment is such that it is bound to be more or less subjective and difficult to formalise.

Thirdly, while an informal and subjective assessment of the validity of research can reflect considerations that are difficult to formalise, it is nevertheless likely to be subject to more or less unknown biases. No matter how hard we try to be objective, there is always a risk that we go by the rule that ”bad studies are ...

those whose results we do not like." (Rosenthal 1991, 130). By assessing validity in terms of formally stated, normative criteria, the role of personal prejudices in the assessment can be minimized. This argument for basing the assessment of the validity of evaluation research on formally stated criteria of validity and a scoring system for those criteria is elaborated in the next chapter.

7 The Pitfalls of Informal Research Syntheses

Meta-analysis is a comparatively recent innovation in scientific methodology. Like many other scientific innovations, it has been greeted by considerable skepticism. When the first meta-analyses were reported in psychology in the mid nineteen seventies, the renowned British psychologist H. J. Eysenck (1978) labelled them ”An exercise in mega-silliness” and rejected the basic concept underlying meta-analysis – that it makes sense to try to combine evidence from several studies by means of quantitative methods – as basically untenable. Related points have been made by numerous other critics. For surveys, see Glass, McGaw and Smith (1981) and Cooper and Hedges (1994).

Critics of meta-analysis are obviously right in claiming that it, like any other scientific technique, can be abused and that it cannot address every conceivable issue that might arise in trying to summarise the state of knowledge in a specific area. What the critics of meta-analysis tend to overlook, is the fact that informal research syntheses are likely to be prone to a number of well known biases that can invalidate their conclusions. By an informal research synthesis is meant a narrative survey of research literature dealing with a subject. An informal research synthesis does not employ any formal techniques for summarising evidence from the studies it includes. In the usual format, a narrative research synthesis consists of a brief presentation and discussion of each study that has been reported. Studies are often presented in chronological order. Following the presentation of each study, general conclusions are drawn based on an informal assessment of study quality and the reviewer’s subjective impression of the results.

Experimental psychology has documented that human beings employ a number of mental heuristics, or simplifying techniques and shortcuts, when trying to make sense of complex data. These heuristics lead to systematic biases that may invalidate the conclusions of analyses that are based primarily on informal techniques, that is on the mental heuristics. In this chapter, a brief summary and illustration of some of these biases will be given. These include:

1 Confirmation bias
2 Hindsight bias
3 Publication bias
4 Belief in the law of small numbers
5 Capitalisation on chance

Confirmation bias denotes the tendency to look for evidence that supports a hypothesis, rather than evidence that disconfirms it. The existence of confirmation bias in hypothesis testing has been found in several experimental studies, starting with Wason's experiments in the nineteen sixties (Wason 1960, 1968), designed to elicit the rules that people applied when testing a hypothesis. Wason found that experimental subjects tended to look for evidence that would support their hypothesis, rather than evidence that would disconfirm it. For a survey of studies of confirmation bias, see Klayman and Ha (1987).

Confirmation bias influences not just what kind of evidence people regard as relevant for testing a hypothesis, but also their interpretation of research findings. An example of an interpretation of the findings of a road safety evaluation study that appears to be based on confirmation bias is found in a report by Blakstad and Giæver (1989). The report compares the accident rate on various types of road in urban and suburban areas. Contrary to prior expectations, Blakstad and Giæver (1989, 12-13) find that the accident rate is higher on roads with a separate track for pedestrians and cyclists than on roads with no such track. However, they dismiss this result, stating that "separate tracks for pedestrians and cyclists have been constructed only along roads where the accident rate was abnormally high, but their safety effects are too small to bring down the accident rate to a level below that for roads without such tracks." They invoke the results of before-and-after studies that have found a decline in the number of accidents when tracks for pedestrians and cyclists were constructed to support this interpretation of the findings.

Later in the report (1989, 18), Blakstad and Giæver report the results of a comparison of accident rates on access roads with and without speed humps. As expected, the accident rate was lower on roads with speeds humps than on roads without them. They readily interpret this an an effect of the measure, stating that "speed reducing devices appear to be effective in residential areas." In other words, when the findings supported their hypothesis, Blakstad and Giæver took them as evidence for the effect of the safety measure. When, on the other hand, the findings did not support their hypothesis, they dismissed them as the result of study artifacts.

Their reasoning is, however, not tenable. If it is correct that tracks for pedestrians and cyclists have been constructed along roads with an abnormally high accident rate, then the results of the before-and-after study that Blakstad and Giæver refer to (a Norwegian study by Ørnes 1981) cannot be used to support their argument, because that study had a fatal methodological flaw. It did not control for regression-to-the-mean, a highly likely source of error in a before-and-after study of a safety measure introduced at locations with an abnormally high accident rate.

It is therefore likely that the interpretations offered by Blakstad and Giæver reflect confirmation bias. This example shows that a rather careful reading of evaluation studies may be needed in order to expose confirmation bias. Moreover, the example shows that in order to determine whether confirmation bias may have influenced the interpretation of research findings, it may be necessary to evaluate the methodological rigor of studies that authors subject to confirmation bias refer

to in order to support their interpretation of the findings of their own study. Blakstad and Giæver's argument sounds plausible at a superficial level and unravels only when examined critically.

It is not always possible to argue that confirmation bias may have influenced the interpretation of research findings in the manner illustrated above. The possible presence of an undetectable confirmation bias in informal research syntheses is a serious source of bias.

Hindsight bias denotes the tendency to discount surprises by adjusting prior expectations to conform to the outcome of an event or experiment. Hindsight bias is typified in the exclamation "I knew it would happen; I could have told you beforehand!" In science, the most common form of hindsight bias is perhaps the tendency to propose *ad hoc hypotheses to explain anomalous findings*. It is nearly always possible to come up with a hypothesis that explains a finding, at least in applied social science, where few, if any, findings can be ruled out a priori by reference to universal laws. Hindsight bias was first studied by Fischhoff (1975; Fischhoff and Beyth 1975), subsequently by Slovic and Fischhoff (1977). Excellent reviews of subsequent research have been given by Hawkins and Hastie (1990) and by Christensen-Szalanski and Willham (1991). In informal research syntheses, the temptation to propose apparently reasonable explanations to unexpected findings is almost irresistible. A subtler form of hindsight bias occurs when researchers *formulate their hypotheses post hoc to make them fit the findings of a study*. The study is then dressed up to make it look as if the hypotheses were derived deductively before the findings were known and were tested as part of the study.

There is no way of knowing exactly how widespread this practice is. One may fear, however, that it is fairly widespread in parts of social science. The temptation to theorise post hoc could of course compromise the scientific integrity of a meta-analysis as well. However, meta-analysis imposes a framework for interpretation of research findings that constrains post hoc theorising. There are, for example, formal tests to determine whether an anomalous finding is really anomalous or simply the product of random variation in study findings. The explanatory value of hypotheses proposed post hoc can also be determined statistically in meta-analysis.

Publication bias denotes the tendency not to publish studies that are believed not to contribute to knowledge, or believed not to have any practical interest. There are, broadly speaking two kinds of publication bias: (1) Bias against results that are not statistically significant at conventional levels, and (2) Bias against results that are regarded as anomalous, go in the "wrong" direction or otherwise seem difficult to interpret on the basis of accepted conventions. Publication bias has been documented in a number of studies (Rosenthal, 1979; Peters and Ceci, 1982; Light and Pillemer, 1984; Coursol and Wagner, 1986; Begg and Berlin, 1988; Berlin, Begg and Louis, 1989; Dickersin and Min, 1993).

Unless there is direct evidence of publication bias, in the form of information in published studies referring to the results of unpublished studies, it may be difficult to detect publication bias in an informal research synthesis. In meta-analysis, on the other hand, there are a number of formal techniques that are designed to detect the presence of publication bias and determine its magnitude (Begg, 1994). By applying these techniques one may, at least partially, adjust for publication bias in meta-analysis.

Belief in the law of small numbers is a misconception of statistics first discovered by Tversky and Kahneman (1971). In short, it means that in making intuitive judgements based on statistical evidence, people do not take sufficient account of the impact of sample size on the reliability of sample statistics. Small samples are believed to provide as reliable estimates of an average value as large samples. In informal research syntheses, belief in the law of small numbers involves assigning the same weight to all studies, irrespective of the sample size they are based on. Study results are tabulated and a simple average computed, disregarding both sample size and the quality of the studies.

In meta-analysis, it is possible to assign weights to studies that depend on sample size and estimate a weighted average. This means that studies based on small samples are given less weight than studies based on large samples.

The final source of error in informal research syntheses to be mentioned is *capitalisation on chance*. This means that random differences are treated as if they were real and explanations are offered for them. A case in point is a study by McGee and Blankenship concerning the safety effects of removing stop signs in intersections in three small towns in the United States (McGee and Blankenship, 1989). The objective of McGee and Blankenship's study was to develop guidelines for converting intersections from stop control to yield control. For this purpose, they broke down their data set according to several variables, finding, for example, that the largest increase in the number of accidents following conversion from stop to yield control occurred in intersections with large traffic volumes.

McGee and Blankenship's data came from the three small cities of Rapid City, Saginaw and Pueblo. In the converted junctions, the number of accidents increased from 12 before conversion to 26 after in Rapid City, from 25 to 68 in Saginaw, and from 4 to 12 in Pueblo. To account for changes expected without conversion, McGee and Blankenship compared the converted intersections to a "control group" of intersections that had even fewer accidents than the converted intersections. Based on these data, McGee and Blankenship concluded that "no statistically significant change was found for Pueblo and Rapid City, whereas a statistically significant increase was observed for Saginaw". In a re-analysis of these data, Hauer (1991) shows that there were no differences in the effect of conversion from stop to yield control between the three cities. McGee and Blankenship were, in effect, both capitalising on chance and succumbing to belief in the law of small numbers by testing for significance the observed changes in the number of accidents in each city separately. The correct method of determining whether the effects of conversion from stop to yield control differed between the

three cities, is to estimate an average effect for all three cities and then test if the effects in each city differ from the average effect by more than chance alone can explain.

In meta-analysis, capitalisation on chance can be avoided by determining the contributions of random and systematic variation to the variance found in a sample of results. Even within the framework of meta-analysis, there is, however, a small risk of capitalising on chance. This can occur when a very large number of variables have been coded for each study included in a meta-analysis and the effects of all these variables are tested as part of the analysis. Some of the tested variables may then turn out to be significant by chance. Using a conservative level of statistical significance when many tests are made will reduce the chances of erroneously interpreting a random effect as real.

8 Operational Criteria of Validity

8.1 Overview

This chapter proposes answers to the questions: What characterises good and bad evaluation studies? When is it defensible to pool the results of a set of evaluation studies in terms of a mean result, or a set of mean results, based on those studies? In what ways can meta-analysis help in answering these questions?

To help answer these questions, table 1 proposes a set of operational criteria of validity in evaluation studies. The criteria refer to four aspects of validity that will be elaborated in this chapter: Statistical conclusion validity, theoretical validity, internal validity and external validity. Some of the criteria of validity apply to each evaluation study, other criteria apply to a set of evaluation studies. Table 1 indicates for each criterion whether it applies to a single study or to a set of studies. To save space, the criteria are stated in short form in the table and will be discussed more in detail in the text. The letter S indicates statistical conclusion validity, the letter T indicates theoretical validity, the letter I indicates internal validity and the letter E indicates external validity. Table 1 contains nine criteria of statistical conclusion validity, four criteria of theoretical validity, four criteria of internal validity and three criteria of external validity. The criteria listed are not altogether independent of each other. Before discussing the relationship between the criteria, however, the meaning of each criterion and its applicability in meta-analysis will be discussed.

8.2 Statistical conclusion validity

Statistical conclusion validity, or simply statistical validity, is defined as the degree to which the numerical results of a study are accurate, reliable and representative of a known population. It includes reliability in the conventional sense of the term, i e the replicability of measurements made by means of a given technique or instrument in a given context. The level of statistical validity attained in an evaluation study, or in a synthesis of a set of evaluation studies, depends on a number of factors. The most important of these factors are listed in Table 1.

Sampling technique (S1) refers to the method used to select study units for inclusion in a study. The term study unit is generic and includes all types of study units, like individuals, physical objects or abstract objects. Based on sampling theory, a distinction can be made between three major sampling techniques. In descending order of validity, these include (1) random sampling or studies that include the whole theoretical population to which one wishes the findings to

apply, (2) systematic sampling according to specific criteria and (3) convenience samples (arbitrary samples) or self selected samples.

Table 1: Operational criteria of validity in evaluation studies

Criterion	Name of criterion	Scoring system	Level of use
S1	Sampling technique	3 = Whole population or random sample 2 = Systematic sample 1 = Convenience or self selected sample	Single study
S2	Sample size	Number of study units or statistical weights of study results	Single study
S3	Measurement reliability	3 = Known and high reliability 2 = Known, but low reliability 1 = Unknown reliability	Single study
S4	Systematic errors	3 = Complete and unbiased reporting 2 = Incomplete reporting; multiple sources of data used 1 = Incomplete and/or biased reporting	Single study
S5	Techniques of analysis	2 = Appropriate techniques used 1 = Inappropriate techniques used	Single study
S6	Dependent variables	3 = Commensurable across studies 2 = Incommensurable, can be converted to commensurable 1 = Incommensurable	Set of studies
S7	Publication bias	2 = No evidence of publication bias 1 = Evidence of publication bias	Set of studies
S8	Shape of distribution	3 = Distribution of results well behaved in terms of modality, skewness and outliers 2 = Distribution of results well behaved in terms of two the three properties 1 = Distribution of results well behaved in terms of one of the three properties	Set of studies
S9	Robustness of mean	2 = Mean result of a set of studies robust with respect to estimation techniques 1 = Mean result of a set of studies sensitive to estimation techniques	Set of studies
T1	Theoretical framework	3 = Explicit causal model and hypotheses formulated 2 = Explicit conceptual framework 1 = No explicit theoretical framework	Single study
T2	Operational concepts	3 = Key concepts operational 2 = Indirect measurements of key concepts 1 = Key concepts not measurable	Single study
T3	Mediating process	3 = Process mediating treatment effects known and measured 2 = Process mediating treatment effects inferred indirectly 1 = Process mediating treatment effects unknown or unspecified	Single study

Table 1: Operational criteria of validity in evaluation studies, continued

Criterion	Name of criterion	Scoring system	Level of use
T4	Support for theory	2 = Theoretical predictions supported 1 = Theoretical predictions rejected or not tested	Single study
I1	Direction of causality	2 = Causal direction clear within study design 1 = Causal direction not clear within study design	Single study
I2	Control of confounders	3 = All known confounders controlled 2 = Some known confounders controlled 1 = Few or no confounders controlled	Single study
I3	Dose-response pattern	2 = Dose-response pattern in relationship between cause and effect 1 = No dose-response pattern or no test of this	Single study
I4	Specificity of effect	2 = Effects found in target group only 1 = Effects dispersed in both target group and other groups	Single study
E1	Stability in time	2 = Results stable over time 1 = Results not stable over time	Set of studies
E2	Stability in space	2 = Results stable across space 1 = Results not stable across space	Set of studies
E3	Stability in contexts	2 = Results stable across contextual variables 1 = Results not stable across contextual variables	Set of studies

In Table 1, this ordering is shown by the numerical values assigned to the different sampling techniques. It has been assumed that an important objective of any evaluation study is to generalise the findings to a certain theoretical population of study units. This objective is, strictly speaking, only attainable when the sample was chosen from a known population by means of random sampling or some other sampling techniques whose properties are known.

In evaluation research, a sampling frame from which random sampling of study units can be made does not always exist. In that case, a systematic sample is often taken. In road safety evaluation studies, systematic samples have sometimes been used in studies that have evaluated the safety effects of traffic engineering measures.

Convenience samples or self selected samples are also common in road safety evaluation studies. It is impossible to know the population to which the findings of studies relying on such samples apply. Statistical tests of significance or estimates of confidence intervals are widely used in studies relying on convenience samples or self selected samples. The use of formal methods of statistical inference in these studies is perhaps best interpreted as an attempt to account for random variation in the data, not as a test of the generality of the findings in a known population.

In meta-analysis, the distinction made between different sampling techniques can be included as a coded variable in the analysis, provided studies describe sampling techniques in sufficient detail to determine which sampling techniques was used.

Sample size (S2) in general refers to the number of study units included in a study. Within the framework of meta-analysis, the term sample size may also denote the sum of statistical weights of study results. This indicator of sample size is relevant in meta-analyses in which the findings of a number of evaluation studies are synthesized in the form of a weighted mean result. In road safety evaluation studies, for example, the study units may be a sample of junctions where some kind of safety treatment has been carried out. The statistical accuracy of the results of the evaluation study depends, however, on the number of accidents recorded in these junctions, not on the number of junctions per se. In synthesising results from multiple junctions, it is therefore convenient to apply statistical weights that depend on the number of accidents in each junction. Sample size is, in both cases, a numerical variable which is subject to the law of large numbers. Hence, the larger the sample, the higher the statistical validity of the results of a study or a set of studies.

Measurement reliability (S3) denotes the replicability of measurements of a given variable made by a given method in a given context. Reliability is high when repeated measurements give identical or nearly identical results. Basically, the reliability of measurements depends on the amount of random variation in the variable that is being measured and on the accuracy of the method used. In accident research, the contribution of random variation is directly related to the number of accidents measurements are based on (Fridstrøm, Ifver, Ingebrigtsen, Kulmala and Thomsen, 1993; 1995). Random fluctuations will be relatively smaller around an expected number of accidents of, say, 100, than around an expected number of accicents of, say, 10. Hence, reliability in accident research depends directly on the size of the accident sample and can be estimated theoretically by relying on the generally accepted assumption that random variation in accident counts can be modelled by means of the Poisson distribution.

In evaluation research in general, however, reliability depends on the accuracy of measuring instruments and not just on the amount of random variation in the variable that is being measured. Instances of inaccurate measurement attributable to the measuring instruments are found in road safety evaluation studies as well, as shown, e g in the discussion of the accuracy of speed measurements in a report by Vaa (1995). Most laymen are likely to believe that it is easy to measure speed. This belief is unfounded. Readers who appreciate the careful discussion presented by Vaa may start wondering how common are the problems he discusses. In most reports, speed measurements are taken at face value and no discussion of their reliability is presented.

Although it is not always possible to determine the level of reliability numerically, a good evaluation study ought to contain a discussion of the problem. The scoring for reliability proposed in Table 1 is based on the assumptions that: (1) it is better to try to measure reliability than not to do so, and (2) if measured, it is better when reliability is found to be high than when it is found to be low.

Systematic errors (S4) refers to the presence of systematic measurement errors and biases in the data on which an evaluation study is based. Low reliability in a study is, by definition, caused by random errors and will not bias the findings, merely reduce their numerical accuracy. Systematic errors, on the other hand, may introduce systematic bias in a study – producing findings that are not just inaccurate, but simply wrong. Needless to say, every evaluation researcher wants to avoid systematic errors in a study. Notwithstanding this, however, systematic errors are likely to be endemic in road safety evaluation studies, due to the vagaries of the official road accident data that most such studies rely on as their major source of data.

Figure 1 traces the sources of error and loss of data in official accident records. Starting with all accidents that actually occur on public roads, the first loss of information occurs because some of these accidents are not defined as reportable to the police. In Norway, accidents that are not reportable include all accidents involving pedestrians only (no vehicles involved) and all accidents in which vehicles are involved, but only an ”inconsequential” (minor) personal injury is sustained (Elvik, Mysen and Vaa, 1997).

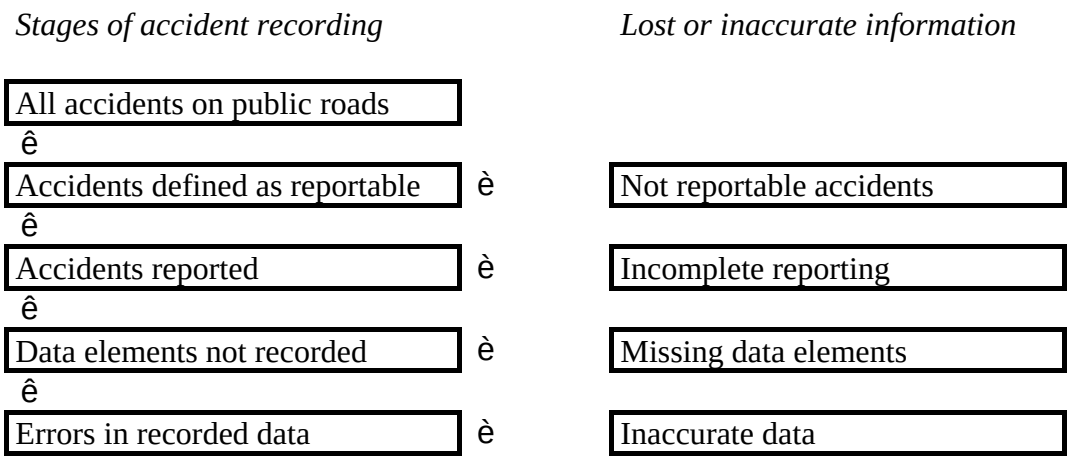


Figure 1: Sources of error and data loss in official accident records

It is well known from a large number of studies, summarised by Borger, Fosser, Ingebrigtsen and Sætermo (1995), that the reporting of injury accidents in official statistics is very incomplete. A large number of potentially important data elements, in particular related to human factors (Elvik and Vaa, 1990), are not recorded. Finally, there is bound to be errors or missing information in some of the recorded data elements.

In road safety evaluation studies that utilize detailed information from official accident records, these sources of systematic error are compounded. Yet, very few studies seem to have probed the implications of these, more or less inevitable, errors. The studies of Hakkert and Hauer (1988; Hauer, 1997), regarding the implications of incomplete and inaccurate accident reporting, are virtually the only studies that have tried to subject this problem to a rigorous analysis.

The problem of incomplete and inaccurate data recording in official statistics is by no means confined to road safety evaluation studies, but concerns evaluation research in general. It is well known that not all crimes are recorded by the police, that not all those of out work register as unemployed, that the gross national product does not include unpaid or ”black labour”, etc, etc. In general, the prevalence of social problems is nearly always underreported in official statistics. Unfortunately, official statistics tend to be the most important, and usually the most easily accessible, source of data in evaluation research. It is remarkable that the potential errors caused by this reliance on notoriously incomplete and inaccurate sources of data are as poorly understood as appears to be the case.

For the purpose of assessing the validity of evaluation studies, a distinction is proposed in Table 1 between studies that rely on complete and accurate reporting, which is in practice unlikely to be attainable, studies that use multiple sources of data in order to check the sensitivity of the results with respect to the source of data, and studies that rely on sources that are known to be subject to incomplete and biased reporting. This variable can be coded and included in a meta-analysis in order to test if study findings are indeed biased by the use of incomplete data sources.

The *choice of techniques of analysis* (S5) for analysing data refers to whether appropriate techniques of analysis for the data at hand have been used or not. This choice is not always strictly determined by statistical theory. Sometimes, more than one technique of analysis can be used. As far as road safety evaluation studies are concerned, it is important to recognise that: (1) Accidents, in particular if there are few of them, are not normally distributed. In large accident samples, however, the Poisson distribution, including generalized Poisson distributions like the negative binomial distribution, approach the normal distribution. (2) The homoskedasticity assumption for residuals in ordinary least squares linear regression (including logarithmic transformations or other models that are linear in parameters) is not correct when the dependent variable is a count of accidents. For accident counts, the amount of residual variance is proportional to the expectation, i e heteroskedastic. (3) The relationship between independent variables and the expected number of accidents is not always linear. Hence, an approach to multivariate modelling that allows different functional forms to be tested, e g by means of Box-Cox transformations, is called for. For a more extensive discussion of these points, the reader is referred to Fridstrøm et al (1993; 1995; see also Fridstrøm, 1998).

In the present context, the main point is that, at least as far as multivariate models based on accident data are concerned, it is possible to assess according to fairly straightforward criteria whether an appropriate technique of analysis has been chosen or not.

The lack of *commensurability of dependent variables* (S6) is a major problem in road safety evaluation research, as well as in evaluation research in general. Commensurability of dependent variables denotes the extent to which the dependent variables used in evaluation studies are identical in terms of their statistical properties and substantive interpretation. It is beyond the scope of this dissertation to discuss in detail the properties and legitimate interpretations of the various dependent variables that are used in evaluation studies. To give the reader an impression of the variety of definitions that exist, Table 2 lists some of the dependent variables commonly found in road safety evaluation studies. The list is not exhaustive.

Table 2: Commonly used dependent variables in road safety evaluation studies

Name of dependent variable	**Formal definition**
Simple odds	U_{at}/U_{bt}
Odds ratio (simple or adjusted)	$(U_{at}/U_{bt})/(U_{ac}/U_{bc})$
Ratio of odds ratios	$[(U_{ati}/U_{bti})/(U_{aci}/U_{bci})]/[(U_{atj}/U_{btj})/(U_{acj}/U_{bcj})]$
Ratio of relative risk	$[U_{ati}/(U_{ati} + U_{bti})]\{[U_{atj}/(U_{atj} + U_{btj})]$
Accident rate ratio	$(U_a/T_a)/(U_b/T_b)$

Notation:
U = number of accidents
T = traffic volume, exposure to risk
a = after, or with, some measure whose effect is evaluated
b = before, or without, some measure whose effect is evaluated
t = test group
c = comparison group
i = category i
j = category j

The definitions of dependent variables depend in part on study design, and therefore on how well the study has controlled for confounding factors. Hence, the interpretation of the various definitions of dependent variables is not merely a statistical problem, but is related to the confidence with which the effects of confounding factors can be ruled out as an interpretation of study findings.

The problems created by incommensurable definitions of dependent variables have been a major stumbling block in the development of meta-analysis. A way around the problem was eventually found by using so called effect sizes as the dependent variable in meta-analyses (Glass, McGaw and Smith, 1981). An effect size is, essentially, the difference in mean value of a certain variable between the test group and the comparison group, divided by the pooled standard deviation. It is the difference measured in number of standard deviations. Several versions of effect sizes have been developed (Rosenthal, 1994) and their statistical properties are today generally well known.

In road safety evaluation studies, the dependent variable is usually the number of accidents or some measure derived from the number of accidents (see Table 2). The different definitions listed in Table 2, however, cannot be pooled in terms of an effect size measure, but have to be treated separately. This, as indicated above,

is because not just the statistical properties, but the substantive interpretation of the various definitions differs.

As far as assessing study validity with respect to commensurability of dependent variables is concerned, a set of studies with commensurable definitions of dependent variables is regarded as more valid from a purely statistical point of view than a set of studies in which there are incommensurable definitions of dependent variables. This does not imply that some of the definitions listed in Table 2 are in general preferred to others.

Publication bias (S7) denotes the tendency not to publish studies whose findings are regarded as unwanted or without value. At least two types of publication bias have been identified: (1) Intolerance of null results, which means that results that are not statistically significant by conventional standards are discarded, and (2) Intolerance of negative results, which means that results that go in the opposite direction of what researchers or the sponsors of research expected or wanted are discarded. An extensive literature dealing with various aspects of publication bias now exists (Rosenthal, 1979; Peters and Ceci, 1982; Light and Pillemer, 1984; Coursol and Wagner, 1986; Begg and Berlin, 1988; Berlin, Begg and Louis, 1989; Dickersin and Min, 1993).

Light and Pillemer (1984) have proposed using inspection of funnel graph plots to test for publication bias. A funnel graph plot is a diagram in which the results of each study are plotted on the abscissa and the sample size each result is based on is plotted on the ordinate. The use of such plots is discussed more in detail in the next chapter. A funnel graph can, at best, give some indications of publication bias, but no hard evidence. Moreover, inspecting such a plot does not constitute a formal test. Hence, it cannot be claimed that there is publication bias on the basis of a funnel graph plot exclusively. Conversely, a funnel graph indicating no publication bias does not constitute evidence that no such bias exists, but it does weaken an argument to the effect that the published findings of evaluation studies are strongly influenced by publication bias.

Rosenthal (1979) has developed a test designed to estimate the number of unpublished studies with so called null results (i e no statistically significant effect) that have to exist in order to affect the mean result of a set of published studies. This test can be used to assess the sensitivity of published results to the potential presence of publication bias.

A good research synthesis applies funnel graphs or Rosenthal's test for the critical number of unpublished studies with null results in order to assess the possible presence of publication bias and discuss its implications. It has to recognized, however, that these tests are imperfect and do not constitute hard evidence.

The *shape of the distribution of results in a set of studies* (S8) refers to whether the distribution of results, as observed in, for example, a funnel graph diagram is unimodal and approximately normal or not. This criterion is related to the possibility of using weighted or unweighted mean results based on a set of studies in order to summarize the central tendency in the findings of those studies. Critics of quantitative research syntheses have claimed that such syntheses tend to mix "apples and oranges", i e to pool results that are substantively different and ought to be kept apart (see, e g, Bangert-Drowns, 1986, for a discussion).

It is obvious that a mean result located, for example, midway between two clearly discernible humps in a bimodal distribution would not be very informative. However, the strength of the"apples and oranges" argument can be assessed empirically. How to do so, is shown in paper 6 of the appended papers, to be discussed more in detail in the next chapter. It is argued that if the distribution of a set of results is well behaved in terms of modality (unimodal), skewness and sensitivity to outliers, then it is defensible and makes sense to summarize the central tendency of the distribution in terms of a weighted or unweighted mean result.

The *robustness of the mean result* of a set of studies (S9) refers to how sensitive the mean result based on a sample of studies is to the technique used to estimate it. Figure 2 gives an overview of the basic techniques that are applicable in quantitative syntheses of road safety evaluation studies. It is based in part on Hauer (1992).

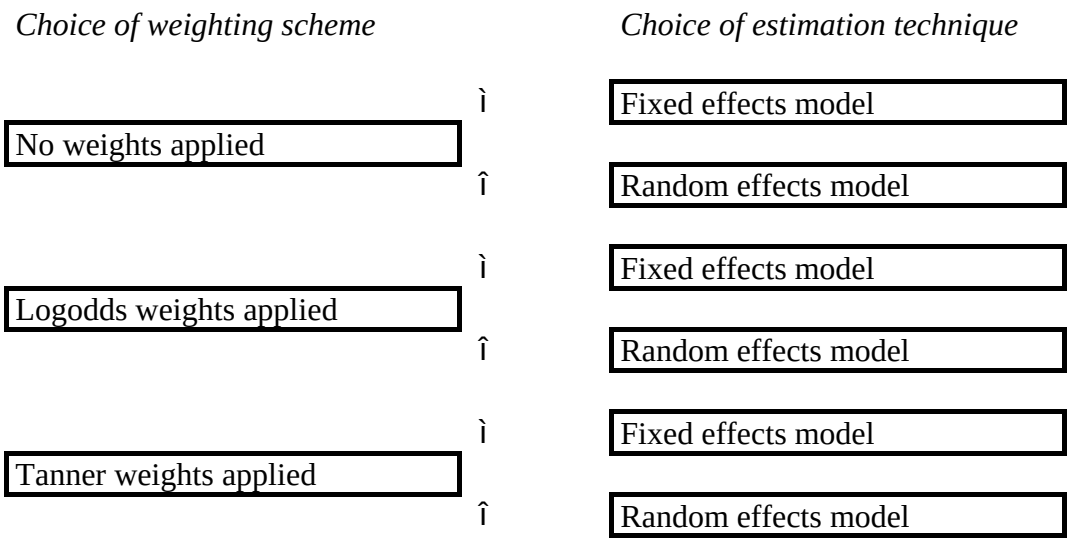


Figure 2: Taxonomy of techniques for estimating mean results in meta-analyses of road safety evaluation studies

A choice first has to be made regarding the weighting scheme to be applied. There are three main possibilities: (1) All results are assigned the same weight (i e an unweighted mean is estimated), (2) The logodds method of combining results is applied, and (3) The Tanner Chi-square technique for combining results is applied. Once the weighting scheme has been chosen, results should be tested for homogeneity in order to choose the right technique for estimating the mean result (Fleiss and Gross, 1991). The basic idea is that if there is significant heterogeneity of results (i e larger than random variations around the mean), a random effects model ought to be applied in estimating the mean result and the uncertainty of this result. If results are homogeneous, on the other hand, a fixed effects model can be used.

An extensive literature exists dealing with these choices and there is no consensus with respect to which model of analysis should be preferred (Tanner, 1958; DerSimonian and Laird, 1986; Kuritz, Landis and Koch, 1988; Berlin, Laird, Sacks and Chalmers, 1988, Griffin, 1989; Fleiss and Gross, 1991; Hauer, 1992; 1997; Shadish and Haddock, 1994). This means that, ideally speaking, a meta-analysis ought to apply all techniques and test the sensitivity of the mean result with respect to the choice of technique. If the estimated mean is the same no matter what technique is used, the choice of technique does not matter. If the estimated mean differs depending on which technique is used to estimate it, then the choice of technique needs to be discussed more in detail and justified in terms of the properties of the data set.

8.3 Theoretical validity

Theoretical validity is the degree to which a study or a set of studies relies on an explicit theoretical foundation that provides an explanation of study findings. The classic example of how theory can provide an explanation to the findings of a study is the Covering Law paradigm of natural science (Hempel, 1965):

E: The water in the radiator of my car is frozen

P1: Water freezes when the temperature drops below zero Celsius
P2: Last night, the temperature dropped below zero Celsius

C: That is why the water in the radiator of my car is frozen

This simple paradigm starts with the result that needs an explanation (E). The explanation consists of a statement of the Covering Law (P1) and the empirical observation made (P2), and is concluded by a statement showing how the two premises of the explanation explain the study finding (C).

It has been pointed out that the lack of an explicit theoretical basis is a major obstacle to cumulative transport research (Brehmer, 1993). An explicit theory, for example in the form of hypotheses set up in advance of an empirical study, is useful in many ways:

1 Theory tells the researcher what is important and what is unimportant, and thus guides the *selection of variables* to be included in a study. The alternative to relying on theory in this respect is to include in a study only those variables for which data happen to be available, or that have turned out to be statistically significant when tested in a preliminary analysis.
2 Theory gives support in designing the plan for collection and analysis of data in a study. It informs the researcher of the *appropriate study design*.

3 Theory gives *support when interpreting the results* of an empirical study. It tells the researcher what results make sense, by stating clearly the results the study is expected to produce. It is, however, appropriate to caution against relying too much on theory in interpreting the results of study, by dismissing all results that contradict the theory. Results that contradict a theory should be taken seriously if the study was appropriately designed.
4 Theory *makes research more cumulative*, by providing a unifying framework for synthesising the findings of multiple studies and integrating new findings with those of previous research.

For these reasons, it is desirable to develop an explicit theoretical foundation for evaluation research. A theoretical foundation for research can be more or less developed. A fully developed theoretical foundation for empirical research will:

1 Identify all relevant concepts and variables and specify how they can best be measured;
2 Sort relevant variables into the categories of independent variables, confounding variables, mediating variables, moderator variables and dependent variables;
3 Propose hypotheses describing the relationships between variables, including: (a) which variables that are related; (b) the direction of the relationship, (c) the strength of the relationship;
4 Identify the most important alternative hypotheses that may explain study findings if the proposed theory is contradicted.

Less well developed theories will not contain all these points. Four criteria of theoretical validity have been proposed. The first criterion, *T1*, refers to how well developed the *theoretical framework* for a study is in terms of the four points listed above. A crude distinction is made between three levels of development.

The second criterion of theoretical validity refers specifically to the use of theoretical concepts and to well *operationalised* these concepts are *(T2)*. The use of theoretical concepts is fruitful only to the extent that these concepts can be measured. Concepts that cannot be measured can only function as labels or heuristic devices in a theory, not as definitions of relevant variables.

The third criterion of theoretical validity (T3) is relevant for evaluation research specifically. It refers to whether a theory specifies the *process mediating effects* from the measure or programme that is evaluated to the dependent variable of interest. With respect to road safety evaluation studies, this usually involves specifying the risk factors for accidents a safety measure is intended to influence. The causal chain from a safety measure to a change in the number or severity of accidents goes through one or more risk factors the measure influences. The point of specifying these factors, and measuring them, is to assess the validity of causal inferences by checking the stages of the causal chain. Suppose, for example, that speed limits are reduced. The more a speed limit is reduced, the more one would expect speed to go down, and the more speed goes down, the more one would expect the number of accidents to go down. If such a pattern is found, it

strengthens a causal inference; if it is not found, it weakens inferring causality in the relationship between speed limit changes and changes in the number of accidents.

The fourth and final criterion of internal validity proposed concerns whether the proposed theory is *supported* or not (T4). Theoretical validity is higher when a theory is supported than when it is rejected.

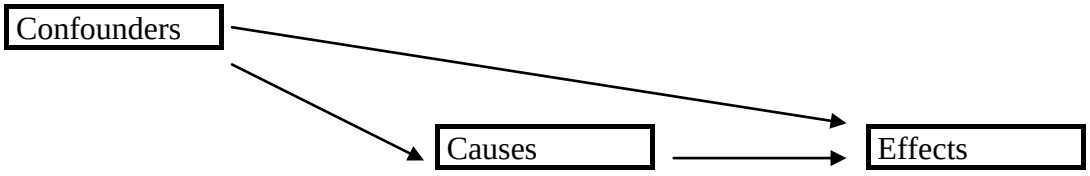
8.4 Internal validity

Internal validity denotes the extent to which a study or a set of studies fulfills the conditions for inferring a causal relationship between the measure or programme whose effects is evaluated and the dependent variable or variables of interest. The criteria of internal validity proposed in Table 1, are based on the following list of commonly accepted conditions for causal inference (Elvik, 1995C), gleaned from the literature (Blalock, 1961; Hill, 1965; Hellevik, 1977; Cook and Campbell, 1979; Elwood, 1988; Cordray, 1993):

1 *Statistical association*
There should be a statistically significant association between the causal variable and the effect variable. This condition is elaborated in points 3 and 4 below.

2 *Clear direction of causality*
It should be possible to determine the direction if causality between the variables subject to a causal relationship, that is whether A causes B or B causes A. The cause is generally assumed to precede the effect in time.

3 *No confounding*
The statistical association between cause and effect should persist when confounding variables are controlled. A confounding variable is any variable that is related to both the causal variable and the effect variable in a way that can either (a) give rise to an artifactual relationship between the causal variable and the effect variable, or (b) mask a true relationship between the causal variable and the effect variable. Confounding is illustrated below:



4 Known causal mechanism
The relationship between a causal variable and an effect variable should be explicable in terms of a known causal mechanism mediating the influence of the causal variable on the effect variable, or in terms of a theory stating why the variables are causally related. The specification of a causal mechanism is illustrated below:

5 Consistency across studies
The relationship between a causal variable and an effect variable should be consistent across studies and be reproduced in repeated studies made in different settings.

6 Dose-response pattern
The effects of the causal variable on the dependent variable should exhibit a dose-reponse pattern. A dose-response pattern is present when large changes in the causal variables are associated with large changes in the effect variable, and the converse.

7 Specificity of relationship
If there are reasons to believe that the relationship between a causal variable and an effect variable applies only to a specific subset of data, a causal inference is strengthened when the presumed specificity of the relationship is found, weakened when this specificity is not found.

The first five of these conditions are the most important, and are nearly always applied in assessing the causality of a relationship. Conditions six and seven may be applied if relevant, otherwise not. The presence of a dose-response pattern or a specificity in the relationship between cause and effect are not necessary conditions for inferring causality, but these conditions are useful when relevant.

From the list of conditions, one can see that in order to infer causality in the relationship between a pair of variables, that relationship should be both (1) Statistically valid, as indicated by condition 1, (2) Theoretically valid, as indicated by condition 4, and (3) Externally valid, as indicated by condition 5. Internal validity therefore partly overlaps the other types of validity; in fact one could say that a relationship between a putative cause and its effect cannot be internally valid unless it is also statistically, theoretically and externally valid.

The criteria of internal validity that are specific to this type of validity are those of conditions 2, 3, 6 and 7. Of these, conditions 2 (direction of causality) and 3 (control of confounding variables) are the most important. Based on the list of conditions for inferring causality, the following criteria of internal validity in evaluation studies have been developed.

Criterion *I1, direction of causality*, refers to the possibility of clearly inferring the direction of causality in a study. This possibility is related to study design. An experimental study, preferably one in which the dependent variable is measured both before and after treatment is introduced, provided the best basis for determining the direction of causality. In non-experimental studies, before-and-after studies are often believed to provide a better basis for inferring direction of causality than cross-section studies. Whether this is in fact the case depends to a large extent on how well a study controls for confounding factors. In a poorly controlled before-and-after study, the direction of causality may be less clear than in well controlled cross-section study. Sometimes, the direction of causality can be inferred from apriori reasoning. Thus, a possible causal relationship between driver gender and accident rates can only go in one direction.

Control of confounding factors (I2) is arguably the most important criterion of internal validity in evaluation research. Several factors make this criterion important: (1) Most of evaluation research uses non-experimental designs that do not guarantee control of all confounding factors; (2) The number of confounding factors that could bias the results of a study is, in principle, infinite; (3) Several studies have shown that lack of control of important confounding factors can seriously bias the results of evaluation studies (for illustrations, see examples given by Elvik, Mysen and Vaa 1997).

Control of confounding factors can be attained both in the design of a study and during the analysis stage of research. The best way of controlling for confounding factors – in fact *the only way* to control *all* confounding factors – is to use an experimental study design. In other study designs, control of confounding factors will be imperfect. However, this does not mean that all non-experimental studies are equally bad in this respect. Since the number of potentially confounding factors is in principle infinite, studies that control for a large number of confounding factors are better than studies that control for just a few or none at all.

On the other hand, it is in fact possible to control for ”too many” confounding factors. This can occur in two ways. The first one is when a variable is related to both the causal variable and the effect variable, but not in a way that confounds the relationship between them. Examples of such cases are given by Kleinbaum, Kupper and Morgenstern (1982). Another case of erroneous control of a confounding variable, is when a mediating variable, that is a variable which is causally influenced by the measure whose effects are evaluated and in turn influences the dependent variable is misconceived as a confounding variable. A case in point would be a study that controlled for changes in driving speed when estimating the effects of a speed limit change on the number of accidents. But a change in speed is likely to be a consequence of the change in speed limit, and is the mediating process through which this measure influences the number of accidents.

Both types of errors can be avoided by basing a study on an explicit causal model that identifies relevant confounding and mediating variables. Non-experimental studies in which the control of confounding variables is based on such a model should therefore be rated as better in terms of control of confounding fac-

tors than studies that base their control of confounding variables on whatever data happened to be available concerning potentially confounding variables.

The presence of a *dose-response pattern (I3)* can further strengthen causal inferences, provided the other conditions of causality are satisfied. In road safety evaluation studies, two kinds of dose-response patterns are conceivable. The first kind is based on the volume or standard of the safety measure that is being evaluated. Examples would be: ”The higher the standard of road lighting, the greater the reduction in nighttime accidents”, or: ”The greater the increase in police enforcement, the greater the reduction in the number of accidents”. The other kind of dose-response pattern is based on the relationship between a risk factor that is influenced by a safety measure and the number and/or severity of accidents. An example would be: ”The greater the reduction in driving speed, the greater the reduction in the number and severity of accidents”. It is not always possible to test for a dose-response pattern in the results of studies that have evaluated the effects of a measure or programme. Some measures are dichotomous and admit of no dose-response pattern: A car either has or has not high mounted stop lamps. However, even if the idea of a dose-response pattern does not make sense at a micro level (that is for each unit of observation in a study), it may still do so at an aggregate level: The higher the proportion of cars that have high mounted stop lamps, the greater becomes the decline in the number of rear-end collisions.

In some cases, the target group of a policy intervention is so clearly defined that it is possible to use the *specificity of an effect to the target group (I4)* as a criterion to support causal inferences. If changes in the expected direction of the dependent variable are found in the target group of the intervention only, that supports a causal inference. If similar changes in the dependent variable are found across the board, the basis for a causal inference is weakened. To illustrate the use of this criterion, consider a study by Broughton (1987) of a prohibition against using large motorcycles (defined as motorcycles with an engine displacement of more than 125 cubic centimetres) for drivers holding a learner’s permit. The observed changes in the number of accidents in this study are shown in Table 3.

It is seen that the largest percentage change in the number of accidents occurred in the target group of the intervention: learner drivers riding motorcycles with an engine displacement of more than 125 ccm. Moreover, the change observed in this group was in the expected direction of fewer accidents. There was an increase in the number of accidents involving learner drivers riding small motorcycles (less than 125 ccm), also expected because of a switch over from larger motorcycles. Only small changes in the number of accidents were observed among experienced motorcycle riders.

Table 3: Changes in the number of accidents following a prohibition against using motorcycles above 125 ccm for learner drivers. Based on Broughton, 1987

	Percent change in the number of accidents		
Groups of riders	Engine displacement	Best estimate	95% confidence limits
Learner drivers	Less than 125 ccm	+24	(+21; +29)
	125 ccm and above	-79	(-80; -77)
	All categories	+2	(-1; +5)
Experienced drivers	Less than 125 ccm	+7	(+2; +12)
	125 ccm and above	-16	(-18; -14)
	All categories	-10	(-13; -8)

This pattern in the results of the study agrees with what one would expect if the policy intervention affected the target group only, or at least had a greater effect within the target group than for other groups. It thus supports a causal inference.

8.5 External validity

External validity denotes the possibility of generalising the results of a set of studies to other contexts than those in which each of the studies in the set were made. The results of a set of studies display high external validity if reproduced to within random error in studies that were made in very different circumstances.

There are two main reasons why external validity is important in evaluation research. In the first place, the weak theoretical foundation of much of evaluation research means that few results can be ruled out on theoretical grounds. Confidence in the results of evaluation studies therefore depends in their having been reproduced in a large number of studies. In the second place, evaluation studies do not always rely on random sampling, but frequently employ convenience samples or self selected samples. Strictly speaking, conventional techniques of statistical inference cannot be used for such samples (because their sampling distribution is unknown). Generalisation of the results of evaluation research cannot rely on statistical testing exclusively, but in addition has to rely on a less formal inductive reasoning based on how often results have been reproduced in evaluation studies.

Three criteria of external validity have been proposed in Table 1. The first criterion concerns the *stability of results in time (E1).* Results that have been reproduced (i e are identical to within random error) in studies reported during a long period score higher for external validity than results that have not been reproduced for a long time. The second criterion concerns the *stability of results in space (E2).* Results that have been reproduced all over the world are more externally valid than results from a single country. The third criterion refers to the *context of a study (E3).* Results that have been reproduced across different study contexts are more externally valid than results that differ from one context to another. The term ”context” is, admittedly, rather vague. It denotes the external

circumstances in which a study was made, not aspects internal to the study. Elements of context for road safety evaluation studies might include the basic rules of the road in a country (like driving on the left versus driving on the right), the level of motorisation (number of cars per inhabitant), and the reporting rules for accidents (the exact definition of reportable accidents). The exact elements of the context that are regarded as relevant in assessing external validity will have to be specified on a case-by-case basis.

8.6 The relationship between types of validity

The four types of validity are not entirely independent and may partly overlap. Figure 3 is an attempt to depict visually the relationship between types of validity.

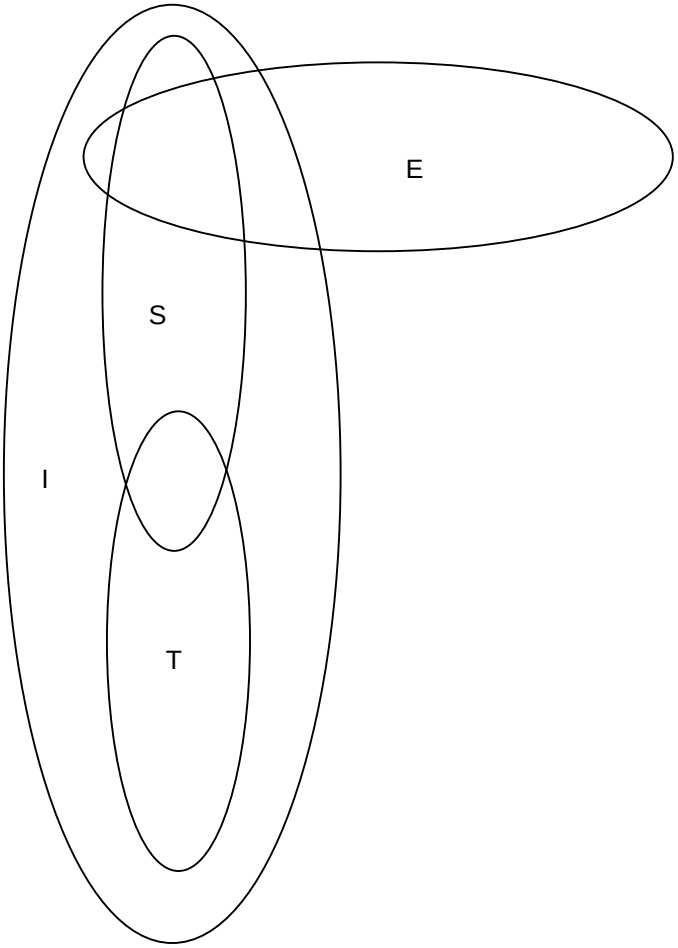


Figure 3: The relationship between types of validity. S = Statistical, T = Theoretical, I = Internal, E = External

There is some overlap between statistical and theoretical validity. Results cannot be theoretically valid without being statistically valid, at least with respect to some of the criteria of statistical validity. There are, on the other hand, aspects of both statistical and theoretical validity that do not overlap. For example, criteria T1 and T2 for theoretical validity do not overlap with statistical validity. Criterion S1 for statistical validity is not a necessary criterion of theoretical validity. Internal validity has been assumed to encompass both statistical and theoretical validity, and in addition partly overlap external validity. There are in addition some specific criteria of internal validity that do not overlap statistical and theoretical validity.

Which is the most basic type of validity? Can strength with respect to one type of validity partly compensate for weakness with respect to another? The importance of the various types of validity will differ depending on the topic for research and research objectives. In basic research in academic disciplines, theoretical validity has traditionally been regarded as very important. In evaluation research, statistical validity is likely to be the most important type of validity, closely followed by internal validity. Statistical validity is the most basic type of validity in empirical research. Results that do not make sense from a statistical point of view are meaningless from any other point of view as well. What can be made of results from research made in small convenience samples, with poor, error ridden data that failed to attain statistical significance? No substantive interpretation is possible for such research.

The following preliminary ranking of the importance of the four types of validity in evaluation research is proposed:

Type of validity	Points for importance
Statistical conclusion validity	4
Internal validity	3
External validity	2
Theoretical validity	1

Statistical conclusion validity is rated as most important, theoretical validity is rated as least important. This ranking reflects the current state of affairs, in particular in road safety evaluation studies. Ideally speaking, it is desirable to increase the importance of theoretical validity and reduce the importance of external validity by developing a more firm theoretical basis for evaluation research.

At present, however, it is necessary to require a high degree of external validity in evaluation research to compensate for the lack of theoretical validity. Results have to be reproduced over and over again before we can believe in them, because there is often no strong theory that informs us that these results must be correct.

9 Summary and Discussion of Appended Papers

Seven papers are appended. In order of appearance, these papers are:

1 The safety value of guardrails and crash cushions: A meta-analysis of evidence from evaluation studies (Elvik, 1995A)
2 A meta-analysis of evaluations of public lighting as an accident counter-measure (Elvik, 1995B)
3 Does prior knowledge help to predict how effective a measure will be? (Elvik, 1996A)
4 A meta-analysis of studies concerning the safety effects of daytime running lights on cars (Elvik, 1996B)
5 Evaluations of road accident blackspot treatment: A case of the Iron Law of evaluation studies? (Elvik, 1997)
6 Evaluating the statistical conclusion validity of weighted mean results in meta-analysis by analysing funnel graph diagrams (Elvik, 1998A)
7 Are road safety evaluation studies published in peer reviewed journals more valid than similar studies not published in peer reviewed journals? (Elvik, 1998B)

This chapter gives a summary and discussion of these papers on the basis of the system for assessing the validity of evaluation research presented in the previous chapters, especially chapter 8. The summary concentrates on how the validity of research has been assessed in these papers. The results of the evaluation studies as such will not be discussed.

The subject of *paper 1* (Elvik, 1995A) is the effects on safety of guardrails and crash cushions. The main focus of the paper is on the substantive issue of how installing guardrails and crash cushion affects road safety. However, research problem 3 as formulated in the paper (Can the evidence from evaluation studies be trusted?) concentrates on the validity of the evaluation studies that have been made with respect to guardrails and crash cushions.

The paper contains a fairly detailed classification of studies with respect to study design and confounding variables controlled. This classification is intended as a basis for assessing studies in terms of internal validity. The paper notes that three conditions should be met for a weighted mean estimate of safety effect based on a number of studies to make sense: (1) There should not be publication bias in the sample of results, (2) The distribution of the individual results around

the weighted mean should be ”well behaved”, and (3) The studies should use identically defined, or at least commensurable, measures of effect.

Six funnel graph diagrammes are presented in the paper in order to test for the possible presence of publication bias. In addition to indicating the possible presence of publication bias, these diagrammes show the modality of the distribution of results, i e whether the results are unimodal, bimodal, multimodal or lack any distinctive mode at all. In general the funnel graphs give no clear indication of publication bias. Some of the funnel graphs are based on rather few data points. No guidelines have been found in the literature concerning the smallest number of data points for which it makes sense to prepare a funnel graph. However, as a rule of thumb, it will in most cases probably be difficult to find a meaningful pattern in graphs based on less than ten data points. Funnel graphs based on less than ten data points are unlikely to provide much useful information.

In two of the funnel graph diagrammes presented in paper 1 (figures 7 and 8), the modal data point (the uppermost data point in the figure, based on the largest statistical weight or sample size) is located to the left of the majority of data points. This means that the modal data point in these graphs is not very representative of the typical result of the studies represented in these funnel graphs. As noted in the paper, these data points contribute more to the statistical weights than any other data points and will therefore unduly influence the weighted mean estimate of effect. The weighted mean estimate of effect will be inflated by these highly atypical modal data points and not be representative of the typical result of an evaluation study.

An approach to this problem, not pursued in *paper 1*, but introduced in *paper 6*, is to define outlying data points in terms of their effects on the weighted mean. An outlying data point is defined as any data point whose exclusion significantly affects the weighted mean. While arbitrary, in the sense that the choice of the level of statistical significance used to assess whether a data point is outlying is a matter of convention rather than analysis, an attractive feature of this definition is that it implicitly accounts for the effects of varying statistical weights on the probability of classifying a data point as outlying. Extreme data points in the tails of a funnel graph are unlikely to be classified as outlying, because they tend to be based on small samples (small statistical weights) and contribute little to the weighted mean.

Figure 7 in *paper 1* shows the results of studies that have evaluated the effects of crash cushions on the odds of sustaining a fatal injury. Ten data points are included in the Figure. A reanalysis of these data, applying the technique introduced in *paper 6* of omitting one data point at a time and estimating the weighted mean based on the remaining n – 1 data points, shows that the modal data point in Figure 7 is not an outlying data point. Its inclusion does nevertheless substantially affect the mean. If included, the weighted mean effect of crash cushions is a 69% reduction in the odds of sustaining a fatal injury. If omitted, the weighted mean effect is reduced to a 54% reduction in the odds of sustaining a fatal injury. The difference between these estimates of the mean effect of crash cushions is, however, not statistically significant.

Paper 1 applies a fixed effects model of meta-analysis. It does not discuss the choice between a fixed effects model and a random effects model. The choice of a fixed effects model can be defended on the grounds that it is a much simpler technique of analysis than a random effects model and that the extensive partitioning of the results into subsets in *paper 1* probably takes account of the effects of most factors that are likely to generate a systematic variation in the effects of guardrails and crash cushions. In *paper 1*, factors contributing to variation in the effects of guardrails and crash cushions are analysed by means of a simple one way analysis of variance. This analysis is carried out in two stages. The first stage is to determine the amount of variation in a set of results. This is done by estimating the coefficient of variation. The second stage of analysis consists of determining the relative contributions of random and systematic variation to the variance in a sample of results.

The approach adopted in *paper 1* relying on analysis of variance has not been applied in subsequent papers. The Chi-square technique of Fleiss (1981) and others is more appropriate for the logodds method of meta-analysis than conventional analysis of variance. This technique for decomposing the variance in a sample of results into random and systematic variation is explained in detail in *paper 6*, which shows a case illustration of the technique. Still, the main findings of the analysis of variance presented in paper 1 are valid and identifies those subsets of the data for which the contribution of systematic variation in study findings is greatest.

Table 1 in chapter 8 lists criteria of validity for evaluation research. The criteria in terms of which studies that have evaluated the safety effects guardrails and crash cushions are assessed formally or informally in *paper 1* include:

S2, sample size, which is shown in each of the funnel graphs and serves as basis for defining the statistical weight of each result included in the meta-analysis;
S6, dependent variable definition, which is discussed in the text as regards the appropriateness of using the odds ratio, defined in terms of levels of injury severity, as a measure of the effect of guardrails and crash cushions on injury severity;
S7, publication bias, which is addressed on the basis of the funnel graph diagrammes;
S8, shape of distribution of results, which is discussed informally on the basis of the funnel graph diagrammes (in terms of skewness and possible outlier bias);
I2, control of confounders, which is tested in terms of the sensitivity of results with respect to study design and control of specific confounding variables;
E1, stability in time, by showing how the results of evaluation studies vary by decade of study publication.

In addition to these criteria, the data assembled for *paper 1* allows a test to be made of a dose-response relationship with respect to the effects of guardrails (criterion I3 in Table 1). More specifically, such a test can be made for median guardrails on divided highways. Three types of guardrails have been studied: (1) Concrete median barriers, that are stiff and unyielding, (2) Steel beam guardrails, that yield upon impact, and (3) Wire guardrails, that yield even more when struck by a motor vehicle than steel guardrails. The more yielding a guardrail is, the more it ”prolongs” a crash by absorbing kinetic energy. The slower the process of absorbing kinetic energy, or transforming it to vehicle deformation, the less likely car occupants are to sustain injury. Hence, one would expect a softer guardrail to reduce the likelihood of injury, especially severe injury, more than a stiff guardrail. Inspection of the results obtained in evaluation studies confirms that this is indeed the case.

To illustrate the logic of this test of a dose-response pattern, consider Figure 4, which is based on (unpublished) data collected for *paper 1*. The figure shows the weighted mean effects of three types of median guardrails on the odds of sustaining a fatal injury or any personal injury, given a crash.

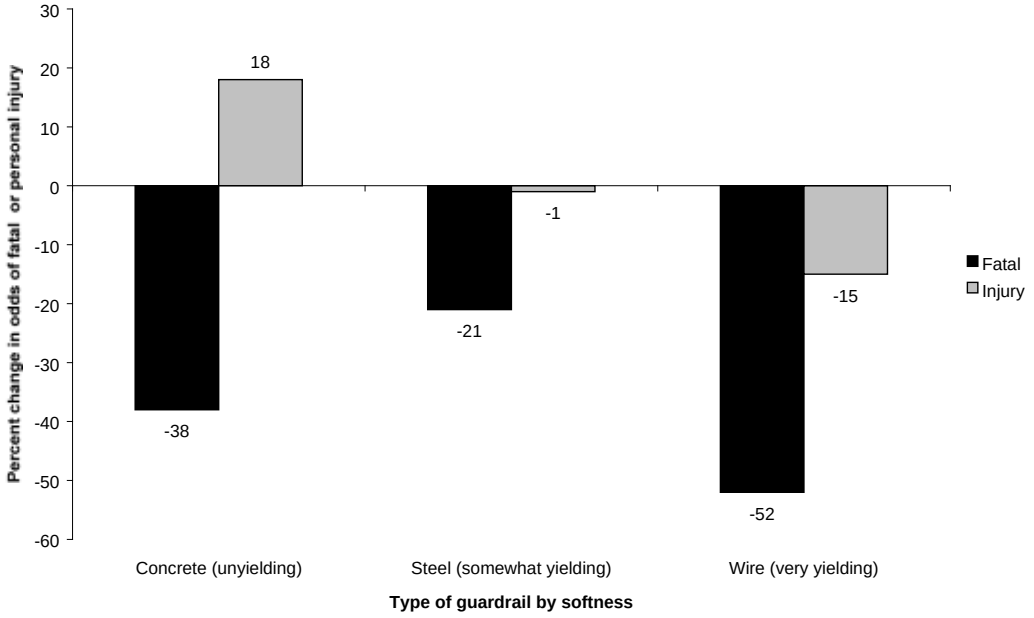


Figure 4: Dose-response pattern in effects of median guardrails

The presence of a dose-response pattern in the effects of median guardrails can be inferred from the following observations: (1) The effect of guardrails is greater for fatal injuries than for personal injuries in general. This tendency is consistent with a dose-response pattern, because an energy absorbing structure like a guardrail will often absorb a sufficient amount of energy to make the crash survivable, but not enough to make it harmless. (2) The effects of guardrails on the probability of sustaining injury increases as the guardrails become more yielding. This pattern is

particularly clear for personal injury of any severity, but a similar tendency, albeit less consistent, is found for fatal injury as well.

Paper 2 (Elvik, 1995B) presents a meta-analysis of studies that have evaluated the safety effects of road lighting. The format of this paper is very similar to *paper 1*. It takes as its starting point previous criticism that has been made of the validity of studies that have evaluated the effects of road lighting. The approach taken to testing the validity of these evaluation studies is essentially the same as in *paper 1*. There are, however, some differences between paper 1 and 2.

Paper 2 contains a brief discussion of various techniques of meta-analysis. Three techniques are compared: (1) A simple vote counting method, (2) A more sophisticated version of the vote counting method, in which account is taken of the statistical significance of results and (3) Methods that estimate a weighted or unweighted mean result based on a sample of evaluation studies. It is argued that methods belonging to group 3 are the most informative. A simple vote count merely tells us the direction in which the majority of results go (increase or decrease). The refined vote counting method in addition informs us of the reliability of the tendency, in terms of the proportion of results that are statistically significant. A meta-analysis in which a mean result is estimated informs us about the size of an effect, not just its direction. Moreover, if the mean result estimated is weighted by the sample size on which each result is based, it will account for the varying levels of statistical reliability of the results that are the basis of the weighted mean. Based on these distinctions, various methods of meta-analysis can be placed on a continuum with respect to the information they provide:

Method of meta-analysis	Simple vote counting method	Refined vote counting method	Methods estimating unweighted mean result	Methods estimating weighted mean result
	←	—	—	→
Information about effects	Direction	Direction Reliability	Direction Size	Direction Size Reliability

The conditions listed in *paper 2* for a weighted mean estimate of effect to make sense are the same as those listed in paper 1 and discussed above. The section of the paper specifically devoted to testing the validity of studies that have evaluated the safety effects of road lighting discusses regression-to-the-mean, secular trends in accident occurrence and the effects of contextual variables as threats to the validity of these studies. Regression-to-the-mean and secular trends are threats to internal validity. The effects of contextual variables, most of which would be termed moderator variables according to the classification of variables introduced in *paper 7*, mainly determines the external validity of the results.

In general the results of studies that have evaluated the effects on road safety of providing road lighting are found to be very robust with respect to the various threats to validity that are examined. In short, this means that the research that has

been performed to find the effects this safety measure is of rather high validity. Summarising the aspects of validity assessed in *paper 2* by reference to Table 1, the following criteria of validity are highlighted in *paper 2*:

S2, sample size, by the weighting scheme used in the meta-analysis;
S6, dependent variable definition, by comparing results defined in terms of the number of accidents and results defined in terms of accident rates;
S7, publication bias, by the use of funnel graph diagrammes;
S8, shape of distribution of results, as can be assessed informally by inspecting the funnel graphs;
I2, control of confounders, by studying how the results of evaluation studies vary according to study design and the control of specific confounding variables;
E1, stability in time, by examining how study results vary depending on decade of publication;
E2, stability in space, by examining how study results vary between countries;
E3, stability in contexts, by examining, for example, how study results vary according to the type of traffic environment where road lighting was installed.

The emphasis put on examining the external validity of studies that have evaluated the effects of road lighting may perhaps seem out of place. Surely, road lighting is an example of a measure for which one would expect the results of reasonably well designed studies to be nearly the same everywhere. Darkness makes it more difficult to see – for everybody all over the world. Road lighting improves visibility at night, which in turn ought to make it easier to avoid accidents.

This line of reasoning is, however, too simple. It is true that road lighting, or at least high quality road lighting, improves visibility at night. Hundreds of studies have been made to determine how various types of road lighting affect visibility and how changes in visibility influences the ability of road users to detect and identify other road users or obstacles on the road (Ketvirtis, 1977). Based on these studies, one would expect reasonably good road lighting to improve road safety. But a hypothesis merely stating that: ”Road lighting can be expected to improve road safety at night” is almost worthless as a theoretical basis for evaluation studies designed to measure the effects of road lighting on safety. Theory is useful as a basis for evaluation research to the extent that it:

1 Makes it possible to rule out certain results, or at least render them highly unlikely,
2 Identifies relevant confounding variables and provides guidance with respect to how best to control for them,
3 Identifies important moderator variables, thus defining a systematic pattern to which results can be expected to conform,

4 Identifies the causal chain through which effects are mediated from an intervention on through one or more mediator variables to the dependent variable of interest.

The hypotheses about the effects on road safety of road lighting that can be derived on the basis of engineering studies that have established the effects of road lighting on factors like luminance levels, subjective rating of visibility or detection distances to specific objects hardly satisfy these requirements. The main reason why the technical studies do not give a satisfactory basis for theory formulation, is that they fail to address the effects of a very important class of variables that partly determines the effects of virtually all road safety measures: Human behavioural adaptation.

When road lighting is installed, a number of changes in road user behaviour may occur. The amount of travel at night may increase, because some people who found it too strenuous or uncomfortable to travel in the dark when roads were unlit will now find the effort worthwhile. The speed of travel may increase, as road users find it easier to see the alignment of the road and objects in it. The level of effort and attention exerted by road users may, perhaps unconsciously and imperceptibly, go down as road users feel that they do not have to make as much effort to see the road and other road users as they had to when the road was unlit. Figure 5 shows a causal chain incorporating these mediating variables.

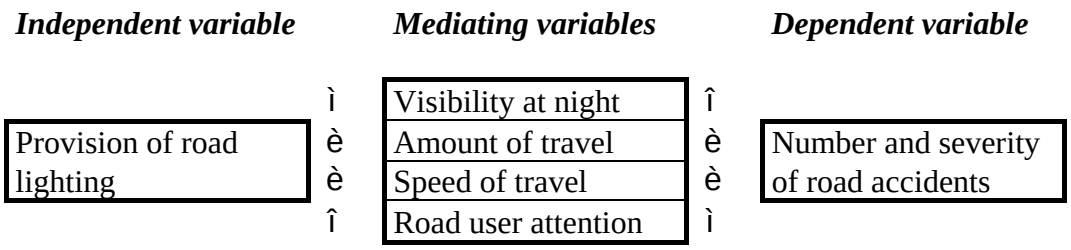


Figure 5: Causal chain for effects of road lighting on the number and severity of road accidents

In a study of behavioural adaptation to road lighting, Bjørnskau and Fosser (1996) have shown that all the three forms of behavioural adaptation listed in Figure 5 occur. It follows that the size and direction of changes in the number and severity of accidents following the provision of road lighting depends on the relative strengths of the effects represented by the various arrows in the model of the causal chain in Figure 5. It is impossible to rule out on theoretical grounds an increase in the number of accidents if, for example, road lighting is of poor quality, while at the same time there is a large increase in nighttime travel, speed goes up and road users pay less attention to traffic.

Although the case of road lighting may at first look like a promising subject for developing a strong theoretical foundation for evaluation studies, in the form of precise hypotheses about the effects of road lighting, based on physics, optical theory and the results of technical experiments, the fact that human behaviour

cannot be taken for granted complicates matter enormously. To predict theoretically the safety effects of road lighting, one would have to predict human behavioural adaptation to it. At the current state of knowledge, such prediction is impossible. Since most technical interventions can be expected to affect human behaviour one way or another, it follows that it is in most cases very difficult to develop a strong theoretical foundation for evaluation research.

Paper 3 (Elvik, 1996A) in a way takes this point of view as a starting point for developing a method for assessing the predictive validity of evaluation studies. By predictive validity is meant the accuracy of predictions of the effects of future applications of a measure based on the results of evaluation studies currently available. Since the effects of future applications of a measure can only be known from evaluation studies, predicting the future effects of a measure is tantamount to predicting the results of future evaluation studies. To assess the predictive validity of evaluation studies is therefore the same as to assess the stability over time of the results of such studies, which is an aspect of their external validity.

Paper 3 introduces a simple approach to testing the predictive validity of evaluation studies. It involves partitioning the evidence from evaluation studies, arranged in chronological order, into fractiles and using the results from an ”early” fractile as a prediction of the results of a subsequent fractile. In *paper 3*, studies are divided into quintiles, based on their statistical weights as a measure of the amount of evidence they provide. The first 20% of evidence accumulated is then used to predict the results of studies representing the next 20% of evidence. In the next stage of analysis, the first 40% of evidence accumulated (in chronological order), is used to predict the next 20%, and so on, until the first 80% of evidence is used to predict the results of the most recent 20% of evidence from evaluation studies. This approach makes it possible to test whether increasing the amount of evidence – that is doing more research – leads to more correct predictions of the effects of a measure. If doing more researchs leads to better predictions, then predictions based on 80% of the evidence currently available should be more accurate than predictions based on the first 20% of the evidence currently available.

According to the analysis in *paper 3*, predictive validity is not guaranteed, but depends on a number of factors as modelled in Figure 6. Some of these factors are assumed to enhance predictive validity, other factors are assumed to reduce it. The actual level of predictive validity depends on the strengths of the effects of the various factors influencing it.

The model presented in Figure 6 can interpreted as a list of some factors that affect the external validity of evaluation research, particularly road safety evaluation studies. High external validity can only be established by doing extensive research over a long period of time in highly different settings. The absence of a strong theoretical foundation for evaluation research means that findings of high generality can only be established by being reproduced a large number of times in highly heterogeneous studies.

A finding which has been replicated in many studies is, ceteris paribus, less likely to be an artifact attributable to poor data or inadequate research design than a finding reported by a single study only. Yet, it is not always the case that doing more research leads to clearer findings. Contradictory findings are common in evaluation research and may lead to confusion rather than clarity. The fact that the research designs employed tend to differ from one study to another compounds the problem of resolving contradictory findings.

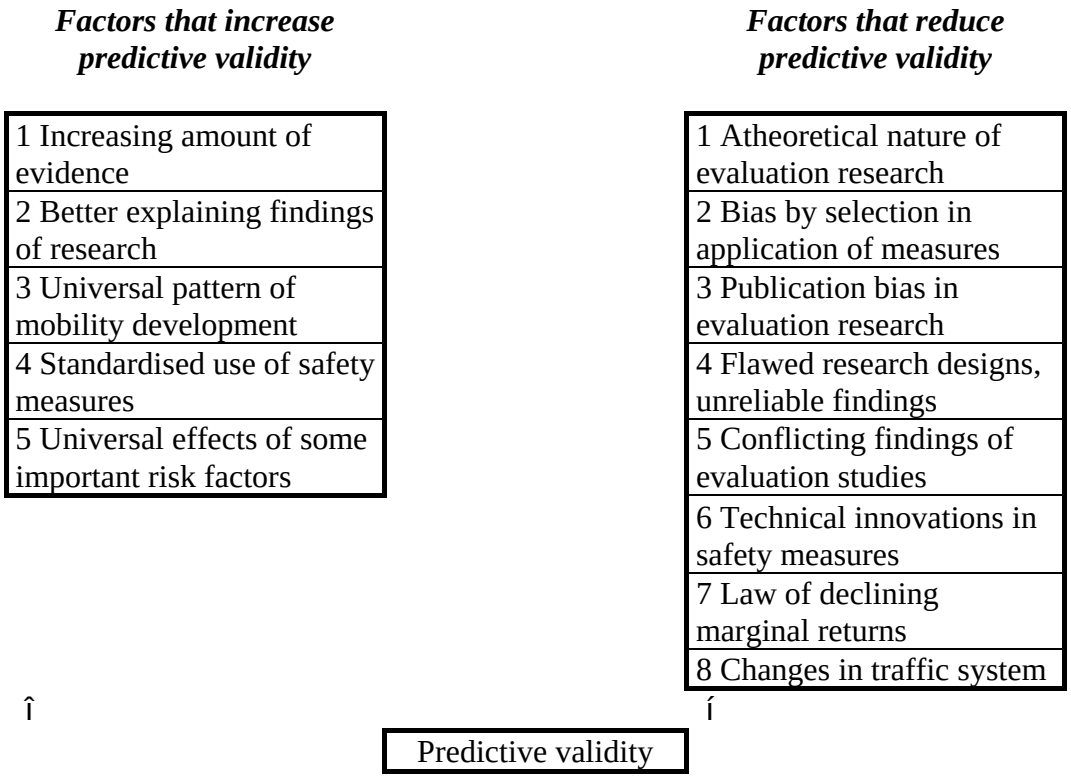


Figure 6: Factors affecting the predictive validity of evaluation studies

Paper 3 shows that doing more research does not necessarily improve the predictive performance of evaluation studies and explains why it is a logical fallacy to believe so. Predictions can be very erroneous and the prospect of explaining why they are so rather poor. In terms of the criteria of validity listed in Table 1, *paper 3* focusses on external validity exclusively, that is on the criteria E1 through E3, with the main emphasis on E1, stability in time of the results of evaluation research.

Paper 4 (Elvik, 1996B) contains a meta-analysis of studies that have evaluated the effects on road safety of daytime running lights on cars. This paper is in many ways similar to *papers 1* and *2*, but assesses other aspects of validity than those papers. Evaluations of daytime running lights have been very controversial. The controversy has focussed on methodological issues.

One of these issues concerns the use of the odds ratio as a measure of the effect of daytime running lights. *Paper 4* compares the odds ratio to two other definitions of the effect of daytime running lights on the number of accidents (the accident rate and the simple odds) and finds that they give broadly speaking the same results. The evaluation studies are, in other words, robust with respect to the definition of the dependent variable used in those studies (criterion S6 in Table 1).

The possible presence of publication bias is assessed by means of a funnel graph diagramme. Studies that have evaluated the effects of daytime running lights are classified in terms of study design. It is found that the results of the studies are very robust with respect to study design. This implies that, at least in evaluations of the intrinsic effects of daytime running lights (the effects for each car using daytime running lights), the influence of uncontrolled confounding factors is rather small. If confounding factors had a major influence on the results of evaluation studies, then studies with a poor control of confounding factors (non-experimental studies with no comparison group) would be expected to obtain different results from studies with a good control of confounding factors (experimental studies).

It is likely, however, that uncontrolled confounding factors have affected the results of studies that have evaluated the aggregate effects of daytime running lights (the effects of laws or campaigns designed to increase the use of daytime running lights). The results of these studies fail to show a dose-response pattern, that is there is no clear relationship between the size of the effect attributed to daytime running lights and the size of the increase in the use of daytime running lights upon the introduction of law requiring their use. There is, however, consistency between the results referring to intrinsic effects and the results referring to aggregate effects as far as the direction and size of the effect attributed to daytime running lights is concerned.

The paper tests the relationship between the intrinsic effects of daytime running lights and the latitude of the country in which effects were studied. This test can perhaps be interpreted as a test of a theoretical prediction (hypothesis), based on how the effects of daytime running lights on vehicle conspicuity vary in different conditions of ambient illumination. The ”latitude hypothesis” gets some support from the data, indicating that there is a systematic pattern in the effects attributed to daytime running lights in evaluation studies. If these effects were entirely caused by statistical artifacts or uncontrolled confounding factors, one would not expect to find this pattern.

Paper 4 is the first of the papers discussed so far that comments on a possible source of bias in meta-analyses, arising from the possibility of including retrieved evaluation studies in a meta-analysis. Four studies that had evaluated the effects of daytime running lights were retrieved, but could not be included in the meta-analysis because they did not report the number of accidents the stated effects were based on. *Paper 4* compares the results of these studies to the results of the studies that were included in the meta-analysis. The results are quite similar, indicating that the omission of the four studies not reporting the number of accidents did not seriously bias the results of the meta-analysis.

The possibility of a study inclusion bias in meta-analysis cannot be ruled out in general, however. A paper by Wagenaar, Zobeck, Williams and Hingson (1995), presenting a meta-analysis of programmes designed to reduce drinking and driving shows that if study inclusion criteria are strict, the large majority of retrieved studies may have to be omitted from a meta-analysis. Figure 7 has been drawn on the basis of Wagenaar et als study.

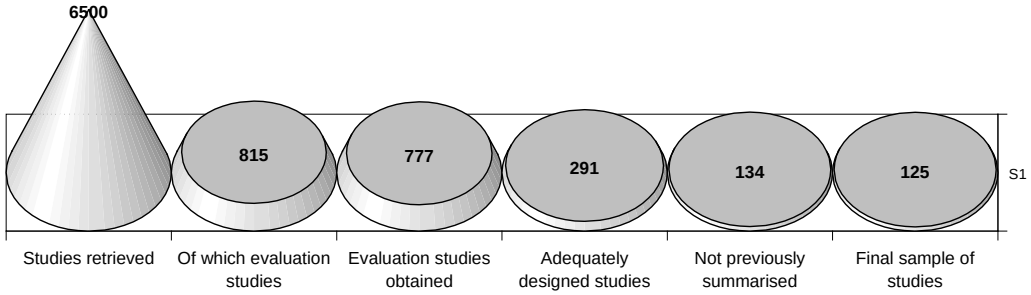


Figure 7: Successive stages of study exclusion in meta-analysis of measures to control drinking and driving. Adapted from Wagenaar et al 1995

A literature search identified 6,500 studies dealing with the subject of drinking and driving. Only 815 of these, however, were evaluation studies. Efforts were made to obtain these studies, but only 777 were obtained. These 777 studies were then screened on the basis of three criteria for methodological quality. Only 291 studies passed this screening. 157 of these were omitted because they were judged to be too old or had been summarised previously. This left 134 studies for analysis, of which 9 were omitted because they used very atypical research designs. This left 125 studies for inclusion in the meta-analysis. When the pruning of studies is as drastic as it was in this case, one may wonder about the representativeness of the studies that were included in the meta-analysis.

Summarising *paper 4* with regard to the criteria of validity assessed in the paper (cf Table 1), the following criteria were emphasised:

S6, dependent variable definition, by comparing study results according to three different definitions of the variable intended to measure the safety effects of daytime running lights;
S7, publication bias, by examining a funnel graph diagramme;

T4, support for theory, by testing the hypothesis about a relationship between the latitude of a country and the effects of daytime running lights;
I2, control for confounders, by comparing study results for research designs embodying varying levels of control of confounding factors;
I3, dose-response pattern, by examining the relationship between the size of the increase in the use of daytime running lights when it is made mandatory and the size of the effect on accidents;
I4, specificity of effect, by discussing (in the text) whether the effect of daytime running lights is confined to multi party daytime accidents, as assumed in the odds ratio measure of effect;
E2, the stability of results in space, by comparing the results of evaluation studies reported in different countries.

Paper 5 (Elvik, 1997) is a case study of the so called Iron Law of evaluation studies, applied to studies that have evaluated the effects on road safety of road accident blackspot treatment. This paper is perhaps the most iconoclastic of the seven appended papers. Proponents of blackspot treatment are likely to read the paper as a one sided and wholly destructive attack on a successful approach to road accident prevention.

The paper concentrates exclusively on criterion I2 of study validity, control for confounders. Four known confounders in non-experimental before-and-after studies are chosen for analysis. The study finds that the effects attributed to black-spot treatment decline to virtually zero as more and more of these confounders are controlled in evaluation studies. This finding supports the Iron Law of evaluation studies.

Paper 5 can serve as basis for a more general discussion of approaches to the control of confounding factors in evaluation studies. Based on a classification of methods for controlling for confounders developed by Elwood (1988, page 94), Figure 8 proposes a preliminary ranking of various methods for removing the effects of confounding variables in evaluation studies.

Stage of control	***Method of control***	***Rank***
Design of a study	Randomization	1
	Matched comparison group	2
	Non-matched comparison group	3
	Restriction of sample	6
Analysis of a study	Multivariate analysis	4
	Stratification	5
	Restriction of sample	7

Figure 8: Approaches to controlling for confounding in evaluation studies

Control for confounding variables can be introduced either in the design of a study or in the analysis of it, or at both stages of the research process. Controls that are introduced early in the research process are generally to be preferred to those that are introduced at later stages. Designing a study to control for confounding variables generally involves using a control or comparison group in addition to the test group that receives the treatment whose effects are evaluated. The best way of defining a control group is by randomization, that is by assigning subjects at random to either the treatment group (or groups) or the control group. Provided the groups are large, randomization ensures that there will be no systematic differences between them except with respect to exposure to the treatment that is evaluated.

Hauer (1997) has proposed using the term comparison group when the control group is not chosen at random, but selected on the basis certain criteria. A matched comparison group is often regarded as better than a non-matched comparison group. However, Hauer (1997) argues that the ranking of matched versus non-matched comparison groups with respect to how well they control for confounding factors depends on their size. A small matched comparison group may perform worse than a large non-matched comparison group.

Restriction of the sample is a procedure that can be applied both at the design and analysis stages of a study. One may control for sex, for example, by confining the study to women. Restriction must be rated as the poorest way of controlling for confounders, because it makes it impossible to generalise the results of a study beyond the restrictions imposed on it. This reduces the external validity of a study.

The second main approach to controlling for confounding variables is to collect data about these variables and measure their effects directly. This approach to controlling for confounding variables is applied at the data collection and analysis stages of a study. The best way of controlling for confounding in analysis, is to use a multivariate technique of analysis. Multivariate analysis allows for the simultaneous control of a large number of confounding variables. Stratifying a sample according to confounding variables rapidly depletes sample size and will therefore normally allow for the control of fewer confounding variables than a multivariate analysis.

Both multivariate analysis and stratification can be of varying quality, depending on how confounding variables are identified for analysis. The best way of identifying confounding variables is by relying on a theoretical model that explicitly identifies relevant confounding variables and models their effects. Another useful approach is to identify confounding variables statistically, as explained by Kleinbaum, Kupper and Morgenstern (1982). Identifying confounding variables statistically prevents the researcher from inadvertently controlling for variables that really are not confounding and need not be controlled, because they do not disturb the effects of the measure that is evaluated.

Paper 6 (Elvik, 1998A) is entirely methodological in its focus and is devoted to how one can assess the statistical conclusion validity of weighted mean results in meta-analysis by analysing funnel graphs and information derived from such

graphs. The paper presents a set of simple techniques that can be applied to assess the statistical conclusion validity of results in a meta-analysis. By doing so, the paper shows how one can use various diagnostic tools in meta-analysis in order to test how appropriate it is to generalise the results of studies included in a meta-analysis.

One of the most common objections to meta-analysis is that it generalises too much; it mixes ”apples and oranges” and estimates meaningless mean results that paste over crucial differences. This criticism is understandable, and fortunately it is possible within the framework of meta-analysis to test whether there is any merit to it. More specifically, *paper 6* shows how one can test for the following threats to the statistical conclusion validity of mean results in meta-analysis:

1 Heterogeneity (systematic variation) in a sample of results,
2 Skewness in a sample of results,
3 The modality of a distribution of results,
4 The sensitivity of the mean to outlying data points in a sample of results,
5 Publication bias in a sample of results,
6 The robustness of a weighted mean to the weighting scheme adopted,
7 The sensitivity of the standard error of the mean to the presence of correlated results in a sample of results.

These threats to statistical conclusion validity mostly refer to criteria S2 (sample size), S7 (publication bias), S8 (shape of distribution of results) and S9 (the robustness of the mean) in the list of criteria of validity given in Table 1.

Heterogeneity in a sample of results simply denotes the presence of systematic variation in effect sizes in the sample. The presence of systematic variation in a sample of results is, by itself, no decisive objection to estimating a weighted mean result based on the sample. It makes perfect sense to conclude that the mean temperature in June is higher than the mean temperature in January, despite the fact that the variation in daily temperatures in each month will no doubt be greater than randomness alone can account for.

If the contribution of systematic variation dominates the total variance in a sample of results, it is well advised to opt for a random effects model of meta-analysis. If, on the other hand, the contribution of systematic variation is minor, little is gained by using a random effects model of meta-analysis. It merely reduces the values of the statistical weights and complicates the analysis without affecting the estimated weighted mean greatly.

By testing in stages first for the presence of systematic variation in study findings, next for publication bias and finally for modality, skewness and possible outlier bias in the distribution of results, the techniques described in *paper 6* can function as diagnostic tools, or screening devices, with respect to the appropriateness of estimating a weighted mean result based on a sample of results. This function is very useful, since the sample of results retrieved for a meta-analysis can rarely be regarded as a random sample from a known sampling frame. Strictly speaking, standard statistical techniques for testing significance or estimating confidence intervals are based on the assumption that the sample was drawn at

random. This assumption is routinely disregarded in current empirical research, as one can quickly ascertain by opening any scientific journal. If, however, the distribution of results in a sample retrieved for meta-analysis is ”well behaved”, that is approximately normal, using standard techniques of statistical inference is perhaps a less serious violation of the assumptions underlying these techniques than if the sample of results is highly skewed and riddled with outliers.

The jackknifing technique described in *paper 6* for removing correlations between multiple results of the same study has not been widely applied in meta-analysis. As indicated in the paper, the idea of a correlation between multiple results of the same study makes sense only when certain assumptions are met; these assumptions are unlikely to be met for the data set used in *paper 6*. Some of the studies included in that data set produced multiple results, sure enough, but the idea of regarding these results as somehow correlated does not seem to make sense. At any rate no method was found to compute the correlation. Multiple results for the same variable can only be correlated if they: (1) represent successive observations in a time series, in which case the idea of an autocorrelation makes sense, or (2) are conceptually or computationally related to each other, like when result A is used to derive result B which in turn is used as input to derive result C.

It should be noted that sophisticated techniques based on linear algebra (Gleser and Olkin, 1994) have been developed in recent years for the treatment of what is generally referred to as ”stochastically dependent effect sizes” in meta-analysis. A comparison of these techniques to the jackknife technique has not been found, but would be very interesting.

The main research problem treated in *paper 7* is rather different from the problems discussed in the other six appended papers. *Paper 7* deals with factors that influence study quality, especially the peer review system of scientific journals. In order to answer the main question posed in *paper 7*, the paper also discusses how study quality can be measured and proposes seven criteria of study validity. These criteria are related to the following criteria of validity in Table 1:

S1, sampling technique, for which an ordinal variable is created;
S2, sample size, as measured by the statistical weight a study represents;
I2, control of confounders, indicated both by the code for research design and the explicit enumeration of relevant confounding variables that ought to be controlled;
I4, specificity of effect, indicated by the coding of moderating variables a study ought to specify;

It should be noted that the studies included in *paper 7* have been rated for validity in terms of methodological strengths and weaknesses only and with no regard to their results. The results are not even mentioned in the paper and are irrelevant in judging the validity of each study. Results are relevant, however, when it comes to judging the external validity of a set a studies, but only with respect to their variability, not their content.

Paper 7 discusses some hypotheses concerning factors that affect study quality. It is hypothesised, for example, that the ”publish or perish” system of universities provide researchers with incentives that lead to higher quality research. The results presented in the paper do not seem to give very strong support to this hypothesis, although the papers published in peer reviewed journals by university professors were rated slightly higher for validity than papers published by authors with other affiliations or not in peer reviewed journals. It is well known that the publish or perish system is despised by most people who are subject to it. The system may actually pervert the incentives to publish to such an extent that researchers churn out a heap of rubbish, and publish it in third rate journals in an attempt to beat the system. A determined author can get any rubbish published. It is almost always possible to find some obscure journal with a sufficiently lax review system to let through even very poor papers. The publish or perish system may lead to fierce competition among researchers, hampering their ability to cooperate and share new ideas with each other and thus, in the long run, slow down scientific progress.

Another hypothesis proposed in paper 7 is that research in traditional academic disciplines benefits from having a much stronger theoretical foundation than most of evaluation research. The trouble with evaluation research is that one can rarely rule out a result on theoretical grounds. On the other hand, the possibility that theory may outrun empirical research to such an extent as to become almost incapable of empirical testing should not be ruled out. A case in point is modern game theory. The most mathematically refined models of game theory seem to bear little relation to everyday life and can only be tested in laboratory simulations. There is simply no way of observing, for example, a repeated Prisoners' Dilemma game in a natural setting in sufficient detail to test hypotheses concerning the propensity to cooperate in the game. When observing human behaviour in a natural setting, one may not even know if the Prisoners' Dilemma is the right model of the interactions studied.

This does not mean that trying to establish a more firm theoretical foundation for evaluation research is futile or should not be encouraged. In most cases, however, one should not expect theory to predict more than the direction of an effect. Theoretical predictions of the size of an effect will, at least at the current stage of social theory, have to rely on rather strong assumptions whose validity cannot always be tested.

The confidence placed in the peer review system by both the scientific community and the general public is perhaps too high. A number of studies have revealed striking weaknesses of the peer review system of scientific journals. In a widely quoted study, Peters and Ceci (1982) resubmitted twelve papers published in prestigious psychology journals, using false names and affiliations (with consent from the original authors), but otherwise changing the papers as little as possible. Only three of the papers were found to be copies of previously published papers. The other nine went through a complete review process. Eight of these papers were rejected, only one accepted for publication.

Coursol and Wagner (1986) show a very great publication bias in studies reporting the outcomes of psychological counseling and psychotherapy. Coursol and Wagner divided papers into those showing a ”positive” outcome, that is an improvement in health state following counseling or therapy, and those showing ”no effect or a negative” outcome. Papers belonging to the former group were more likely than papers in the latter group both to be submitted to a journal, and, once submitted, to get published. 66% of papers showing a positive outcome were published, but only 22% of papers showing no effect or a negative outcome were published. The peer review process strengthened publication bias rather than reducing it. Other studies of publication bias include those of Begg and Berlin (1988) and Dickersin and Min (1993).

Hargens (1988) shows that journal rejection rates are closely related to scholarly consensus, that is to whether referees agree on the fate of a paper or not. He shows how editorial decisions with respect to publication can be predicted almost perfectly from a simple decision model using only referee recommendations as input. In a similar vein, Cullen and Macauley (1994) studied the relationship between agreement between referees about publication and editorial decisions concerning publication in the *Journal of Clinical Anesthesia*. Their study comprised 422 papers in total. Referee recommendations were coded as: (1) Accept as submitted, (2) Accept with revisions, (3) Reject in present form, and (4) Reject outright. They found that referees agreed perfectly in their recommendations for 169 papers. They differed by one category (for example between categories 2 and 3) for 168 papers, by two categories (like 1 versus 3) for 73 papers and by three categories (1 versus 4) for 12 papers. Disagreement among referees is, in other words, quite common. But the majority of papers that got mixed reviews were published, except in cases where one of the referees recommended outright rejection. Even 33% of the papers for which both referees recommended rejection in the present form were published. It seems that editors are more inclined than referees are to give authors the benefit of doubt. This means that many journals are likely to contain a quite a few papers that the majority of the readers of those journals will find worthless.

The unreliability of peer review has also been demonstrated in a study by Cicchetti (1991). Referees fail to detect even outright fraud in scientific papers (Rennie 1994). Rennie (1994) tells the story of Robert Slutsky, who during a period of seven years (1978-1985) published 137 scientific papers in medical journals. 48 of those papers were subsequently found to be of questionable validity, another 12 were found to be fraudulent. Before the fraud was exposed, all papers by Slutsky were cited at the same rate. Once fraud was exposed, however, the citation rate dropped by 67% for the fraudulent papers. But all these papers had been published and quoted in good faith.

In view of these studies, it should perhaps not come as a surprise that road safety evaluation studies published in peer reviewed journals do not score much higher for study quality than similar studies not published in peer reviewed journals. In addition to the failings of peer review, the incentives facing evaluation research in general are, as noted in *paper 7*, not conducive to high quality research. On a continuum going from pure market incentives on one end to pure intellectual curiosity for its own sake on the other, evaluation research is pretty close to the market end. As pointed out by Stephan (1996), knowledge is a public good and competitive markets generally provide poor incentives for the production of a public good. She claims, however, that science has developed a reward structure that overcomes this problem and provides incentives for scientists to behave in socially responsible ways. What stimulates intellectual curiosity, according to Stephan, is the recognition awarded by the scientific community to scientists who are the first to make a discovery or propose a new theory. This incentive can hardly be said to play an important part in evaluation research. Evaluation research concentrates on well-defined problems and often aims to add only a little to previous knowledge. It is not the arena for grand discoveries.

Table 4 summarises the criteria of validity that have been addressed explicitly and implicitly in the seven appended papers.

Table 4: Criteria of validity used to assess studies in meta-analysis. Based on Papers 1-7. Criteria used explicitly denoted by E, criteria used implicitly denoted by I. Criteria taken from Table 1

	Appended paper number						
Criteria of validity	**1**	**2**	**3**	**4**	**5**	**6**	**7**
S1 Sampling technique							E
S2 Sample size	E	E		I		E	E
S3 Measurement reliability							
S4 Systematic errors in data							
S5 Techniques of analysis							
S6 Commensurability of dependent variables	I	E		E			
S7 Publication bias	E	E		E		E	
S8 Shape of distribution of results	E	E				E	
S9 Robustness of mean						E	
T1 Explicit theoretical framework							
T2 Operationality of key concepts							
T3 Specification of mediating process							
T4 Support for theory				E			
I1 Unequivocal direction of causality							
I2 Control of confounding factors	E	E		E	E		E
I3 Dose-response pattern in results				E			
I4 Specificity of effect to target group				E			E
E1 Stability of results over time	E	E	E				
E2 Stability of results in space		E	E	E			
E3 Stability of results across study contexts		E	E				

Between them, the appended papers have assessed the validity of road safety evaluation studies in terms of all listed criteria of validity, except for:

S3, Measurement reliability;
S4, Systematic errors in data;
S5, Techniques of analysis;
T1, The presence of an explicit theoretical framework for a study;
T2, The operationality of key concepts;
T3, Specification of the mediating process between cause and effect;
I1, An unequivocal direction of causality.

These are seven out of the total of the twenty criteria of validity listed in Table 3 and previously discussed in Chapter 8.

As far as the statistical conclusion validity of evaluation studies is concerned, meta-analysis seem best suited to test aspects related to:

- The definition and commensurability of dependent variables,
- The possible presence of publication bias,
- The shape of the distribution of a sample of results, and
- The robustness of an estimated mean effect with respect to techniques of meta-analysis

These are all aspects of validity that have been extensively discussed in the meta-analysis literature. *Measurement reliability (S3)*, which is not explicitly discussed in any of the appended papers, is another aspect of validity that has received extensive attention in textbooks of meta-analysis. There exists, for example, a well-developed statistical theory specifying how various sources of unreliability affect correlation coefficients and how one can adjust the value of correlation coefficients for these sources of unreliability (see, for example, the instructive discussion in Hunter and Schmidt, 1990, part II). In principle, therefore, it is possible to assess measurement reliability within the framework of meta-analysis and rate studies according to this criterion.

In road safety evaluation studies, an important source of unreliability is, as mentioned before, random fluctuations in the number of accidents. In meta-analyses using the logodds method, this source of unreliability is accounted for in the estimation of the statistical weights of the results going into the meta-analysis. Results based on a small number of accidents are more unreliable than results based on a larger number of accidents, and are assigned a smaller statistical weight in meta-analyses using the logodds method.

Unreliability in the measurement of independent or mediating variables can also affect the results of a study. Unless it is possible to model statistically this kind of unreliability, it is rather difficult to assess it formally in a meta-analysis. To the extent that unreliability is related to sample size, it is always possible to account for it in meta-analysis. If unreliability is attributable to random variation (sampling variation) in the variable that is measured, it is related to sample size and will be less important in large samples than in small samples. If unreliability

is related to random errors of measurement (errors of coding, misreading an instrument, etc), it is not obvious that such errors will be less frequent in large samples than in small samples. To fully account for measurement errors, one would have to know their frequency and nature, which is rarely the case.

This point of view applies to *systematic errors in data (S4)* as well. As noted in chapter 8, most road safety evaluation studies rely on official accident statistics. It is known that official accident statistics is subject to incomplete and inaccurate reporting. Hauer and Hakkert (1988; see also Hakkert and Hauer 1988) show that: (1) the more incomplete the reporting, the more unreliable become the results of studies relying on officially reported accidents, and (2) the more imprecisely known the level of reporting is, the more unreliable become the results of studies relying on officially reported accidents. Unless one has access to an accident recording system known to be complete, there is really no fully satisfactory way of solving this problem.

To try to account for varying levels of accident reporting in road safety evaluation studies within the framework of meta-analysis, one can test the homogeneity of results as shown in *paper 7*. If the results in a meta-analysis are statistically homogeneous, meaning that they vary no more than chance fluctuations, one can conclude that varying levels of accident reporting do not affect the results of the analysis. If, on the other hand, the individual results are statistically heterogeneous, meta-analysis can proceed by using a random-effects model.

A random-effects model accounts for *varying* levels of accident reporting across studies. It does, however, not account for *incomplete* accident reporting in each study. Hauer and Hakkert have shown how one can account for this, provided that: (1) the reporting level is known and (2) the uncertainty in the estimate of reporting level is known. Unfortunately, this knowledge is rarely likely to be available at the level of detail that is required for meaningful use of the corrections described by Hauer and Hakkert. The level of accident reporting varies, among other things, according to injury severity, group of road user, type of accident and age of victim. Moreover, it may change over time. It could therefore be misleading to correct for incomplete accident reporting in a specific study by using an overall mean reporting level for the country in which the study was reported. For further discussion, see Elvik (1999).

In most road safety evaluation studies, only simple *techniques of analysis (S5)* are used. In non-experimental studies, however, advanced multivariate techniques of analysis are increasingly used. It is possible to code studies with respect to the techniques of analysis used and use this as a variable in meta-analysis. Although none of the appended papers include this variable, it is possible in a meta-analysis to assess the validity of studies with respect to choice of technique of analysis.

The theoretical validity of evaluation research, described in terms of four criteria in Tables 1 and 3, is hardly assessed at all in the appended papers. These papers do not address questions like: Do the results of these studies make sense from a theoretical point of view? To what extent can a theoretical explanation of study findings be given? Were the essential concepts used in an evaluation study adequately defined? Did the evaluation studies contribute to the development of new

theory or new concepts, or are they merely ”puzzle solving” within a highly developed theoretical framework? Or do these studies simply not rely on an explicitly stated theory at all?

In one of the appended papers (*paper 3*), it is stated that evaluation research is atheoretical and that very few results can be ruled out on theoretical grounds. As an illustration of the difference between evaluation research and natural science, the case of heating an iron rod is used. If it does not expand, we would not reject the theory which states that iron expands when heated. We would rather start wondering if there was something wrong with the thermometer used to measure the temperature of the iron rod or the ruler used to measure its length. In evaluation research, on the other hand, researchers are rarely able to rule out certain results in the same manner by invoking a well-established theory.

It is an exaggeration, however, to say that evaluation research is entirely atheoretical. Although evaluation researchers rarely try to establish an elaborate theoretical foundation for their studies, these studies nevertheless frequently use theoretical concepts and rely on implicit hypotheses about the relationships between variables. Examples of theoretical concepts frequently used in road safety evaluation studies include the concepts of attention, driver expectancy, degree of surprise, motives underlying driver behaviour, driver behavioural adaptation, road surface friction, visibility, and risk of apprehension. These concepts have been taken from basic academic disciplines like psychology, economics, physics and probability theory. Their function in evaluation studies is, however, mostly as heuristic devices. Most evaluation studies are not designed primarily for the purpose of testing propositions derived from the theoretical concepts. Their main objective is simply to measure the effects of a measure or programme designed to alleviate a certain social problem, like crime, poverty or accidents.

In most evaluation studies, both the researchers and the sponsors of research have certain prior expectations about study findings. Roughly speaking, the expectation is generally that the measures or programmes that are evaluated will contribute to reducing the problem they were designed to reduce. These prior expectations can, of course, often be stated in the form of hypotheses to be tested. One reason why this is rarely done, at any rate in road safety evaluation studies, is that the hypotheses are too obvious or too trivial to be stated. In the case of road lighting, for example, one could hypothesise that: H1: Road lighting improves visibility at night, and H2: Improved visibility at night reduces the number of accidents. But these hypotheses embody very few theoretically interesting implications; in fact they are truisms bordering on the tautological.

Interest in obtaining a theoretical explanation of study findings often arises only when an evaluation study does not confirm prior expectations. When the provision of road lighting leads to more accidents, one starts wondering what is going on. In recent years, there has been a surge in attempts to model driver behaviour theoretically, spurned to a major extent by an increasing number of ”anomalous” findings in road safety evaluation research. A report issued by the OECD (1990) gives an excellent survey of these models. It remains doubtful, however, if any of the recently developed models of driver behaviour are really

able to establish a firmer theoretical basis for road safety evaluation studies. At their present stage of development, these models can only serve as the basis for non-testable predictions like: ”A road safety measure that is intended to reduce the number of accidents by modifying risk factor A, will have the intended effect unless drivers adapt their behaviour to the measure in a way that completely offsets this effect by modifying risk factors B, C, and D etc”. To make such predictions testable, one would have to specify both when offsetting behavioural adaptation is expected to occur and when it is not expected to occur, and the forms behavioural adaptation will take. It is only when hypotheses become specific about this that they can be falsified, and only falsifiable hypotheses can help in the interpretation of evaluation studies. Otherwise, they serve only as a source of non-testable ad hoc and post hoc explanations.

The preliminary conclusion of this discussion is that there is not much point in trying to assess the theoretical validity of evaluation studies in meta-analysis when the theoretical foundation of these studies is as weak as it is today. Theoretical validity is simply not a relevant criterion of validity for most evaluation studies.

Turning to internal validity, most of the criteria listed in Table 1 have been used to assess the validity of road safety evaluation studies in the appended papers. The only exception concerns criterion *I1, direction of causality*. This criterion states that in order to support causal inferences, an evaluation study must be able to determine the direction of causality between the variables to which a causal inference applies. More specifically, it must be the case that the measure or programme being evaluated is the cause (or one of the causes) of changes in the dependent variable, and not the other way around.

This criterion of validity can to some extent be satisfied by choosing an appropriate study design. In an experimentally designed study (a controlled trial with random assignment), the direction of causality is clear. In all other study designs, however, the direction of causality is not always clear. It is widely believed that direction of causality is clear in before-and-after studies. This belief is unfounded. If, for example, a totally ineffective road safety measure is introduced because an abnormally high number of accidents has been recorded, one will normally find a subsequent decline in the number of accidents due to regression-to-the-mean. But this is a case of reversed causation. It was the high prior number of accidents that caused the introduction of the safety measure, not the safety measure that caused the decline in the number of accidents.

Before-and-after studies do, however, sometimes offer an opportunity to test for direction of causality. Such an opportunity arises when, in a set of before-and-after studies, there are cases both of introducing the measure and of removing it. In this case, one would expect the direction of changes in the dependent variable to depend on whether the measure was introduced or removed. A case illustrating this point is shown in Figure 9.

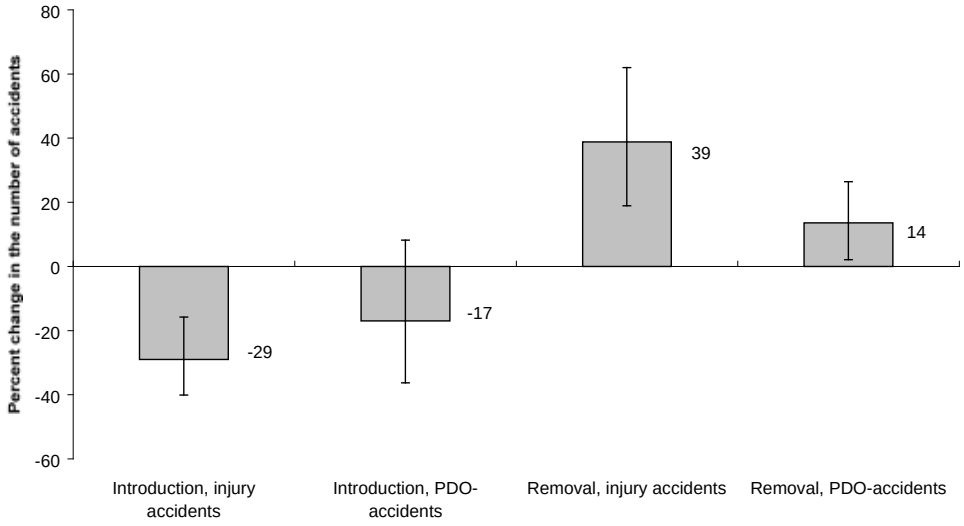


Figure 9: Changes in the number of accidents following the introduction and removal of stop signs at junctions

Figure 9 shows the percentage changes in the number of accidents following the introduction of stop signs in junctions that used to have give way signs, and following the return back to give way signs in junctions that used to have stop signs (Elvik, Mysen and Vaa 1997). It is seen that the changes in the number of accidents go in opposite directions depending on whether the measure is introduced or removed. Moreover, the sizes of the effects are similar and in both cases greater for injury accidents than for property-damage-only accidents (PDO-accidents). These changes indicate that the direction of causality goes from the safety measure to the number of accidents, and not the other way around.

In cross-section studies it is difficult to test the direction of causality directly in this manner. Sometimes it is possible to infer the direction of causality theoretically. As an example, driver gender may causally influence accident rates, but not the other way around. If the direction of causality cannot be inferred theoretically, testing for it in cross-section studies will in most cases have to take the form of assessing the robustness of a statistical relationship between a putative cause and its effect with respect to confounding variables. If the statistical relationship stands up when a large number of confounding variables are controlled in a recursive model, there is more reason to believe that it is a causal relationship in the postulated direction than if it does not stand up to control of confounding variables.

It may be concluded that all the criteria of internal validity proposed in this study are amenable to formal assessment within the framework of meta-analysis.

As far as external validity is concerned, the discussion can be brief. All the criteria of external validity introduced in Table 1 are easily applied in meta-analysis. Testing studies for external validity in meta-analysis relies on the same basic approach as that used to test other aspects of validity. Studies are coded with respect to the variables that describe various aspects of external validity: time, location and study context. In meta-analysis, studies are then stratified with respect to these variables and the results of studies compared across strata. If results are highly similar, external validity is high, meaning that the results of evaluation studies can be generalised in time, across locations and with respect to other aspects of study context.

Testing for external validity is important in assessing evaluation studies. To some extent, the lack of a strong theoretical basis for evaluation research can be compensated for by a high level of external validity. If a finding has been re-produced in a large number of studies made over a long period in different countries and different social settings, and employing different study designs, there is more reason to believe in it than if it has not been reproduced in this manner. Sometimes, there is even reason to believe that a finding represents a lawlike relationship if it has been reproduced a large number of times in different settings.

10 Conclusions, Future Prospects and Research Needs

10.1 Conclusions

The main conclusions of this study will be stated as answers to the main research problems formulated in Chapter 2 and elaborated in subsequent chapters. The first problem that was posed was this:

> Is it possible at all to establish objective criteria of validity in research? Or do the criteria accepted at any time merely reflect the dominant prejudices among researchers?

The arguments of epistemologic relativism to the effect that no objective criteria of scientific knowledge can be established, and that, a fortiori, there are no objective criteria for what counts as good or bad science were discussed. The position taken in this dissertation with respect to epistemologic relativism can be summarised as follows:

1 There are probably not any universally valid criteria of scientific knowledge, if by ”universally valid” one thinks of criteria that have been accepted by everybody throughout history. It is a fact that what counts as scientific knowledge, as opposed to superstition or pseudoscience, has changed over time and is even today in dispute. Moreover, scientists have not always complied perfectly with their own conception of what constitutes good science.
2 These observations do not imply, however, that it is in principle impossible to establish criteria of scientific quality. It is essential to bear in mind that such criteria are *normative* only; they are *not* meant as a *description of how research is actually done*. Moreover, the claim to objectivity made for such criteria signifies only that (a) the criteria are publicly stated and precise, in the sense that they do not admit of multiple and conflicting interpretations, and (b) the criteria are widely, if perhaps not unanimously, accepted by researchers in the field to which they apply.
3 It is recognised that criteria of scientific quality (validity) satisfying these conditions may change over time and may apply only to specific areas of science, not to science in general. The criteria of validity proposed in this dissertation are intended to apply only to evaluation research and reflect the current state-of-the-art with respect to the possibility of formally assessing validity. The criteria reflect the conception of science advocated by logical empiricism.

In other words, the main conclusion is that it is possible to establish objective criteria of validity in research, but that these criteria may change over time and differ between scientific disciplines. The second main problem raised was this:

> Provided that criteria of validity can be established, what is the relevance of those criteria for assessing evaluation research? Should evaluation research be assessed strictly in terms of its validity, or are other bases for assessment more relevant?

It is obvious that, as a matter of fact, the value of evaluation research is not assessed strictly in terms of its validity, at least not as defined in this dissertation. Some researchers have even claimed that validity is largely irrelevant. What counts is the practical utility of evaluation research; the extent to which its results can contribute to solving social problems.

This point of view is not shared in this dissertation. Research that is not valid, for example because it is riddled with methodological shortcomings, is useless for practical purposes. Bad studies simply do not show the effects of the measures or programmes one might like to introduce to curb crime, raise income or reduce the number of accidents. Bad studies are more likely to show the effects of uncontrolled confounding factors or poor data. They have no practical utility. The position taken in this dissertation is that there exists a true effect of programmes introduced to solve social problems; it is the task of evaluation research to reveal this effect. It is of course impossible to claim that a certain evaluation study shows the true effects of a measure. The best one can do, is to give arguments for believing that the findings are as close to the truth as one can get by using the imperfect methods of empirical research. To claim, as some researchers have done, that no objective reality exists is simply to drop out of the world of science and into a world of fancy and opinion in which not even a claim that gravity does not exist can be dismissed as nonsensical.

The third main research problem stated in Chapter 2 was this:

> What forms of knowledge, and which aspects of the research process, can be incorporated into formal criteria of validity? Is any formal list of criteria of validity likely to be supported by the majority of researchers and by the public?

Traditionally, epistemology has been built around a subjective conception of knowledge, often defined as ”justified, true belief”. It is the term ”belief” that renders this conception of knowledge subjective. Knowledge resides in the head of a knowing subject; it consists of statements the subject believes in because they have been shown to be true. A subjective conception of knowledge may not permit very strong criteria of validity to be established. A certain piece of scientific evidence that convinces one person may fail to convince another. Except for the most basic principles of logic and mathematics, there are probably few elements of scientific reasoning that everybody regards as convincing (i e that leads them to believe in statements justified by invoking those elements of reasoning).

According to the subjective conception of knowledge, one might say that there is little knowledge in a subject area if few people are acquainted with the research that has been made in the area. This may seem somewhat odd. In this dissertation, the concept of objective knowledge, as introduced by Karl Popper, has been used to characterise the form of knowledge to which the formal criteria of validity are intended to apply. The criteria of validity are intended to apply only to a written body of knowledge available to all in the form of reports and papers.

As far as the second part of the question posed above is concerned, a standard definition of validity does not seem to exist. The different definitions that have been proposed are, however, not fundamentally at odds with each other. Different definitions of validity emphasise different aspects of the same underlying concept. In this dissertation, a deliberate choice was made to adopt the validity framework of Cook and Campbell (1979), because it includes more aspects of validity than any other conceptions found in the literature.

The fourth problem stated in Chapter 2 was:

> Provided widely accepted formal criteria of validity can be established, is meta-analysis the best approach to assessing the extent to which research conforms to these criteria? Will different approaches to meta-analysis give different results?

This question is a restatement of the main problem of this dissertation:

> To what extent is it possible to assess the validity of evaluation research by conducting meta-analysis of evaluation research studies?

There are two ways of trying to assess the validity of a set of evaluation studies. One approach, which was the only one used until meta-analysis was invented some twenty years ago, is to review studies informally, perhaps sorting them into a few groups, and form an opinion about their validity based on an informal assessment. The other approach is to code studies according to formal criteria of validity and use meta-analysis to assess studies according to these criteria. Informal research syntheses were discussed in Chapter 7, formal criteria of validity designed for use in meta-analysis were introduced in Chapter 8. Applications of these criteria in seven appended studies were discussed in Chapter 9. The main conclusions of these three chapters can be summarised as follows:

1 *Problems of informal research syntheses*
Informal research syntheses are subject to numerous sources of bias that are difficult to detect unless a formal analysis is made. Important sources of bias in informal research syntheses include: (a) Confirmation bias, which means that results confirming prior expectations are treated as more valid than results not confirming prior expectations, even if there is no basis for such a preference in terms of study methodology; (b) Hindsight bias, which denotes a tendency to invent ad hoc explanations of unexpected findings, or insidiously formulating hypotheses after inspecting the data and dressing up the study to

make it look as if these hypotheses were tested as part of the study; (c) Publication bias, which denotes the tendency not to publish studies whose results are believed not be useful, either because they are not statistically significant at conventional levels or because they are in the ”wrong” direction; (d) Belief in the law of small numbers, denoting a tendency to disregard sample size when assessing the relative contributions various studies have made to current knowledge; (e) Capitalisation on chance, which means that random differences in study findings are erroneously interpreted as if they were real. Meta-analysis makes it possible to avoid these pitfalls, at least to some extent.

2 Criteria of validity designed for meta-analysis

A total of twenty criteria of validity designed to assess the validity of evaluation research by means of meta-analysis were proposed. These criteria refer to four types of validity: (a) Statistical conclusion validity, denoting the numerical accuracy and representativeness of a study result or the mean of a set of study results. Nine criteria of statistical conclusion validity were proposed; (b) Theoretical validity, which denotes the extent to which studies are based on an explicit theoretical basis that is supported by study findings. Four criteria of theoretical validity were proposed; (c) Internal validity, which refers to the extent to which a study or a set of studies satisfies commonly accepted conditions for attributing causality to the relationship between the measure or programme that is evaluated and the dependent variable of interest. Four criteria of internal validity were proposed; (d) External validity, which refers to the extent to which the findings of evaluation studies can be generalised to other contexts than those in which each study was made. Three criteria of external validity were proposed. In principle, all the twenty criteria of validity can be used in meta-analysis to formally assess study validity. The simplest approach to doing so, is to code studies with respect to the criteria of validity and stratify them according to the criteria during analysis. If: (i) most studies score high on the criteria for validity, and (ii) study results are similar across the categories of the criteria of validity, it may be concluded that studies are highly valid.

3 Application of the criteria of validity in seven studies

The criteria of validity have been applied in seven studies presented in the appended papers. Thirteen of the twenty criteria were applied formally or informally in these papers. Seven of the criteria were not applied. The studies reported in the appended papers show that the criteria of validity that are most difficult to apply in meta-analysis are those that refer to the possible presence of systematic errors in data and those that refer to theoretical validity. To assess how systematic errors in data or techniques of analysis affect the results of evaluation studies, it is necessary to either (a) have access to data that are known not to contain systematic errors and compare results obtained with these data to results obtained with data containing errors, or (b) statistically model the effects of systematic errors in data, in order to adjust for their effects during analysis. Neither of these options is widely available. It is there-

fore often not possible to assess study validity with respect to errors in data within the framework of meta-analysis. As far as theoretical validity is concerned, it is concluded that this criterion is of comparatively little relevance to evaluation research, because the theoretical foundation of this research is often poorly developed and studies do not aim to test theoretical propositions.

4 *Possible problems in the application of meta-analysis*
This study has also uncovered some problems and limitations in the use of meta-analysis to assess the validity of evaluation research. One possible problem is *study inclusion bias* in meta-analysis, which arises when criteria for inclusion in a meta-analysis are so strict that many relevant studies have to be omitted. Whenever a large number of relevant studies have to be omitted, it is necessary to try to test for study inclusion bias in the meta-analysis. A second problem is the *garbage in, garbage out* problem, which can arise when all evaluation studies that have been reported in an area are really quite bad. Meta-analysis can never improve the quality of original studies, except in those rather few cases when a reanalysis is possible. The garbage in, garbage out problem is, however, common to all formal techniques of analysis. In general, poor data should be analysed by means of simple techniques only, whereas good data can be subjected to more sophisticated analyses. A third limitation in using meta-analysis to assess study validity is the fact that *no widely accepted overall measure of study validity exists*. In this dissertation, validity has been assessed in terms of twenty criteria referring to four types of validity. It will sometimes be the case, however, that studies which are strong by one criterion are weak by another. How should the overall validity of such studies be assessed? The meta-analyses presented in the appended papers have assessed study validity by rating studies according to one criterion at a time. Finally, a fourth problem in the use of meta-analysis is that there exists *several techniques of meta-analysis* that do not always give identical results. The choice of technique is not always obvious.

The main conclusion of the study stated in broad terms is that it is to a certain extent possible to assess the validity of evaluation research by means of meta-analysis. But it is probably too optimistic to believe that the use of meta-analysis to assess the validity of evaluation research will resolve all controversies surrounding such research. It may therefore not lead out of the mess created by the perennial controversies involving evaluation research in the United States. Some of these controversies are not about validity at all. Formal criteria of study validity will not help in resolving those controversies.

Some aspects of study validity can be formally assessed by means of meta-analysis, others are less amenable to formal assessment. There will always be subtle, qualitative aspects of research that influence our assessment of its validity, but are impossible to code formally in a way that makes sense. The style of presentation used in a paper is one of these qualitative aspects. Somehow, most of us place greater confidence in a paper when the authors are clearly aware of the limitations of their research and point them out, than in an otherwise similar paper presented in a less humble way. In science, humility instills confidence. Hubris destroys confidence. But humility and hubris are qualities that cannot be reduced to numbers.

Meta-analysis is best suited to empirical research. It is a lot more difficult to use meta-analysis to assess the validity of theoretical models. Consider, for example, the models of driver behaviour that have been proposed in road safety research in recent years (for a survey, see Bjørnskau, Midtland and Sagberg 1993). It is not obvious how to assess the validity of these models at all, let alone how to use meta-analysis to do so.

10.2 Future prospects and research needs

Meta-analysis is only about twenty years old. It is therefore still in its infancy. The use of meta-analysis is growing rapidly. Hundreds of meta-analyses have by now been reported and the scope of problems subjected to meta-analysis is expanding all the time. The expanding use of meta-analysis is probably related to several trends that characterise modern science:

1 The volume of research is expanding. In some subject areas, there are hundreds of studies. Summarising these studies in the traditional narrative format is nearly impossible.
2 It is increasingly important to separate the wheat from the chaff in research. The expanding volume of research means that more excellent studies are done, but also that more bad studies are done. Sorting studies by quality is an essential part of extracting and synthesising knowledge from previous studies.
3 Research syntheses are performed with two major objectives in mind: (a) To find the main tendency (”average finding”) in the findings of previous research, and (b) To identify factors that influence the findings of previous research (moderating factors).

Meta-analysis is excellently suited to these needs. It is therefore safe to predict that the use of meta-analysis will continue to grow and become ever more sophisticated. To make meta-analysis even more useful as a tool for summarising research and assessing its quality, there are several aspects of it that need further development. These aspects include:

1 *Multivariate techniques of meta-analysis*
There is a need for developing multivariate techniques of meta-analysis adapted to different weighting schemes. In the appended papers, the logodds method of meta-analysis has been applied. The analyses in the appended papers proceed by stratifying the data set according to the variables of interest. Multivariate techniques of analysis are clearly superior to the stratification technique, but no description of such techniques developed for the logodds method of meta-analysis has been found in the literature.

2 *Overall measure of validity*
It is desirable to develop an overall measure of validity that summarises all aspects of the concept in the form of a general assessment. In order to develop such a measure, it is necessary to rate the importance of various types of validity, to establish rules for trading off one type of validity against another and to develop a uniform system for coding all criteria of validity.

3 *Choice of technique of meta-analysis*
For many problems, there is a choice of technique of meta-analysis, that is several techniques can be used and it is not always obvious which one is the best. There is a need for testing the sensitivity of the results of meta-analyses with respect to choice of technique. It may discredit meta-analysis if the results of such analyses turn out to be very sensitive to the choice of technique, and if that choice is, essentially, arbitrary.

References

Bangert-Drowns, R. L. Review of Developments in Meta-analytic Method. Psychological Bulletin, 99, 388-399, 1986.

Begg, C. B. Publication Bias. In Cooper, H.; Hedges, L. V. (Eds): The Handbook of Research Synthesis, Chapter 25, 399-409. New York, NY, Russell Sage Foundation, 1994.

Begg, C. B.; Berlin, J. A. Publication Bias: a Problem in Interpreting Medical Data. Journal of the Royal Statistical Society, Series A, 151, Part 3, 419-463, 1988.

Berlin, J. A.; Begg, C. B.; Louis, T. A. An assessment of publication bias using a sample of published clinical trials. Journal of the American Statistical Association, 84, 381-392, 1989.

Berlin, J. A.; Laird, N. M.; Sacks, H. S.; Chalmers, T. C. A comparison of statistical methods for combining event rates from clinical trials. Statistics in Medicine, 8, 141-151, 1989.

Bjørnskau, T.; Fosser, S. Bilisters atferdstilpasning til innføring av vegbelysning. Resultater fra en før- og etterundersøkelse på E-18 i Aust-Agder. TØI rapport 332. Oslo, Transportøkonomisk institutt, 1996.

Bjørnskau, T.; Midtland, K.; Sagberg, F. Beskrivelse og drøfting av aktuelle modeller for bilføreres atferd. Arbeidsdokument TST/0472/93. Oslo, Transportøkonomisk institutt, 1993.

Black, J. A.; Champion, D. J. Methods and Issues in Social Research. New York, NY, John Wiley, 1976.

Blakstad, F.; Giæver, T. Ulykkesfrekvenser på vegstrekninger i tett og middels tett bebyggelse. Rapport STF63 A89005. Trondheim, SINTEF Samferdselsteknikk, 1989.

Blalock, H. M. Causal Inferences in Nonexperimental Research, New York, NY, W. W. Norton, 1961.

Borger, A.; Fosser, S.; Ingebrigtsen, S.; Sætermo, I-A. Underrapportering av trafikkulykker. TØI rapport 318. Oslo, Transportøkonomisk institutt, 1995.

Boruch, R. F. The Future of Controlled Randomized Experiments: A Briefing. Evaluation Practice, 15, 265-274, 1994.

Brehmer, B. Vad är det för fel på transportforskningen? Innlegg på VTIs og KFBs Forskardagar, Linköping, Januar, 1993.

Broughton, J. The effect on motorcycling of the 1981 Transport Act. TRRL Research Report 106. Crowthorne, Berkshire, Transport and Road Research Laboratory, 1987.

Carmines, E. G.; Zeller, R. A. Reliability and validity assessment. Series: Quantitative applications in the social sciences. Beverly Hills/London, Sage Publications, 1979.

Christensen-Szalanski, J. J. J.; Willham, C. F. The Hindsight Bias: A Meta-Analysis. Organizational Behavior and Human Decision Processes, 48, 147-168, 1991.

Cicchetti, D. V. The Reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. The Behavioral and Brain Sciences, 14, 119-186, 1991.

Cook, T. D.; Campbell, D. T. Quasi-Experimentation. Design and Analysis Issues for Field Settings. Chicago, Ill, RandMcNally, 1979.

Cooper, H.; Hedges, L. V. (Eds): The Handbook of Research Synthesis. New York, NY, Russell Sage Foundation, 1994.

Cordray, D. S. Strengthening Causal Interpretations of Nonexperimental Data: The Role of Meta-analysis. New Directions for Program Evaluation, No 60, 59-95, 1993.

Coursol, A.; Wagner, E. E. Effect of Positive Findings on Submission and Acceptance Rates: A Note on Meta-Analysis Bias. Professional Psychology: Research and Practice, 17, 136-137, 1986.

Crossen, C. Tainted Truth. The Manipulation of Fact in America. New York, NY, Simon and Schuster, 1994.

Cullen, D. J.; Macauley, A. Consistency of Peer Reviewers Who Evaluate Scientific Articles. In Weeks, R. A.; Kinser, D. L. (Eds). Editing the Refereed Scientific Journal. Practical, Political, and Ethical Issues, 13-16. New York, NY, The IEEE Press, 1994.

DerSimonian, R.; Laird, N. Meta-Analysis in Clinical Trials. Controlled Clinical Trials, 7, 177-188, 1986.

Dickersin, K.; Min, Y-I. Publication Bias: The Problem That Won't Go Away. In Warren, K. S.; Mosteller, F. (Eds): Doing more good than harm: the evaluation of health care interventions, 135-148. Annals of the New York Academy of Sciences, Volume 703, 1993.

Elvik, R. The safety value of guardrails and crash cushions: a meta-analysis of evidence from evaluation studies. Accident Analysis and Prevention, 27, 523-549, 1995A.

Elvik, R. Meta-Analysis of Evaluations of Public Lighting as Accident Countermeasure. Transportation Research Record, 1485, 112-123, 1995B.

Elvik, R. Evaluation of Risto Kulmala's doctoral dissertation: "Safety at rural three- and four-arm junctions. Development and applications of accident prediction models". Reprint No 86. Oslo, Institute of Transport Economics, 1995C.

Elvik, R. Does knowledge of safety effect help to predict how effective a measure will be? Accident Analysis and Prevention, 28, 339-347, 1996A.

Elvik, R. A meta-analysis of studies concerning the safety effects of daytime running lights on cars. Accident Analysis and Prevention, 28, 685-694, 1996B.

Elvik, R. Evaluations of road accident blackspot treatment: a case of the Iron Law of evaluation studies? Accident Analysis and Prevention, 29, 191-199, 1997.

Elvik, R. Evaluating the statistical conclusion validity of weighted mean results in meta-analysis by analysing funnel graph diagrams. Accident Analysis and Prevention, 30, 255-266, 1998A.

Elvik, R. Are road safety evaluation studies published in peer reviewed journals more valid than similar studies not published in peer reviewed journals? Accident Analysis and Prevention, 30, 101-118, 1998B.

Elvik, R. Incomplete accident reporting: A meta-analysis of studies made in thirteen countries. Paper 990047. Submitted for the 1999 Annual Meeting of the Transportation Research Board, Washington DC, January 10-16, 1999.

Elvik, R.; Mysen, A. B.; Vaa, T. Trafikksikkerhetshåndbok. Oversikt over virkninger, kostnader og offentlige ansvarsforhold for 124 trafikksikkerhetstiltak. Tredje utgave. Oslo, Transportøkonomisk institutt, 1997.

Elvik, R.; Vaa, T. Human factors, road accident data and information technology. Report 67. Oslo, Institute of Transport Economics, 1990.

Elwood, J. M. Causal Relationships in Medicine. A Practical System for Critical Appraisal. Oxford, Oxford University Press, 1988.

Eysenck, H. J. An exercise in mega-silliness. American Psychologist, 33, 517, 1978.

Feyerabend, P. Against Method. Outline of an Anarchist Theory of Knowledge. London, Verso, 1975.

Feyerabend, P. Science in a Free Society. London, Verso, 1978.

Feyerabend, P. Farewell to Reason. London, Verso, 1987.

Fischhoff, B. Hindsight ¹ Foresight: The effects of outcome knowledge on judgment under uncertainty. Journal of Experimental Psychology: Human Perception and Performance, 1, 288-299, 1975.

Fischhoff, B.; Beyth, R. "I knew it would happen" – remembered probabilities of once-future things. Organizational Behavior and Human Performance, 13, 1-16, 1975.

Fleiss, J. L. Statistical Methods for Rates and Proportions. Second Edition. New York, NY, John Wiley and Sons, 1981.

Fleiss, J. L.; Gross, A. J. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. Journal of Clinical Epidemiology, 44, 127-139, 1991.

Fridstrøm, L. An Econometric Model of Energy Consumption, Road Use, and Traffic Accidents (preliminary title). Draft doctoral dissertation. Oslo, Institute of Transport Economics, 1998.

Fridstrøm, L.; Ifver, J.; Ingebrigtsen, S.; Kulmala, R.; Krogsgård Thomsen, L. Explaining the variation in road accident counts. Report Nord 1993:35. Copenhagen, Nordic Council of Ministers, 1993.

Fridstrøm, L.; Ifver, J.; Ingebrigtsen, S.; Kulmala, R.; Krogsgård Thomsen, L. Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. Accident Analysis and Prevention, 27, 1-20, 1995.

Glass, G. V.; McGaw, B.; Smith, M. L. Meta-Analysis in Social Research. Beverly Hills/London, Sage Publications, 1981.

Gleser, L. J.; Olkin, I. Stochastically Dependent Effect Sizes. In Cooper, H.; Hedges, L. V. (Eds): The Handbook of Research Synthesis, Chapter 22, 339-355. New York, NY, Russell Sage Foundation, 1994.

Griffin, L. I. III. Using before-and-after data to estimate the effectiveness of accident countermeasures implemented at several treatment sites. Unpublished manuscript. Texas Transportation Institute, The Texas A&M University System, College Station, Tx, December 1989.

Guba E. G.; Lincoln, Y. S. The Countenances of Fourth-Generation Evaluation: Description, Judgment, and Negotiation. In Palumbo, D. J. (Ed). The Politics of Program Evaluation 202-234. Newbury Park, Ca, Sage Publications, 1987.

Hakkert, A. S.; Hauer, E. The extent and implications of incomplete and inaccurate road accident reporting. In Rothengatter, J. A.; deBruin, R. (Eds): Road User Behaviour: Theory and Research, 2-11. Van Gorcum, Assen/Maastricht, 1988.

Hargens, L. L. Scholarly consensus and journal rejection rates. American Sociological Review, 53, 139-151, 1988.

Hauer, E. A Case for Science-Based Road Safety Design and Management. Paper presented at the conference ”Highway Safety: At the Crossroads”, San Antonio, Texas, March 1988. Proceedings published by American Society of Civil Engineers.

Hauer, E. The behaviour of public bodies and the delivery of road safety. In Koornstra, M. J.; Christensen, J. (Eds): Enforcement and Rewarding. Strategies and Effects, Proceedings of the International Road Safety Symposium in Copenhagen, Denmark, September 19-21, 1990, 134-138. Leidschendam, SWOV Institute for Road Safety Research, 1991.

Hauer, E. Should Stop Yield? Matters of Method in Safety Research. ITE-Journal, September 1991, 25-31.

Hauer, E. A note on three estimators of safety effect. Traffic Engineering and Control, 33, 388-393, 1992.

Hauer, E. Observational Before-After Studies in Road Safety. Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety. Oxford, Pergamon Press, 1997.

Hauer, E.; Hakkert, A. S. Extent and Some Implications of Incomplete Accident Reporting. Transportation Research Record, 1185, 1-10. National Research Council, Washington DC, 1988.

Hawkins, S. A.; Hastie, R. Hindsight: Biased Judgments of Past Events After the Outcomes Are Known. Psychological Bulletin, 107, 311-327, 1990.

Heckman, J. J.; Smith, J. A. Assessing the Case for Social Experiments. Journal of Economic Perspectives, 9, 85-110, 1995.

Hedges, L. V.; Olkin, I. Statistical Methods for Meta-Analysis. San Diego, Ca, Academic Press, 1985.

Hellevik, O. Forskningsmetode i sosiologi og statsvitenskap. Oslo, Universitets-forlaget, 1977.

Hempel, C. G. Aspects of Scientific Explanation and other Essays in the Philosophy of Science. New York, NY, The Free Press, 1965.

Hill, A. B. The Environment and Disease: Association or Causation. Proceedings of the Royal Society of Medicine, Section of Occupational Medicine, Meeting January 14 1965, 295-300.

Hovi, J.; Rasch, B. E. Samfunnsvitenskapelige analyseprinsipper. Oslo, Fagbok-forlaget, 1996.

Hunter, J. E.; Schmidt, F. L. Methods of Meta-Analysis. Correcting Error and Bias in Research Findings. Newbury Park, Ca, Sage Publications, 1990.

Ketvirtis, A. Road Illumination and Traffic Safety. Prepared for Road and Motor Vehicle Traffic Safety Branch, Transport Canada. Ottawa, Transport Canada, 1977.

Klayman, J.; Ha, Y-W. Confirmation, Disconfirmation, and Information in Hypo-thesis Testing. Psychological Review, 94, 211-228, 1987.

Kleinbaum, D. G.; Kupper, L. L.; Morgenstern, H. Epidemiologic Research. Principles and Methods. New York, NY, Van Nostrand Reinhold, 1982.

Kuritz, S. J.; Landis, J. R.; Koch, G. G. A general overview of Mantel-Haenszel Methods: Applications and recent developments. Annual Review of Public Health, 9, 123-160, 1988.

Light, R. J.; Pillemer, D. B. Summing Up. The Science of Reviewing Research. Cambridge, Mass, Harvard University Press, 1984.

McGee, H. W.; Blankenship, M. R. Guidelines for converting stop to yield control at intersections. National Cooperative Highway Research Program Report 320. Washington DC, Transportation Research Board, 1989.

Mohr, L. B. Impact Analysis for Program Evaluation. Newbury Park, Ca, Sage Publications, 1992.

OECD Scientific Expert Group. Behavioural adaptations to changes in the road transport system. Paris, OECD, 1990.

Palumbo, D. J. Politics and Evaluation. In Palumbo, D. J. (Ed). The Politics of Program Evaluation 12-46. Newbury Park, Ca, Sage Publications, 1987.

Palumbo, D. J. (Ed). The Politics of Program Evaluation. Newbury Park, Ca, Sage Publications, 1987.

Peters, D. P.; Ceci, S. J. Peer-review practices of psychological journals: The fate of published articles, submitted again. The Behavioral and Brain Sciences, 5, 187-255, 1982.

Pollard, W. E. Bayesian statistics for evaluation research. An introduction. Beverly Hills, Ca, Sage Publications, 1986.

Popper, K. R. Objective Knowledge. An Evolutionary Approach. Revised Edition. Oxford, Oxford University Press, 1979.

Rennie, D. The Failure of Scientists to Identify Fraudulent Papers, and the Decline in Citation Rates of Papers After They Have Been Publicly Identified as Being Fraudulent. In Weeks, R. A.; Kinser, D. L. (Eds). Editing the Refereed Scientific Journal. Practical, Political, and Ethical Issues, 73-75. New York, NY, The IEEE Press, 1994.

Rosenthal, R. The ”File Drawer Problem” and Tolerance for Null Results. Psychological Bulletin, 86, 638-641, 1979.

Rosenthal, R. Meta-Analytic Procedures for Social Research. Applied Social Research Methods Series Volume 6. Newbury Park, Ca, Sage Publications, 1991.

Rosenthal, R. Parametric Measures of Effect Size. In Cooper, H.; Hedges, L. V. (Eds): The Handbook of Research Synthesis, Chapter 16, 231-244. New York, NY, Russell Sage Foundation, 1994.

Rossi, P. H.; Freeman, H. E. Evaluation. A Systematic Approach. Third Edition. Beverly Hills, Ca, Sage Publications, 1985.

Russell, B. In praise of idleness, and other essays. London, Routledge, 1935.

Siegel, H. Farewell to Feyerabend. Inquiry, 33, 343-369, 1989.

Shadish, W. R.; Haddock, C. K. Combining estimates of effect size. In Cooper, H.; Hedges, L. V. (Eds): The Handbook of Research Synthesis, Chapter 18, 261-281. New York, NY, Russell Sage Foundation, 1994.

Slovic, P.; Fischhoff, B. On the Psychology of Experimental Surprises. Journal of Experimental Psychology: Human Perception and Performance, 3, 544-551, 1977.

Stephan, P. E. The Economics of Science. Journal of Economic Literature, 34, 1199-1235, 1996.

Stern, P. C.; Kalof, L. Evaluating social science research. Second edition. New York, NY, Oxford University Press, 1996.

Tanner, J. C. A problem in the combination of accident frequencies. Biometrika, 45, 331-342, 1958.

Tversky, A.; Kahneman, D. Belief in the law of small numbers. Psychological Bulletin, 76, 105-110, 1971.

Vaa, T. Politiets fartskontroller: Virkning på fart og subjektiv oppdagelsesrisiko ved ulike overvåkingsnivåer. TØI-rapport 301. Oslo, Transportøkonomisk institutt, 1995.

Wagenaar, A. C.; Zobeck, T. S.; Williams, G. D.; Hingson, R. D. Methods used in studies of drink-drive control efforts: a meta-analysis of the literature from 1960 to 1991. Accident Analysis and Prevention, 27, 307-316, 1995.

Wason, P. C. On the failure to eliminate hypotheses in a conceptual task. Quarterly Journal of Experimental Psychology, 12, 129-140, 1960.

Wason, P. C. On the failure to eliminate hypotheses – a second look. In Wason, P. C.; Johnson-Laird, P. N. (Eds). Thinking and reasoning, 165-174. Harmondsworth, Middlesex, Penguin Books, 1968.

Wason, P. C.; Johnson-Laird, P. N. (Eds). Thinking and reasoning. Harmondsworth, Middlesex, Penguin Books, 1968.

Weeks, R. A.; Kinser, D. L. (Eds). Editing the Refereed Scientific Journal. Practical, Political, and Ethical Issues. New York, NY, The IEEE Press, 1994.

Weiss, C. H. Evaluation Research. Methods for Assessing Program Effectiveness. Englewood Cliffs, NJ, Prentice Hall, 1972.

Wolf, F. M. Meta-Analysis. Quantitative Methods for Research Synthesis. Series: Quantitative applications in the social sciences. Newbury Park/London, Sage Publications, 1986.

Ørnes, A. L. Trafikksikkerhetseffekten av gang- og sykkelveger. Oppdragsrapport 56. Trondheim, Norges Tekniske Høgskole, Forskningsgruppen, Institutt for samferdselsteknikk, 1981.