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).

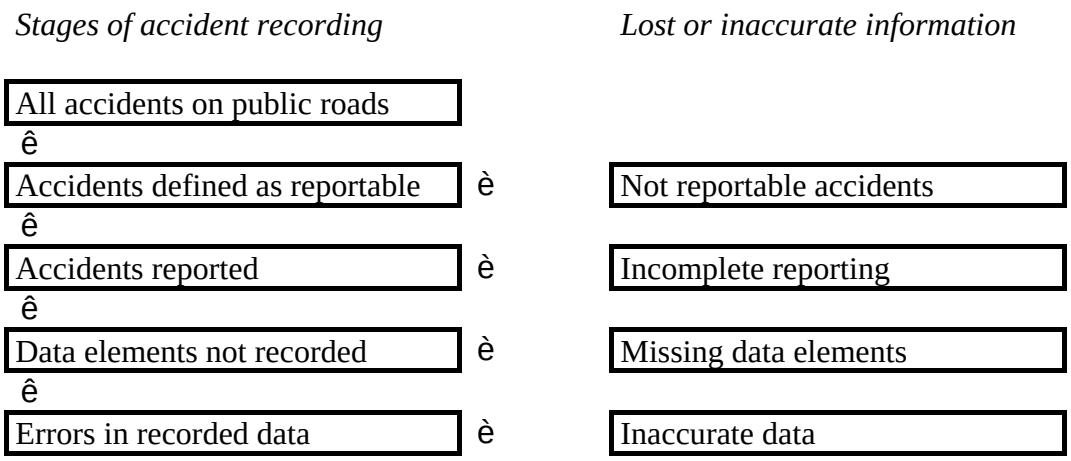


Figure 1: Sources of error and data loss in official accident records

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).

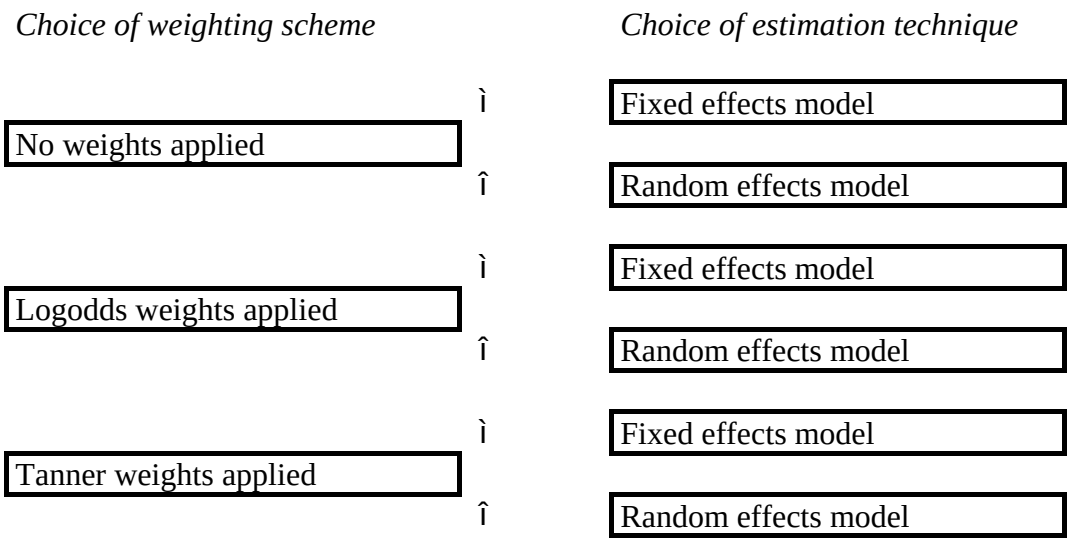


Figure 2: Taxonomy of techniques for estimating mean results in meta-analyses of road safety evaluation studies

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.

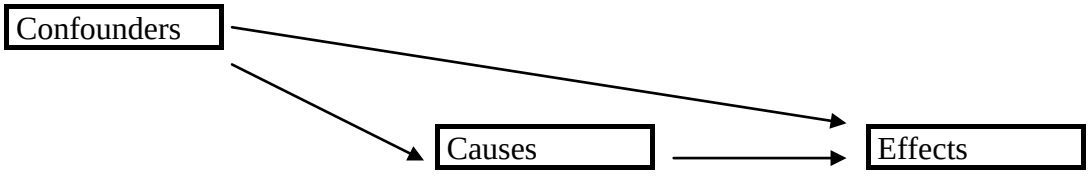
8.4 Internal validity

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.

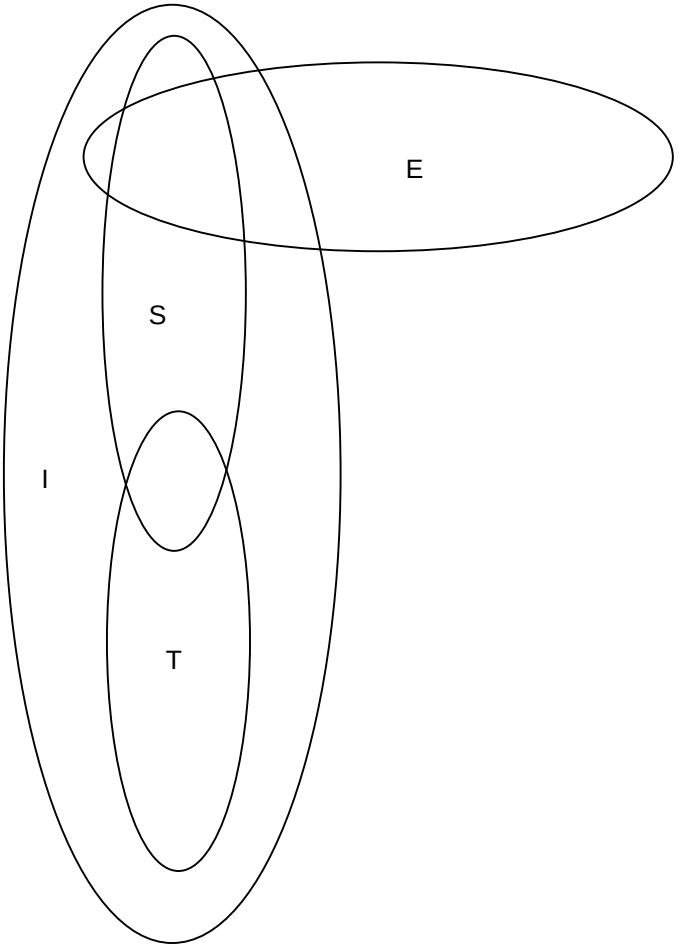


Figure 3: The relationship between types of validity. S = Statistical, T = Theoretical, I = Internal, E = External

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.

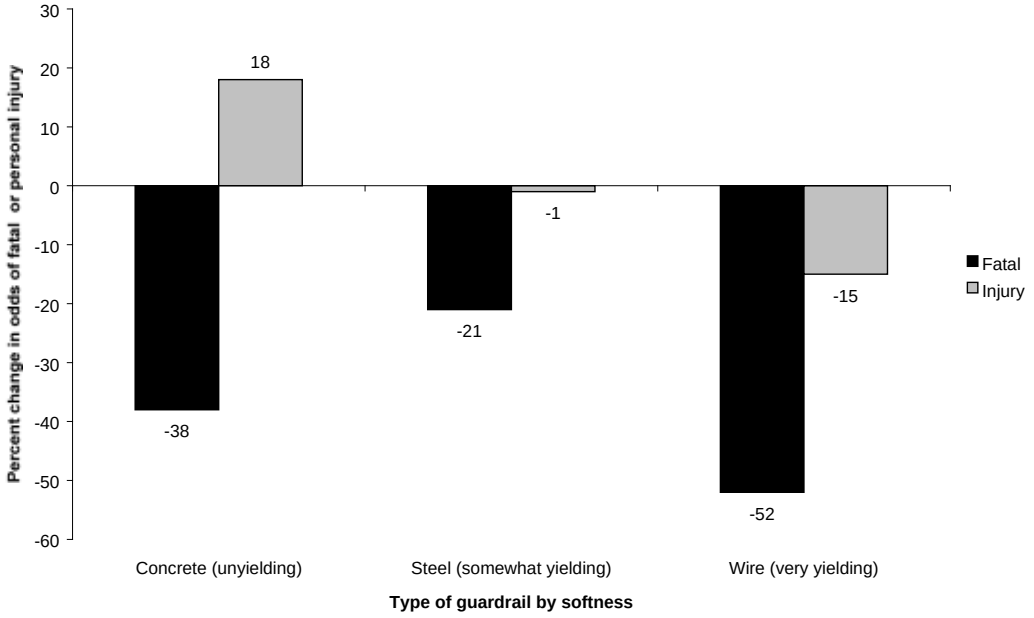


Figure 4: Dose-response pattern in effects of median guardrails

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.

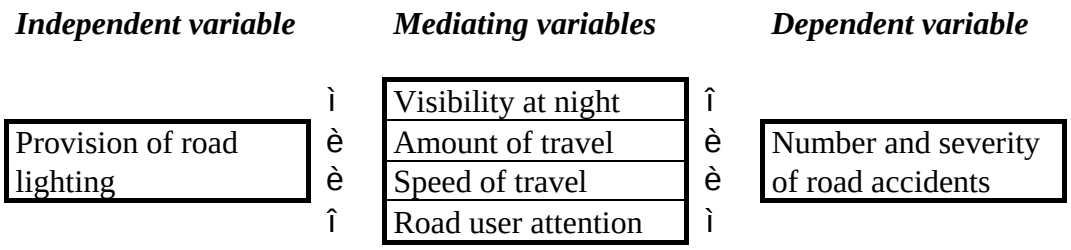


Figure 5: Causal chain for effects of road lighting on the number and severity of road accidents

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.

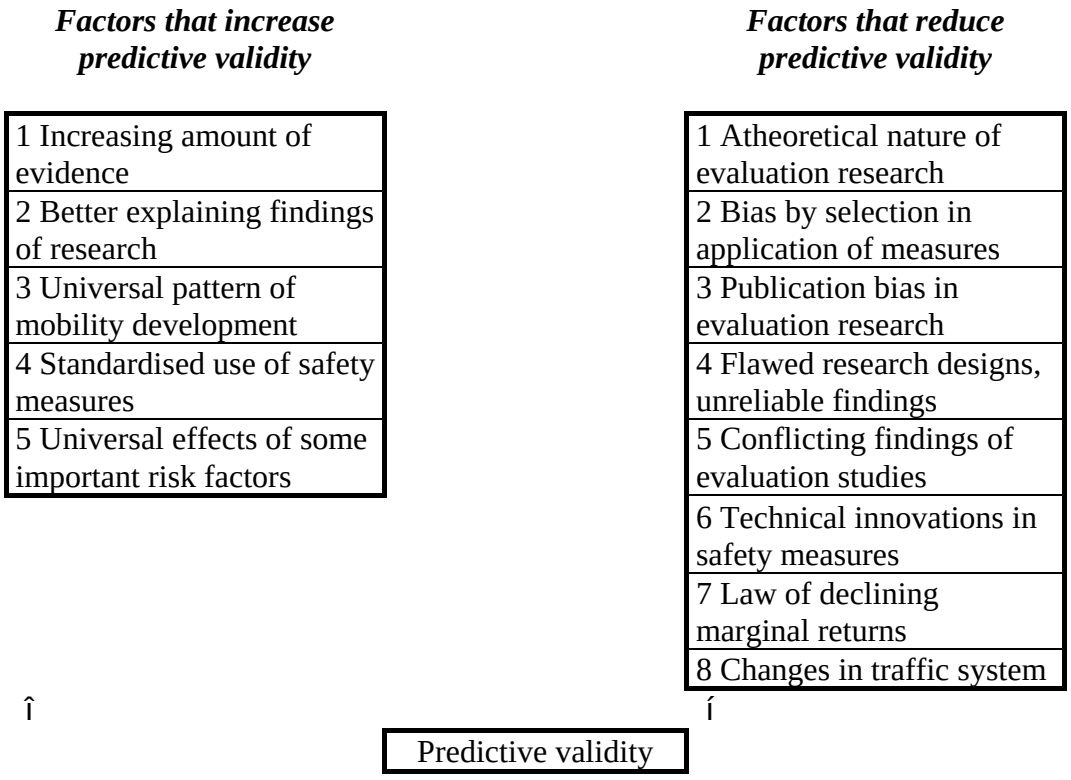


Figure 6: Factors affecting the predictive validity of evaluation studies

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.

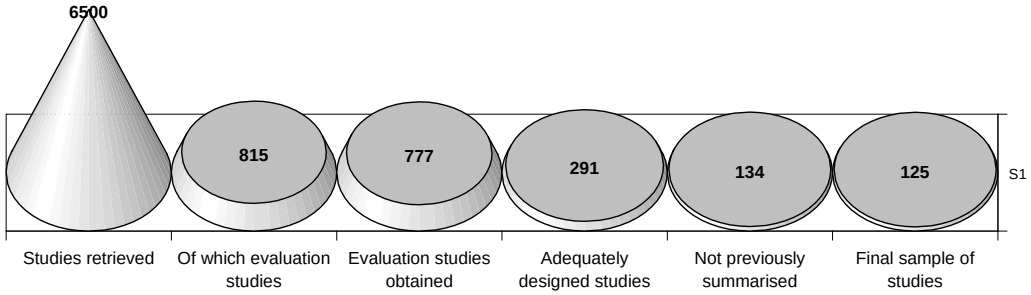


Figure 7: Successive stages of study exclusion in meta-analysis of measures to control drinking and driving. Adapted from Wagenaar et al 1995

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.

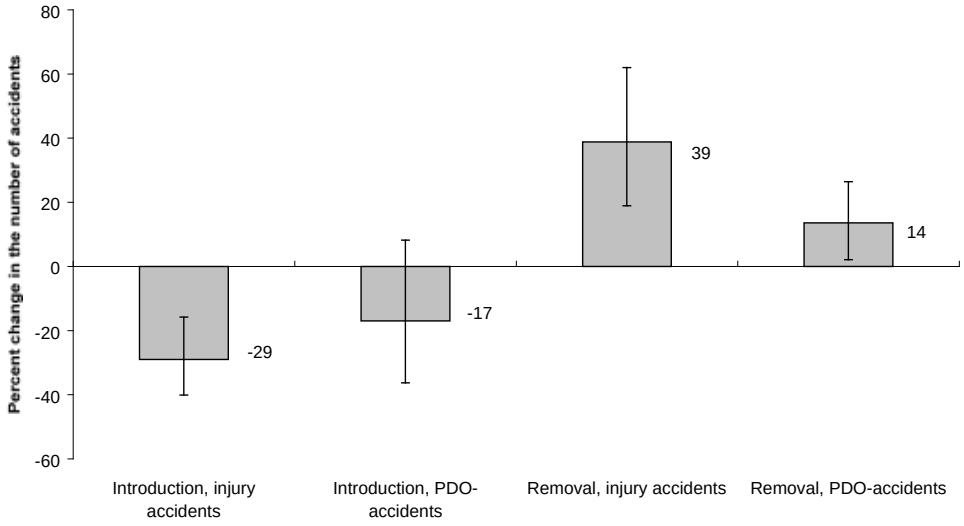


Figure 9: Changes in the number of accidents following the introduction and removal of stop signs at junctions

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.

Paper 1

Pergamon

Accid. Anal. and Prev., Vol. 27, No. 4, pp. 523–549, 1995

0001-4575/95 $9.50 + .00

0001-4575(95)00003-8

THE SAFETY VALUE OF GUARDRAILS AND CRASH CUSHIONS: A META-ANALYSIS OF EVIDENCE FROM EVALUATION STUDIES

RUNE ELVIK

Institute of Transport Economics, PO Box 6110 Etterstad, N-0602, Oslo, Norway

(Accepted 2 November 1994)

Abstract—Evidence from 32 studies that have evaluated the safety effects of median barriers, guardrails along the edge of the road, and crash cushions (impact attenuators) is summarized by means of a meta-analysis. Two hundred and thirty-two (232) estimates of safety effects are included in the meta-analysis. The presence of publication bias is tested by means of the funnel graph method. For most subsets of the data, no evidence of publication bias is found. Weighted mean estimates of safety effects are computed by means of the logodds method. Median barriers are found to increase accident rate, but reduce accident severity. Guardrails and crash cushions are found to reduce both accident rate and accident severity. The effects of guardrails and crash cushions on accident rate have been less extensively studied than the effects on accident severity. Current estimates of the effects on accident rate are highly uncertain because of methodological shortcomings of available studies. The effects of guardrails on accident severity are found to be quite robust with respect to study design and the number of confounding variables controlled in each study. In general, random variation in the number of accidents is the most important source of variation in study results.

Keywords—Guardrail, Meta-analysis, Evaluation studies, Safety effectiveness

INTRODUCTION

Guardrails are widely used in all motorized countries to reduce the consequences of accidents in which vehicles run off the road or cross the median on divided highways. It is universally accepted (Michie, Calcote, and Bronstad 1971) that guardrails should be installed only where the consequences of striking the guardrail are judged to be less serious than the consequences of striking the guarded object. Implementing this guideline in practice is, however, difficult. Michie, Calcote and Bronstad (1971, p. 10) note that "an ideal guardrail system—that is, one that safely redirects errant vehicles without endangering other traffic and without causing injuries or fatalities among the occupants—would improve safety at most highway sites, with the possible exception of those with flat embankments that are clear of obstacles. However, such ideal systems do not exist; guardrail and median barrier systems are intrinsic roadside hazards and provide the errant vehicles with only a relative degree of protection." Various guardrail designs have been subjected to extensive crash tests, but the laboratory-like environment of such tests greatly simplify the situations leading to real crashes. Moreover, crash tests of guardrails say nothing about the consequences of striking any of the objects guardrails are designed to protect vehicles from striking. There is, in other words, no substitute for real-world experience comparing the severity of accidents in which guardrails were hit to that of accidents in which other objects were hit in order to find out when hitting guardrails reduces the consequences of accidents.

This paper reports the results of a meta-analysis of 32 evaluation studies that have quantified the effects of guardrails and crash cushions (also known as impact attenuators) on the probability and severity of accidents. The objective of the analysis is to summarize the evidence from these evaluation studies with respect to the following questions:

1. Does installing median barriers, guardrails (along the edge of the road), and crash cushions affect the probability of accident occurrence—that is, the number of accidents per vehicle kilometre of travel along the site or section where median barriers, guardrails, or crash cushions were installed?

2. How do median barriers, guardrails, and crash cushions affect the severity of accidents? When do these devices reduce the chances of a fatality, given that an accident has occurred? When do they reduce the chances of an injury, given that an accident has occurred?
3. Can the evidence from evaluation studies be trusted? Do the results of these studies vary according to study design and other variables characterizing study quality or the context where studies were made? Which are the best estimates of the safety value of median barriers, guardrails, and crash cushions?

DATA AND METHOD

Retrieval of evaluation studies

The studies included in the meta-analysis were retrieved by means of a systematic literature survey. The literature survey consisted of scanning selected journals, like *Highway Research Record, Transportation Research Record,* and *Accident Analysis and Prevention.* In addition, studies referred to in papers published in the journals that were scanned were obtained. A more detailed description of the literature survey is given elsewhere (Elvik 1994). A total of 32 studies, containing a total of 232 numerical estimates of the effects of median barriers, guardrails, or crash cushions on the probability and/or severity of accidents were retrieved. Complete data for the studies included are given in Appendix A.

Data describing each estimate of safety effect

In the meta-analysis, each estimate of safety effect constitutes the unit of analysis. Thus, sample size in the meta-analysis is 232. For each estimate, the following data were recorded:

1. Authors of study
2. Year of publication. Various years from 1956 through 1993 represented
3. Country where data used in study were collected. Six countries represented
4. Study design. Coded variable. See Table 1
5. Confounding variables controlled in study. Coded variable. See Table 1
6. Guarded object. Coded variable. See Table 1
7. Type of guardrail. Coded variable. See Table 1
8. Accident severity. Coded variable. See Table 1
9. Number of accidents of given severity before or without guardrail
10. Number of accidents of given severity after or with guardrail
11. Effect on accident rate. Numerical estimate. See Table 1
12. Effect on chances of fatality. Numerical estimate. See Table 1
13. Effect on chances of personal injury. Numerical estimate. See Table 1

Table 1 gives more detailed information concerning each of the variables recorded.

There are two main kinds of study design: before-and-after designs and case-control designs. In the former kind of design, accident rate and the severity of accidents are compared at given sites before and after median barriers, guardrails, or crash cushions have been installed. In the latter kind of design, accident rate and the severity of accidents at sites with median barriers, guardrails, or crash cushions (cases) are compared to accident rates and severity measures at sites without median barriers, guardrails, or crash cushions (controls). For both designs, the validity of results depends, among other things, on how successful a study is in eliminating the effects of various confounding variables on accident rate and the severity of accidents. Table 1 lists a number of important confounding variables that any study ought to take account of. The list in Table 1 is by no means complete. It includes just the confounding variables that were judged to be most important.

The list of guarded objects refers in particular to case-control studies of guardrails, in which the accident rate and the severity of accidents at sites with guardrails are compared to sites where one of the listed objects was struck.

Description of countermeasures

A distinction is made between three countermeasures: (i) Median barriers, that is guardrails in the median of divided highways; (ii) Guardrails along the edge of the road; (iii) Crash cushions (impact attenuators), that is energy absorbing structures placed in front of, for example, bridge piers or exit ramps to reduced the severity of crashes. For median barriers and guardrails, a further distinction is made between different kinds according to their rigidness. Figure 1 shows different kinds of median barriers and guardrails. Figure 2 shows an example of an impact attentuator.

Measures of safety effect

Three measures of the safety effect of median barriers, guardrails, and crash cushions have been defined. The net effect on safety is defined as the

Table 1. Information recorded for each study included in the meta-analysis

Variables	Categories of each variable
Authors	Listed alphabetically
Year of publication	1956 through 1993
Country	Six countries represented
Study design	2131 = Before-and-after study with matched comparison group and traffic volume data for both groups 25 = Case-control studies where the effects of confounding variables are estimated by means of multivariate analysis 26 = Case-control studies where cases and controls are stratified according to one or more confounding variables 27 = Case-control studies where cases and controls have been matched in pairs according to one or more confounding variables 31 = Before-and-after studies with no comparison group, but traffic volume data before and after
Confounding variables	For probability of accident occurrence: I.A Traffic volume I.B Type of road (access control, number of lanes, presence of median) I.C Alignment (horizontal and vertical curvature) I.D Cross section (lane width, shoulder width) I.E Enviroumental risk factors (rain, slippery road surface, etc) For severity of accidents: II.A Vehicle mass (weight) II.B Vehicle occupancy (number of persons in vehicle) II.C Use of restraint systems (seat belts) II.D Departure angle II.E Vehicle trajectory after departure (vaulting, rollover, etc) II.F Type of guardrail (see below) II.G Guarded object (see below) II.H Impact angle II.I Distance to object from edge of road II.J Impact speed
Guarded object	A Different type of guardrail (in cases of replacement) B Median C Part of bridge structure D Ditch (manmade) E Embankment (natural sideslope along roads) F Rockside G Tree H Utility pole I Highway sign J Unspecified object
Type of guardrail	A Concrete barrier B Steel W-beam guardrail C Steel wire guardrail
Accident severity	A Fatal accident (accident where at least one person dies) B Injury accident (accident where at least one person is injured) C Property-damage-only accident
Number of accidents	A Before or without guardrail or crash cushion B After or with guardrail or crash cushion
Effect on accident rate	Per cent change in number of accidents per million vehicle kilometres of travel (accident rate)
Effect on accident severity	A Per cent change in conditional probability of fatal accident, granted that an accident has occurred (number of fatal accidents) B Per cent change in conditional probability of injury accident, granted that an accident has occurred (number of injury accidents)

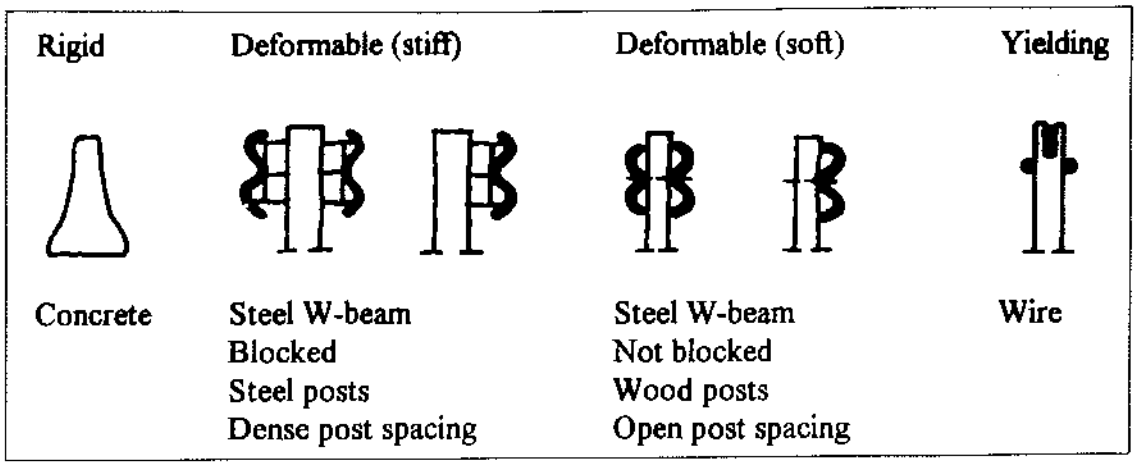


Fig. 1. Types of guardrail according to rigidness.

product of the effect on the probability of accident occurrence and the effect on the severity of accidents:

Net safety effect = Change in probability of accidents × change in severity of accidents

The probability of accidents is measured in terms of the accident rate:

$$\text{Accident rate} = \frac{\text{Number of accidents of all degrees of severity}}{\text{Number of vehicle kilometres of travel}}$$

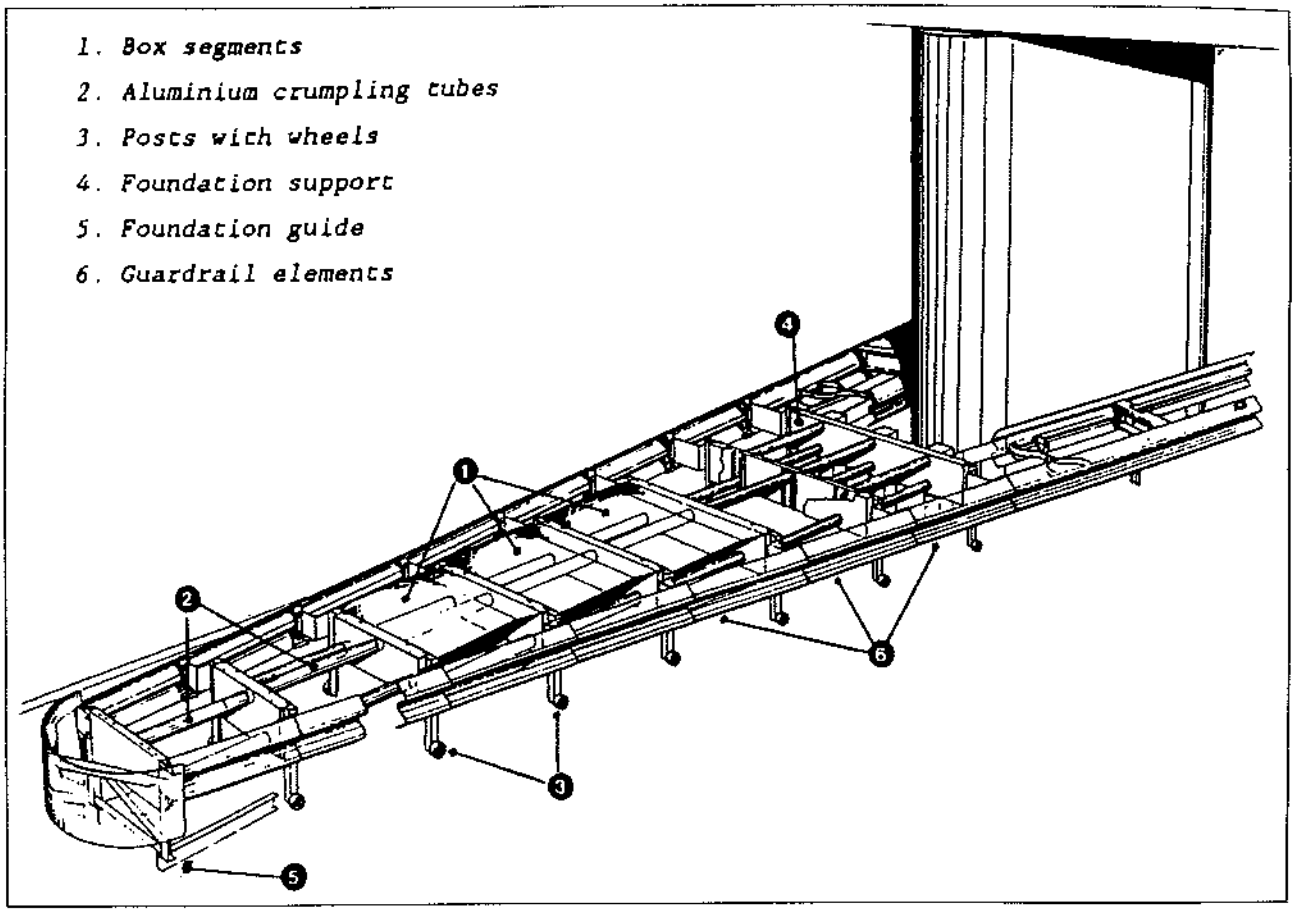


Fig. 2. Example of a crash cushion (impact attenuator). Dutch RIMOB-system. (Schoon, 1990). *Source:* SWOV Institute for Road Safety Research/Dutch ministry of Transport and Public Works. Reprinted by permission.

Effects are defined as changes in accident rate. Effects on the consequences of accidents are defined in terms of the odds of a fatal accident or an injury accident with guardrail compared to that without guardrail (the odds ratio):

$$\text{Change in chance of fatal accident} = \frac{\left(\frac{\text{Number of fatal accidents with guardrail}}{\text{Number of injury and PDO-accidents with guardrail}}\right)}{\left(\frac{\text{Number of fatal accidents without guardrail}}{\text{Number of injury and PDO-accidents without guardrail}}\right)}$$

In the corresponding odds ratio for injury accidents, the numerator includes fatal and injury accidents and the denominator includes just property-damage-only (PDO) accidents. If the odds ratio is below 1.0, the chance of a fatal or injury accident has been reduced. If it is above 1.0, the chances have increased.

A possible objection to using the odds ratio as a measure of the effect of guardrails and crash cushions on accident severity, is that it can give apparently biased results when the accident rate, or the total number of accidents, changes. A numerical example will clarify the point. Suppose that before guardrails are installed, there are 100 accidents, of which 40 are injury accidents. The odds on having an injury accident is 40/60 = 0.667. Suppose, further, that when guardrails are installed, the number of accidents increases to 110, of which 30 are injury accidents. The odds on having an injury accident is now 30/80 = 0.375. The odds ratio is 0.375/0.667 = 0.562, implying a reduction of about 44% in the number of injury accidents. However, as the total number of accidents has increased, the actual number of injury accidents has decreased by just 25%, from 40 to 30.

As an alternative to using the odds ratio as a measure of effect on accident severity, one might consider using the actual number of accidents of a certain severity. Using the actual number of accidents of a certain severity as the measure of the effect of guardrails and crash cushions on accident severity would, however, give biased results. To continue the example given above, suppose that the total number of accidents declined to 80 when guardrails were put up, of which 30 were injury accidents. The number of injury accidents would then be reduced by 25% (from 40 to 30). The odds on having an injury accident, given that accident has occurred, would be reduced by 10% (from 0.667 to 0.600).

Measures of the effect on accident severity need to be defined in terms of the distribution of a given number of accidents by levels of severity. Thus, the measure of effect on accident severity chosen in this paper says nothing about the actual number of accidents. It refers strictly to the conditional probability of sustaining a fatality or a personal injury, given that an accident has occurred.

Testing for publication bias and modality of the distribution of results

The objective of this study is to summarize evidence from several evaluation studies in the form of a weighted mean estimate of the safety effects of median barriers, guardrails, and crash cushions. For a weighted mean estimate of safety effect to make sense, three requirements must be fulfilled: (i) there should not be publication bias, (ii) the assumption that all results belong to a distribution having a well defined mean value should be reasonably well supported, (iii) all studies should use comparable measures of safety effect.

The term *publication bias* refers to the tendency not to publish results that are unwanted or believed not to be useful, for example because they show an increase in accidents or because they are not statistically significant (Light and Pillemer, 1984). Light and Pillemer (1984) have developed a graphical technique of testing for publication bias, called the funnel graph method. It relies on visual inspection of a diagram in which each study result is plotted in a coordinate system. The horizontal axis shows each result. The vertical axis shows the sample size each result is based on. The idea is that if there is no publication bias, the scatter plot of study results should resemble the form of a funnel turned upside down. The dispersion of points in the diagram should narrow as sample size increases, since large samples provide more precise estimates of effects than small samples. If the tails of the scatter plot are symmetrical, this indicates that there is no publication bias.

The shape of scatter plots in funnel graph diagrams indicates whether it makes sense to estimate a weighted mean safety effect. If the funnel graph is bimodal (has two humps) or multimodal, or if there is no clear pattern in the scatter plot, a weighted mean will not be very informative. If a funnel pattern is clearly visible, estimating a weighted mean safety effect will be informative and indicate the size of the effect that studies tend to converge to as sample size increases. The measures of safety effect used in different studies are considered comparable, in that they are identically defined and, as far as is known, rely on police reported accidents in all studies.

Technique of meta-analysis

There are several techniques of meta-analysis (Fleiss 1981; Light and Pillemer 1984; Hedges and

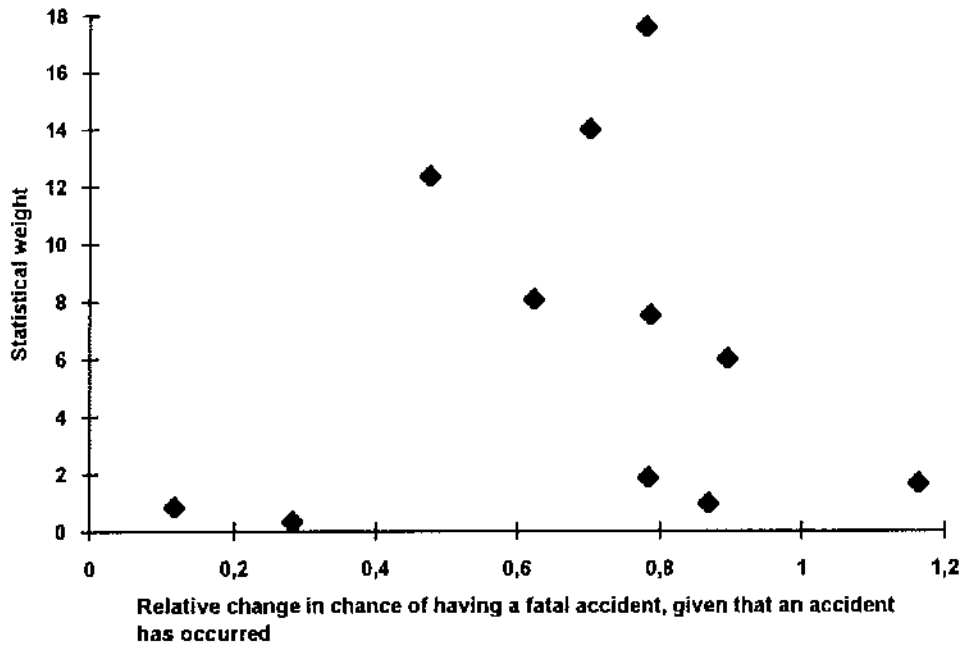


Fig. 3. Funnel graph diagram for change in number of fatal accidents associated with median barriers.

Olkin 1985; Hunter and Schmidt, 1990; Rosenthal, 1991). In this paper, the logodds method, described by Fleiss (1981) has been applied. Briefly stated, this method involves computing the natural logarithm of the odds ratio that measures safety effect in each study and weighting results by means of weights that are proportional to the inverse of the variance of each result. A short technical description of the method is given in Appendix B.

PUBLICATION BIAS AND THE REPRESENTATIVENESS OF WEIGHTED MEAN SAFETY EFFECTS

In order to test for publication bias and the representativeness of weighted mean safety effects, six funnel graph diagrams have been prepared. Figures 3 and 4 refer to median barriers, Figures 5 and 6 refer to guardrails and Figures 7 and 8 refer to crash cushions. All figures show the effects of the various safety devices on accident severity, since the most common warrant for installing median barriers, guardrails, or crash cushions is to reduce accident severity, rather than the number of accidents. Statistical weight is taken as a measure of sample size (see Appendix B).

Figure 3 shows the results of studies evaluating the effects of median barriers on the chance of having a fatal accident, given that an accident has occurred. The shape of a funnel turned upside down can be discerned in the figure. The distribution of data points does not give any clear indication of publication bias. The scatter plot is unimodal, indicating that a weighted mean safety effect would be representative of the results that studies tend to converge to as sample size increases.

In Fig. 4, referring to injury accidents, the distribution of data points is more skewed, possibly indicating the presence of publication bias in favour of results showing a decline in the chance of having an injury accident. On the other hand, the result that is based on the largest sample size indicates an increase in the chance of having an injury accident. Besides, there is an outlier indicating a fourfold increase in the chance of having an injury accident. It is concluded that the scatter of data points is not sufficiently skewed to justify rejecting the evidence as merely reflecting publication bias, rather than the effect of median barriers.

Figures 5 and 6 refer to guardrails. Figure 5 shows results for fatal accidents, Figure 6 for injury accidents. Three outlying data points, based on small samples and showing substantial increases in the number of accidents, were omitted from each figure. The preponderance of results based on small samples is apparent in both figures. No clear indication of publication bias can be found. For fatal accidents, there is great dispersion of results even from large samples, as shown by the two data points uppermost in the figure. For injury accidents, there is less dispersion of results. It is concluded that a

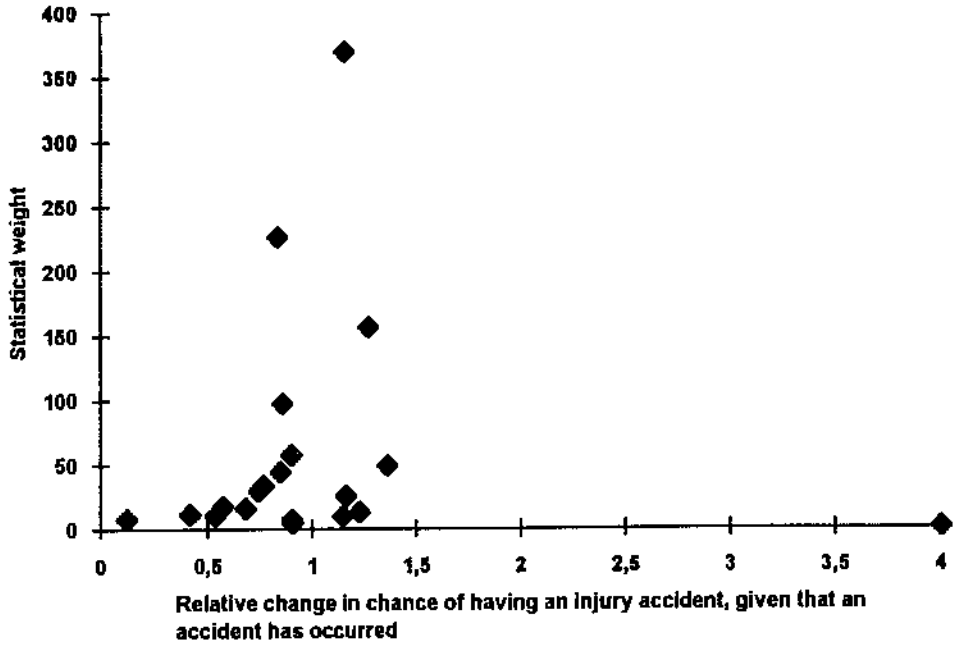


Fig. 4. Funnel graph diagram for change in number of injury accidents associated with median barriers.

weighted mean estimate of safety effects makes sense both for fatal accidents and injury accidents.

Figures 7 and 8 refer to crash cushions. Both figures contain few data points, showing great dispersion. In both figures, the data point based on the largest accident sample is located to the left of the centre of gravity of the scatter plot. This means that a weighted mean safety effect based on the data points of Figs. 7 and 8 is likely to be misleading. It will not be representative of the safety effects shown by most studies. Apart from the nonrepresentativeness of the data point based on the largest accident sample, there is no clear indication of publication bias in Figs. 7 and 8.

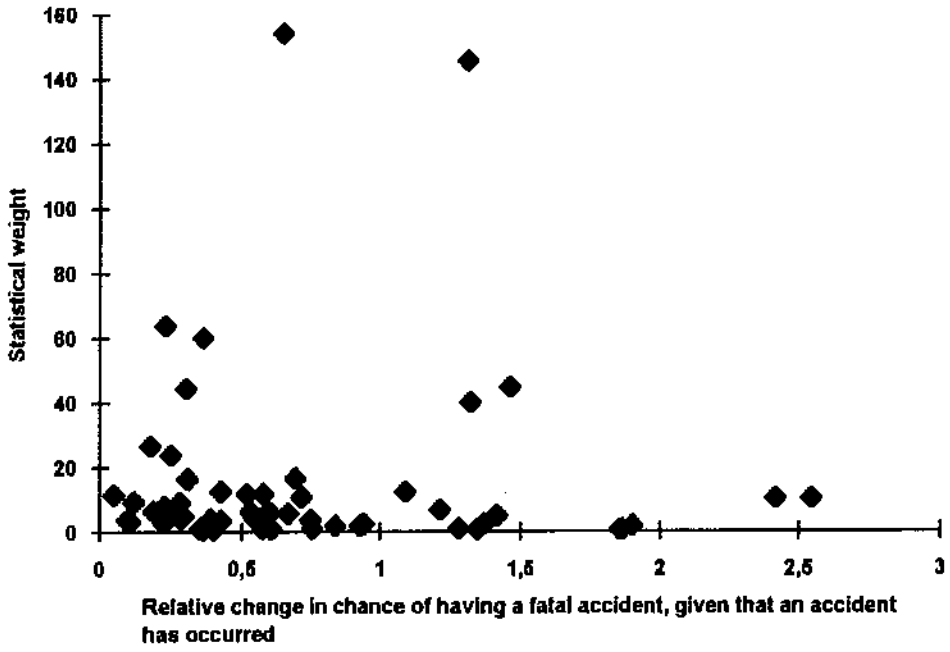


Fig. 5. Funnel graph diagram for change in number of fatal accidents associated with guardrails.
Note: Three data points (increase in number of accidents) have been omitted.

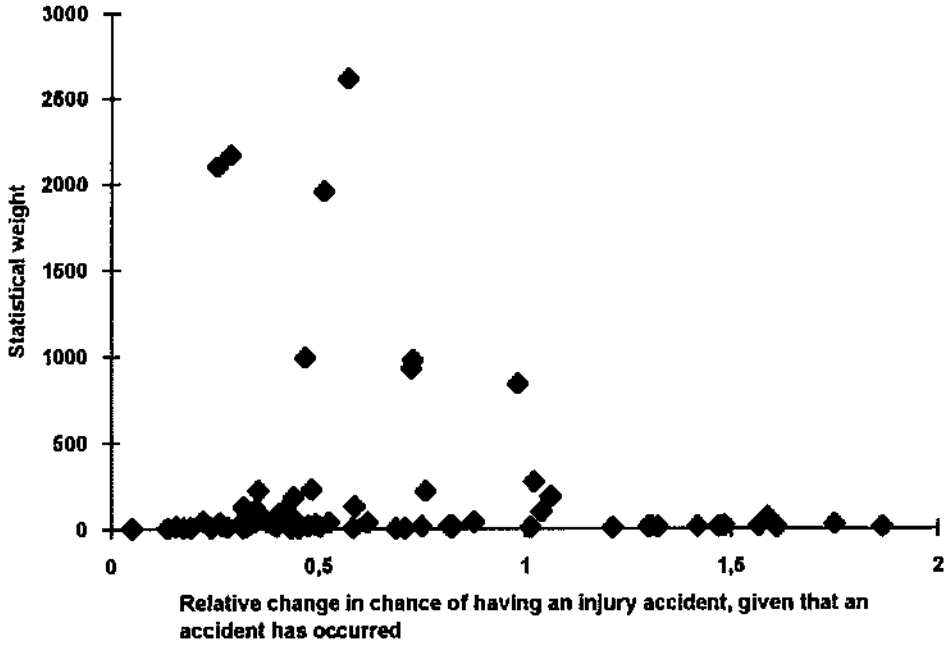


Fig. 6. Funnel graph diagram for change in number of injury accidents associated with guardrails. *Note:* Three data points (increase in number of accidents) have been omitted.

MEAN SAFETY EFFECTS OF MEDIAN BARRIERS, GUARDRAILS AND CRASH CUSHIONS

Table 2 presents estimates of the mean weighted safety effects of median barriers, guardrails, and crash cushions. In order to save space in the table, effects on the probability of accident occurrence are referred to as "accident rate". Effects on the conditional probability of a fatal accident, given that an accident has occurred, are referred to as "fatal accidents". The corresponding effects for injury accidents are referred to as "injury accidents". A distinction is made between new installations and re-

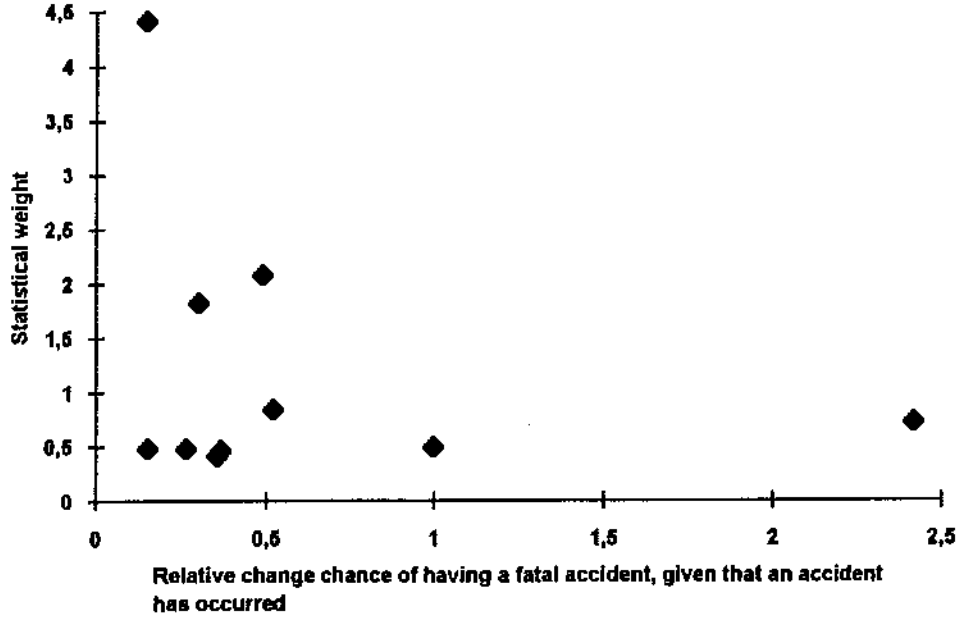


Fig. 7. Funnel graph diagram for change in number of fatal accidents associated with crash cushions.

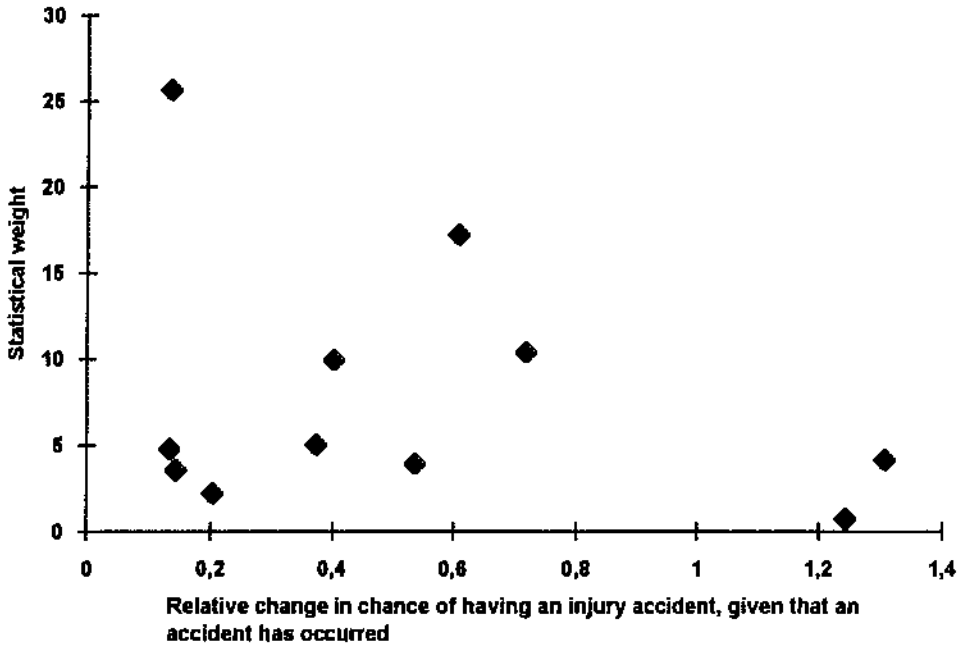


Fig. 8. Funnel graph diagram for change in number of injury accidents associated with crash cushions.

placement of old installations. Results in the latter category refer to replacing an existing barrier or guardrail with a more deformable or yielding type of barrier or guardrail.

Table 2 shows that installing median barriers increases the total number of accidents by about 30%. This effect is found both for new installations and for replacement of existing median barriers. The increase in the number of accidents is statistically significant at the 5% level. The severity of accidents is reduced. New median barriers reduce the probability of fatal accidents, given the total number of accidents, but apparently have no effect on the probability of injury accidents. Replacing existing barriers reduces the probability of injury accidents, but has no statistically significant effect on the probability of fatal accidents.

Guardrails reduce both the number of accidents and their severity. This applies both to new installations and to replacements of old installations. The evidence concerning the effect of guardrails on accident rate is far less extensive than the evidence concerning the effects on accident severity. This can be seen from the contribution of the various results to the statistical weights. Results that refer to the effects on accident rate constitute just 1.2% of the statistical weights of all results included in Table 2. Results concerning the effects on accident severity, on the other hand, constitute 64.7% of the statistical weights.

Crash cushions reduce both the number and severity of accidents. There are, however, few studies, some of which are of rather doubtful validity, as will be discussed in more detail in a subsequent section of the paper.

THE VALIDITY OF EVALUATIONS OF MEDIAN BARRIERS, GUARDRAILS, AND CRASH CUSHIONS

The results of studies that have evaluated the safety effects of median barriers, guardrails, and crash cushions vary substantially, as shown by Figs. 3–8. There are at least four potential sources of variation in study results: (i) publication bias, (ii) random variation in the number of accidents, (iii) variation related to the design and quality of data of each evaluation study, (iv) systematic variation in the effect of the countermeasures, depending on, for example, the design of guardrails and the kind of objects they protect from.

The possibility of publication bias was discussed previously. Although some indications of publication bias were found, the conclusion was that publication bias is unlikely to be a major threat to the validity of the results of the studies.

Random variation in the number of accidents contributes to most of the variation in study results. This is seen in Table 3. In Table 3, two measures of the variability of study results have been esti-

Table 2. Summary of weighted mean estimates of the safety effects of median barriers, guardrails, and crash cushions

Type of guardrail	Type of installation	Measure of safety effect	Proportion of statistical weights	Per cent change in measure of safety effect		
				Lower 95%	Best estimate	Upper 95%
Median barrier	New installation	Accident rate*	0.1858	+25	+29	+32
		Fatal accidents§	0.0028	-14	-32	-46
		Injury accidents#	0.0466	+4	-2	-7
	Replace old	Accident rate	0.0693	+31	+37	+44
		Fatal accidents	0.0011	-24	+10	+61
		Injury accidents	0.0303	-21	-26	-31
Guardrail	New installation	Accident rate	0.0121	-18	-27	-35
		Fatal accidents	0.0336	-40	-44	-48
		Injury accidents	0.6074	-51	-52	-53
	Replace old	Accident rate	0.0000	NA	NA	NA
		Fatal accidents	0.0005	+2	-41	-66
		Injury accidents	0.0059	-21	-33	-43
Crash cushion	New installation	Accident rate	0.0007	-74	-84	-90
		Fatal accidents	0.0005	-46	-69	-83
		Injury accidents	0.0034	-60	-68	-74

NA = Not available
The statistical weights sum to 1.0000

*Number of accidents per million vehicle kilometres of travel.
§Conditional probability of sustaining fatal injury, given that an accident has occurred.
#Conditional probability of sustaining personal injury, given that an accident has occurred.

Table 3. Variability and sources of variation in weighted mean estimates of the safety effects of median barriers, guardrails, and crash cushions

Type of guardrail	Type of installation	Measure of safety effect	Proportion of statistical weights	Cofficient of variation (#)	Proportion of random variance (§)
Median barrier	New installation	Accident rate*	0.1858	0.676	0.020
		Fatal accidents§	0.0028	0.236	1.000
		Injury accidents#	0.0466	0.244	0.467
	Replace old	Accident rate	0.0693	0.194	0.027
		Fatal accidents	0.0011	0.320	1.000
		Injury accidents	0.0303	0.260	0.179
Guardrail	New installation	Accident rate	0.0121	1.629	0.044
		Fatal accidents	0.0336	1.619	1.000
		Injury accidents	0.6074	0.837	0.035
	Replace old	Accident rate	0.0000	NA	NA
		Fatal accidents	0.0005	1.728	1.000
		Injury accidents	0.0059	0.121	1.000
Crash cushion	New installation	Accident rate	0.0007	0.244	1.000
		Fatal accidents	0.0005	1.798	1.000
		Injury accidents	0.0034	1.000	0.528

NA = not available
(#) = standard deviation divided by mean safety effect
(§) = proportion of variance accounted for by random variation in the number of accidents
The statistical weights sum to 1.0000

*Number of accidents per million vehicle kilometres of travel.
§Conditional probability of sustaining fatal injury, given that an accident has occurred.
#Conditional probability of sustaining personal injury, given that an accident has occurred.

mated. The first one, which shows the amount of variation, is the coefficient of variation. The coefficient of variation is the (weighted) standard deviation divided by the (weighted) mean. Table 3 shows that the coefficient of variation is above 1.0 in many cases.

The second measure of variability given in Table 3, is the proportion of variance in study results accounted for by random variation in the number of accidents. Appendix 2 shows how this proportion was estimated. A value of, for example, 0.467 in Table 3, means that 46.7% of the total variance in the study results can be attributed to random variation in the number of accidents. Table 3 shows that random variation in the number of accidents accounts for all of the variance in study results in six cases (out of 14) and for more than half of the variance in one case. For these cases, no further analysis of sources of variation in study results is possible. Such analyses would merely capitalize on chance.

The best possibilities for analysing the contributions of study design and systematic variation to the variation in study results is given by results referring to the effects of new guardrails on the probability of injury accidents. Study results vary substantially (coefficient of variation 0.837), but a minor proportion of this variation is accounted for by random variation in the number of accidents (0.035). Table 4 shows how the weighted mean effects of new guardrails on the probability of injury accidents vary according to four confounding variables: (i) study design, (ii) confounding variables controlled for in each study design, (iii) type of object guarded by guardrail, and (iv) decade of publication of study.

Nearly all studies have used a case-control design (design 26). In these studies, accident experience at sites with guardrails is compared to accident experience at sites without guardrails, for example sites where trees or utility poles were struck by vehicles leaving the roadway. In studies of this kind, it is important to ensure that the guardrail sites are as similar to the comparison sites as possible. Otherwise, any differences found in accident experience may have been caused by other variables affecting the probability and severity of accidents, not guardrails. In Table 4, studies have been grouped according to which of the confounding variables affecting accident severity and listed in Table 1 that they

Table 4. Effects of selected confounding variables on weighted mean effects of guardrails on the number of injury accidents

Confounding variables	Categories of each variable	Proportion of statistical weights	Per cent change in number of injury accidents		
			Lower 95%	Best estimate	Upper 95%
Study design (cf Table 1)	Design 26	0.9987	-51	-52	-53
	Design 27	0.0009	-44	-68	-81
	Design 31	0.0004	+26	-42	-74
Variables controlled (cf Table 1)	Design 26 - G	0.0411	-30	-35	-40
	Design 26 - FG	0.4593	-41	-42	-43
	Design 26 - BFG	0.0717	-24	-28	-32
	Design 26 - EFG	0.0096	-60	-66	-71
	Design 26 - AFGH	0.4171	-63	-64	-65
	Design 27 - FGI	0.0004	+24	-42	-72
	Design 27 - ACFG	0.0004	-62	-83	-92
	Design 31 - FG	0.0004	+26	-42	-74
Guarded object	Highway sign	0.1310	-1	-6	-10
	Ditch	0.0654	-24	-29	-33
	Part of bridge	0.0417	-29	-34	-39
	Embankment	0.0988	-52	-54	-56
	Utility pole	0.3442	-57	-58	-59
	Rockside (cutting)	0.0143	-57	-62	-67
	Tree	0.3037	-61	-62	-63
	Unspecified object	0.0009	-44	-68	-81
Decade of study	1960s	0.0323	-69	-72	-74
	1970s	0.1224	-31	-35	-37
	1980s	0.4283	-37	-39	-40
	1990s	0.4170	-63	-64	-65
Note: the statistical weights sum to 1.0000 for each variable					

have taken account of. It is readily seen that no study has controlled for all the confounding variables listed in Table 1.

On the other hand, it is reassuring that the effects of guardrails do not disappear in studies controlling for more confounding variables. The so called Iron Law of Evaluation Studies (Rossi and Freeman 1985)—which states that the better an evaluation study is technically, the less likely it is to show positive program effects—predicts that as more confounding variables are controlled, the estimated safety effects of a countermeasure are likely to become smaller. Apparently this law does not apply to studies that have evaluated the safety effects of guardrails.

There is systematic variation in the effects of guardrails with respect to the kind of object they protect errant vehicles from striking. The largest effects are found for trees, rock sides (road located in rock cutting), and utility poles. The effects of guardrails do not appear to have diminished over time. However, more recent studies are often technically better than older studies. Thus, the effects of study decade cannot be separated from those of study quality.

DISCUSSION

Traditionally, literature surveys have been informal, narrative, and uncritical. Often they have merely presented summaries of previous research, without attempting to synthesize results or discuss their validity and reliability. Using meta-analysis, it is possible improve the quality of literature surveys, but not the quality of surveyed literature.

The safety effects of median barriers, guardrails, and crash cushions have been studied extensively. The most extensive research has addressed the effects of guardrails on accident severity. The effects of guardrails on accident rate have been less studied. Most studies that have evaluated the effects of guardrails on accident severity have not evaluated their effects on accident rate. With respect to median barriers, the opposite is true. Their effects on accident rate have been studied more extensively than their effects on accident severity. Few studies have evaluated crash cushions. As far as the effect on accident rate is concerned, just one study, a before-and-after study at sites with a bad accident record (Houh, Epstein, and Lee 1986), was found. The results of this study are probably flawed, as the authors did not take account of the regression-to-the-mean effect likely to occur at the study sites.

Based on the studies included in this meta-analysis, it is likely that median barriers increase accident rate, but reduce accident severity. Guardrails appear to reduce both accident rate and accident severity, but their effects on accident rate have not been evaluated extensively. Crash cushions appear to reduce both accident rate and accident severity, but the true effects are probably overstated in the reviewed evaluation studies. The estimated effect on accident rate contains uncontrolled regression-to-the-mean effects. The weighted mean effects on accident severity are unduly influenced by a few nonrepresentative results of studies based on greater accident samples than the other studies.

The technical quality of studies that have evaluated the effects of median barriers, guardrails, and crash cushions is not as good as one would like it to be. In particular, studies dealing with effects on accident severity have not taken account of a number of important confounding variables known to affect accident severity. However, the results appear to be quite robust with respect to the effects of confounding variables. The results of studies that have taken account of different confounding variables were compared. The estimated effects of guardrails on accident severity did not disappear as more confounding variables were controlled.

CONCLUSIONS

The main results of the research reported in this paper can be summarized as follows:

1. By means of a systematic literature survey, 32 studies that have evaluated the safety effects of median barriers, guardrails, and crash cushions were retrieved. The 32 evaluation studies contained a total of 232 numerical estimates of safety effects.
2. A distinction was made between effects on accident rate (number of accident of all degrees of severity per million vehicle kilometres of travel) and effects on accident severity (the conditional probability of a fatal accident, or an injury accident, given that an accident as occurred).
3. Weighted mean safety effects were estimated by means if the logodds method. Median barriers increase accident rate, but reduce accident severity. Guardrails reduce both accident rate and accident severity. Crash cushions reduce accident rate and accident severity.
4. The numerical estimates of the effects of crash cushions are particularly uncertain due to methodological shortcomings of the evaluation studies. The weighted mean effects of

guardrails on accident severity are quite robust with respect to study design and the number of confounding variables controlled in each study.

5. The most important source of variation in study results is random variation in the number of accidents. This source of variation in study results is particularly important in studies relying on small accident samples.
6. Based on the studies included in the meta-analysis, the best current estimates of the effects of median barriers are a 30% increase in accident rate, a 20% reduction in the chance of sustaining a fatal injury, given an accident, and a 10% reduction in the chance of sustaining a personal injury, given an accident.
7. Guardrails reduce the chance of sustaining a fatal injury by about 45%, given that an accident has occurred. The chance of sustaining a personal injury is reduced by about 50%.

REFERENCES

Andersen, K. B. Uheldsmønsteret på almindelige 4-sporede veje. Rapport 20. Rådet for Trafiksikkerhedsforskning, 1977.

Beaton, J. L.; Field, R. N.; Moskowitz, K. Median barriers: One year's experience and further controlled full-scale tests. Highway Research Board Proceedings 41:433–468; 1962.

Billion, C. E. Effect of median barriers on driver behavior. Highway Research Board Bulletin 137:1–17; 1956.

Billion, C. E.; Parsons, N. C. Median accident study—Long Island, New York. Highway Research Board Bulletin 308;64–79; 1962.

Billion, C. E.; Taragin, A.; Cross, E. C. Effect of parkway medians on driver behavior—Westchester County parkways. Highway Research Board Bulletin 308:36–63; 1962.

Bryden, J. E.; Fortuniewicz, J. S. Performance of highway traffic barriers. In Carney, J. F., III (editor): Effectiveness of highway safety improvements. New York, NY: American Society of Civil Engineers; 1985:242–252.

Elvik, R. Metaanalyse av effektmålinger av trafikksikkerhetstiltak. TØI-rapport 232. Oslo, Transportøkonomisk institutt, 1994.

Fleiss, J. L. Statistical methods for rates and proportions. 2nd edition. New York, NY: Wiley, 1981.

Galati, J. V. Study of box-beam median barrier accidents. Highway Research Board Special Report 107, Highway Safety. Washington, DC: Highway Research Board; 1970;133–139.

Glennon, J. C.; Tamburri, T. N. Objective criteria for guardrail installation. Highway Research Record 174:184–206; 1967.

Good, M. C.; Joubert, P. N. A review of roadside objects in relation to road safety. Melbourne: University of Melbourne, Department of Mechanical Engineering; 1971 (Published by Australian Government Publishing Service, 1973, as Report no NR/12 by Expert Group on Road Safety)

Griffin, L. I. How effective are crash cushions in reducing deaths and injuries? Public Roads March: 132–134; 1984.

Hall, J. W. Guardrail installation and improvement priorities. Transportation Research Recurd 868:47–53; 1982.

Hedges, L. V.; Olkin, I. Statistical methods for meta-analysis. San Diego, CA: Academic Press; 1985.

Houh, M. Y.; Epstein, K. M.; Lee, J. Crash cushion improvement priority and performance evaluation. Transportaton Research Record 1065:87–97; 1986.

Hunter, J. E.; Schmidt, F. L. Methods of meta-analysis. Correcting error and bias in research findings. Newbury Park, CA: Sage Publications; 1990.

Hunter, W. W.; Stewart, J. R.; Council, F. M. A comparative performance study of longitudinal roadside barriers and end treatments. Preprint 23/9 Traffic safety prepared for Conference Strategic Highway (SHRP) andTraffic Safety on Two Continents, The Hague, The Netherlands, September 22–24, 1993.

Johnson, H. D. Cross-over accidents on all-purpose dual carriageways. Supplementary report 617. Crowthorne, Berkshire, U.K.: Transport and Road Research Laboratory; 1980.

Johnson, R. T. Effectiveness of Median Barriers. Highway Research Record 105:99–109; 1966.

Kurucz, C. N. An analysis of the injury reduction capabilities of breakaway light standards and various guardrails. Accident Analysis and Prevention 16:105–114; 1984.

Light, R. J.; Pillemer, D. B. Summing up. The science of reviewing research. Cambridge, MA: Harvard University Press; 1984.

Michie, J. D., Calcote, L. R., and Bronstad, M. E. Guardrail performance and design. National Cooperative Highway Research Program Report 115. Washington, DC: Highway Research Board; 1971.

Moore, R. L.; Jehu, V. J. OTA Study Week Theme II. Recent developments in barrier design. Traffic Engineering and Control 10:421–429; 1968.

Moskowitz, K.; Schaefer, W. E. California median study 1958. Highway Research Board Bulletin 266:34–62; 1960.

Perchonok, K.; Ranney, T. A.; Baum S.; Morris, D. F.; Eppich, J. D. Hazardous effects of highway features and roadside objects. Volume 2: Findings. Report FHWA-RD-78-202. Washington DC: U.S. Department of Transportation, Federal Highway Administration: 1978.

Pettersson, R. Avkörningsolyckor och vägens sidoutrymme. Etapp 2. Olycksrisk samt samband mellan skadeföljd nch utformingen av vägens sidoutrymme. VTI-rapport 127. Linköping, Statens Väg-och Trafikinstitut, 1977.

Ray, M. H.; Troxel, L. A.; Carney, J. F., III. Characteristics of fixed-roadside-object side-impact accidents. Journal of Transportation Engineering 117:281–297; 1991.

Rosenthal, R. M. Meta-analytic procedures for social research. Applied social research methods series, Volume 6. Newbury Park, CA: Sage Publications; 1991.

Rossi, P. H.; Freeman, H. E. Evaluation. A systematic approach. 3rd Edition. Beverly Hills, CA: Sage Publications; 1985.

Sacks, W. L. Effect of guardrail in a narrow median upon Pennsylvania drivers. Highway Research Record 83:114–131; 1965.

Schandersson, R. Avkörningsolyckor och vägens sidoutrymme. Etapp 3. Olyckskostnader samt beräkning av olycksrisker och olyckskostnader för objekt i sidoutrymmet. VTI-rapport 185. Linköping, Statens Väg- och Trafikinstitut, 1979.

Schoon, C. C. After seven years: RIMOB in practice. An evaluation of the Dutch impact attenuator RIMOB. SWOV Report R-90-49. Leidschendam, The Netherlands: SWOV Institute for Road Safety Research; 1990.

Schultz, L. C. Pennsylvania's guide rail standards: A cost-effective change. Transportation Research Record 1065:12–18; 1986.

Statens Vägverk. Trafiksäkerhet på vägar med midträcke. Rapport TU 143. Borlänge, Statens Vägverk, 1980.

Tamburri, T. N.; Hammer, C. J.; Glennon, J. C.; Lew, A. Evaluation of minor improvements. Highway Research Record 257:34–79, 1968.

Tye, E. J. Median barriers in California. Traffic Engineering 25:28–29; September 1975.

Viner, J. G.; Tamanini, F. J. Effective highway barriers. Accident Analysis & Prevention 5:203–214; 1973.

Williston, R. M. Motor vehicle traffic accidents: Limited access expressway system. Connecticut State Highway Department, Bureau of Traffic, Technical Report 10. 1969 (quoted in Good and Joubert 1971, op. cit.).

Woods, D. L.; Bohuslav, B.; Keese, C. J. Remedial safety treatment of narrow bridges. Traffic engineering, March: 11–16; 1976.

APPENDIX I
Data for 32 Studies of the Safety Effects of Guardrails and Crash Cushions

Author	Year	Country	Design	Confounding variables	Type of object	Type of guardrail	Accident severity	Accidents before/without	Accidents after/with	Effect on accident severity-1	Effect on accident severity-2	Effect on accident rate-1	Effect on accident rate-2
Billion	1956	USA	2131	I.ABCDE	Median	Concrete	Inj	49	54	0,598	0,575		
			2131	II.FG			PDO	82	151			1,019	1,160
			2131	I.ABCDE	Median	Concrete	Inj	36	34	1,278	1,228		
			2131	II.FG			PDO	115	85			0,549	0,625
Moskowitz, Schaefer	1960	USA	26	I.ABCDE	Median	Concrete	Fat	93	9	0,623	0,623		
			26	II.FG			Inj	1844	333	1,274	1,274		
			26				PDO	2778	385			1,470	1,470
Beaton,Field, Moskowitz	1962	USA	31	I.ABCD	Median	Wire	Inj	55	56	0,682	0,682		
			31	II.FG			PDO	65	97			1,204	1,204
			31		Median	Wire	Inj	71	105	1,163	1,163		
			31				PDO	114	145			1,258	1,258
			31		Median	W-beam	Inj	26	41	1,147	1,147		
			31				PDO	48	66			1,400	1,400
			31		Median	W-beam	Inj	158	155	0,742	0,742		
			31				PDO	84	111			1,052	1,052
			27	I.-	W-beam-	Wire-	Inj	41	56	0,929	0,929		
			27	II.FG	guardrail	guardrail	PDO	66	97				
			27		W-beam-	Wire-	Inj	95	105	0,922	0,922		
			27		guardrail	guardrail	PDO	121	145				
Billion, Parsons	1962	USA	26	I.ABCDE	Median	W-beam	Fat	10	2	1,163	1,163		
			26	II.FG			Inj	597	120	1,360	1,360		
			26				PDO	717	106			1,093	1,093

(continued)

(*Appendix A continued*)

Author	Year	Country	Design	Confounding variables	Type of object	Type of guardrail	Accident severity	Accidents before/without	Accidents after/with	Effect on accident severity-1	Effect on accident severity-2	Effect on accident rate-1	Effect on accident rate-2
Billion, Taragin, Cross	1962	USA	31	I.ABCD	Median	W-beam	Inj	42	11	0,124	0,124		
			31	II.FG			PDO	144	304			6,738	6,738
			26	I.ABCDE	Median	W-beam	Inj	4	2	4,000	4,000		
			26	II.FG			PDO	24	3			0,667	0,667
Sacks	1965	USA	31	I.ABCD	Median	W-beam	Fat	1	0,5	0,282	0,282		
			31	II.FG			Inj	17	29	0,904	0,904		
			31				PDO	32	58			1,640	1,640
			31	I.ABCD	Median	W-beam	Fat	6	1	0,118	0,118		
			31	II.FG			Inj	82	100	0,769	0,769		
			31				PDO	199	297			1,280	1,280
Johnson	1966	USA	31	I.ABCD	Median	Wire	Fat	31	21	0,477	0,477		
			31	II.FG			Inj	682	883	0,834	0,834		
			31				PDO	873	1327			1,316	1,316
			31	I.ABCD	Median	W-beam	Fat	31	27	0,701	0,701		
			31	II.FG			Inj	1173	1585	1,158	1,158		
			31				PDO	1488	1718			1,200	1,200
			27	I.-	W-beam-	Wire-	Fat	27	21	1,162	1,162		
			27	II.FG	guardrail	guardrail	Inj	1585	883	0,726	0,726		
			27				PDO	1718	1327				
Glennon, Tamburri	1967	USA	26	I.-	Embankment	W-beam	Fat	0,5	14	6,360	6,360		
			26	II.FG	H=1		Inj	30	147	1,304	1,304		
			26		V>1,5:1(-)		PDO	42	170				
			26		Embankment	W-beam	Fat	0,5	14	1,855	1,855		
			26		H=1		Inj	11	147	0,824	0,824		
			26		V<1,5:1(+)		PDO	10	170				
			26		Embankment	W-beam	Fat	2	14	1,899	1,899		
			26		H=1,5-3		Inj	45	147	0,626	0,826		
			26		V>1,5:1(-)		PDO	41	170				

(*continued*)

(*Appendix A continued*)

Author	Year	Country	Design	Confounding variables	Type of object	Type of guardrail	Accident severity	Accidents before/without	Accidents after/with	Effect on accident severity-1	Effect on accident severity-2	Effect on accident rate-1	Effect on accident rate-2
			26		Embankment	W-beam	Fat	3	14	1,369	1,369		
			26		H=1,5-3		Inj	61	147	0,474	0,474		
			26		V<1,5:1(+)		PDO	32	170				
			26		Embankment	W-beam	Fat	1	14	0,574	0,574		
			26		H=3-6		Inj	7	147	0,710	0,710		
			26		V>2:1(-)		PDO	6	170				
			26		Embankment	W-beam	Fat	8	14	1,413	1,413		
			26		H=3-6		Inj	162	147	0,524	0,524		
			26		V<2:1(+)		PDO	94	170				
			26		Embankment	W-beam	Fat	10	14	0,667	0,667		
			26		H=6-9		Inj	96	147	0,491	0,491		
			26		V<2:1(+)		PDO	55	170				
			26		Embankment	W-beam	Fat	3	14	0,942	0,942		
			26		H=9-12		Inj	47	147	0,322	0,322		
			26		V<2:1(+)		PDO	17	170				
			26		Embankment	W-beam	Fat	2	14	0,927	0,927		
			26		H=12-15		Inj	31	147	0,316	0,316		
			26		V<2:1(+)		PDO	11	170				
			26		V<2:1(+)		PDO	11	170				
			26		Embankment	W-beam	Fat	5	14	0,424	0,424		
			26		H=15-21		Inj	36	147	0,277	0,277		
			26		V<2:1(+)		PDO	12	170				
			26		Embankment	W-beam	Fat	2	14	0,839	0,839		
			26		H=21-31		Inj	30	147	0,237	0,237		
			26		V<2:1(+)		PDO	8	170				
			26		Embankment	W-beam	Fat	1	14	1,281	1,281		
			26		H=31-46		Inj	24	147	0,189	0,189		
			26		V<2:1(+)		PDO	5	170				
			26		Embankment	W-beam	Fat	4	14	0,221	0,221		

(*continued*)

(*Appendix A continued*)

Author	Year	Country	Design	Confounding variables	Type of object	Type of guardrail	Accident severity	Accidents before/without	Accidents after/with	Effect on accident severity-1	Effect on accident severity-2	Effect on accident rate-1	Effect on accident rate-2
			26		H=46-61		Inj	17	147	0,135	0,135		
			26		V<2:1(+)		PDO	3	170				
			26		Embankment	W-beam	Fat	6	14	0,103	0,103		
			26		H=61-150		Inj	13	147	0,050	0,050		
			26		V<2:1(+)		PDO	1	170				
			26	I.ABCDE	Bridge rail	W-beam	Fat	19	16	0,225	0,225		
			26	II.FG	ends		Inj	79	191	0,265	0,265		
			26				PDO	25	199			1,163	1,163
			26	I.ABCDE	Bridge pier	W-beam	Fat	51	8	0,593	0,593		
			26	II.FG			Inj	183	35	0,396	0,396		
			26				PDO	59	28			0,640	0,640
			26	I.ABCDE	Utility pole	W-beam	Fat	26	1	0,754	0,754		
			26	II.FG			Inj	401	23	1,319	1,319		
			26				PDO	305	13			4,098	4,098
			26	I.ABCDE	Highway sign	W-beam	Fat	11	1	0,350	0,350		
			26	II.FG			Inj	112	35	1,417	1,417		
			26				PDO	146	31			0,391	0,391
			26	I.ABCDE	Highway sign	W-beam	Fat	7	15	0,281	0,281		
			26	II.FG			Inj	27	220	1,013	1,013		
			26				PDO	17	116			0,256	0,256
Moore,Jehu	1968	GB	2131	I.ABCDE	Median	W-beam	Fat	12	23	0,790	0,785		
			2131	II.FG			Inj	127	272	0,775	0,848		
			2131				PDO	115	315			1,111	1,163
Tamburri,Hammer,Glennon,Lew	1968	USA	31	I.ABCDE	Embankment	W-beam	Fat	1	1	1,864	1,864		
			31	II.FG	(curves)		Inj	23	10	0,686	0,686		
			31				PDO	18	12			0,450	0,450
			31	I.ABCDE	Bridge ends	W-beam	Fat	7	1	0,607	0,607		
			31	II.FG			Inj	31	5	0,451	0,451		
			31				PDO	20	7			0,453	0,453

(*continued*)

(Appendix A continued)

Author	Year	Country	Design	Confounding variables	Type of object	Type of guardrail	Accident severity	Accidents before/without	Accidents after/with	Effect on accident severity-1	Effect on accident severity-2	Effect on accident rate-1	Effect on accident rate-2
Williston [Good,Joubert]	1969	USA	26	I.-	Utility pole	All	Fat	15	25	0,277	0,277		
			26	II.FG		(W-beam)	Inj	208	753	0,402	0,402		
			26				PDO	186	1615				
			26	I.-	Tre	All	Fat	16	25	0,118	0,118		
			26	II.FG		(W-beam)	Inj	102	753	0,314	0,314		
			26				PDO	77	1615				
			26	I.-	Highway sign	All	Fat	7	25	0,270	0,270		
			26	II.FG		(W-beam)	Inj	59	753	0,876	0,876		
			26				PDO	120	1615				
			26	I.-	Bridge	All	Fat	9	25	0,188	0,188		
			26	II.FG	guardrail	(W-beam)	Inj	65	753	0,618	0,618		
			26				PDO	95	1615				
			26	I.-	Bridge pier	All	Fat	26	25	0,049	0,049		
			26	II.FG		(W-beam)	Inj	69	753	0,259	0,259		
			26				PDO	51	1615				
			26	I.-	Ditch	All	Fat	6	25	0,290	0,290		
			26	II.FG		(W-beam)	Inj	90	753	0,376	0,376		
			26				PDO	75	1615				
			26	I.-	Median	All	Fat	11	14	0,894	0,894		
			26	II.FG		(W-beam)	Inj	190	257	0,902	0,902		
			26				PDO	190	284				
Galati	1970	USA	31	I.ABCD	Median	W-beam	Fat	2	2	0,868	0,868		
			31	II.FG			Inj	39	31	0,537	0,537		
			31				PDO	40	60			1,098	1,098
Good,Joubert	1971	GB	2131	I.ABCDE	Median	W-beam	Inj	20	21	0,840	0,909		
			2131	II.FG			PDO	16	20			0,978	1,001
Good,Joubert	1971	AUS	26	I.B	Embankment	All	Fat	29	0,5	0,360	0,360		
			26	II.G			Inj	691	22	0,502	0,502		
			26				PDO	771	48				

(continued)

(*Appendix A continued*)

Author	Year	Country	Design	Confounding variables	Type of object	Type of guardrail	Accident severity	Accidents before/without	Accidents after/with	Effect on accident severity-1	Effect on accident severity-2	Effect on accident rate-1	Effect on accident rate-2
			26	I.B	Tree	All	Fat	13	0,5	0,398	0,398		
			26	II.G			Inj	270	22	0,752	0,752		
			26				PDO	454	48				
	1971	AUS	26	I.I	Utility pole	All	Fat	130	14	0,424	0,424		
			26	II.G			Inj	2261	376	0,438	0,438		
			26				PDO	1660	626				
			26	I.I	Tree	All	Fat	13	14	1,212	1,212		
			26	II.G			Inj	575	376	0,586	0,586		
			26				PDO	553	626				
			26	I.I	Highway sign	All	Fat	10	14	0,601	0,601		
			26	II.G			Inj	114	376	1,588	1,588		
			26				PDO	316	626				
			26	I.I	Embankment	All	Fat	128	14	1,091	1,091		
			26	II.G	(all)		Inj	4429	376	0,761	0,761		
			26				PDO	5565	626				
Viner,Tamanini	1973	USA	27	I.-	Exit ramp	Crash cushion	Fat	1	1	1,000	1,000		
			27	II.FGI			Inj	29	22	0,716	0,716		
			27				PDO	99	106				
Tye	1975	USA	26	I.ABCD	Concrete	W-beam	Fat	3	18	2,096	2,096		
			26	II.FG	guardrail	guardrail	Inj	216	682	1,177	1,177		
			26				PDO	431	1170			0,970	0,970
			26	I.ABCD	W-beam	Wire	Fat	18	36	0,912	0,912		
			26	II.FG	guardrail	guardrail	Inj	682	1064	0,613	0,613		
			26				PDO	1170	2998			1,566	1,566
Woods,Bohuslav,Keese	1976	USA	31	I.ABCD	Bridge ends	W-beam	All	20	4			0,168	0,168
Andersen	1977	DK	2131	I.ABCDE	Median	W-beam	Inj	124	113			0,652	1,038

(*continued*)

(*Appendix A continued*)

Author	Year	Country	Design	Confounding variables	Type of object	Type of guardrail	Accident severity	Accidents before/without	Accidents after/with	Effect on accident severity-1	Effect on accident severity-2	Effect on accident rate-1	Effect on accident rate-2
Petterson	1977	S	26	I.-	Rock side	W-beam	Fat	13	21	0,577	0,577		
			26	II.BFG			Inj	212	342	0,346	0,346		
			26				PDO	135	630				
			26	I.-	Tree	W-beam	Fat	83	21	0,305	0,305		
			26	II.BFG			Inj	726	342	0,316	0,316		
			26				PDO	444	630				
			26	I.-	Utility pole	W-beam	Fat	23	21	0,716	0,716		
			26	II.BFG			Inj	257	342	1,039	1,039		
			26				PDO	505	630				
			26	I.-	Highway sign	W-beam	Fat	0,5	21	19,963	19,963		
			26	II.BFG			Inj	48	342	4,918	4,918		
			26				PDO	414	630				
			26	I.-	Embankment	W-beam	Inj	20	363	1,469	1,469		
			26	II.BFG	>2,7:1 H<2		PDO	51	630				
			26	I.-	Embankment	W-beam	Inj	25	363	1,567	1,567		
			26	II.BFG	>2,7:1 H 2-4		PDO	68	630				
			25	I.-	Embankment	W-beam	Inj	24	363	1,296	1,296		
			26	II.BFG	>2,7:1 H>4,1		PDO	54	630				
			26	I.-	Embankment	W-beam	Inj	21	363	1,866	1,866		
			28	II.BFG	2,7-1,6:1 H<2		PDO	68	630				
			26	I.-	Embankment	W-beam	Inj	28	363	1,482	1,482		
			28	II.BFG	2,7-1,6:1 H 2-4		PDO	72	630				
			26	I.-	Embankment	W-beam	Inj	31	363	0,818	0,818		
			26	II.BFG	2,7-1,6:1 H>4		PDO	44	630				
			26	I.-	Embankment	W-beam	Inj	3	363	2,497	2,497		
			26	II.BFG	1,6-1,2:1 H<2		PDO	13	630				

(*continued*)

(*Appendix A continued*)

Author	Year	Country	Design	Confounding variables	Type of object	Type of guardrail	Accident severity	Accidents before/without	Accidents after/with	Effect on accident severity-1	Effect on accident severity-2	Effect on accident rate-1	Effect on accident rate-2
			26	I.-	Embankment	W-beam	Inj	10	363	1,210	1,210		
			26	II.BFG	1,6-1,2:1 H2-4		PDO	21	630				
			26	I.-	Embankment	W-beam	Inj	4	363	0,432	0,432		
			26	II.BFG	1,6-1,2:1 H>4		PDO	3	630				
Perchonok,Ranney,Baum,Morris,Eppich	1978	USA	26	I.-	Embankment	W-beam	Fat	18	5	0,386	0,386		
			26	II.EFG			Inj	216	85	0,341	0,341		
			26				PDO	172	194				
			26	I.-	Tree	W-beam	Fat	48	5	0,231	0,231		
			26	II.EFG			Inj	405	85	0,219	0,219		
			26				PDO	214	194				
			26	I.-	Utility pole	W-beam	Fat	14	5	0,748	0,748		
			26	II.EFG			Inj	292	85	0,443	0,443		
			26				PDO	292	194				
			26	I.-	Part of bridge structure	W-beam	Fat	14	5	0,095	0,095		
			26	II.EFG			Inj	52	85	0,155	0,155		
			26				PDO	22	194				
			26	I.-	Highway sign	W-beam	Fat	1	5	1,344	1,344		
			26	II.EFG			Inj	16	85	1,610	1,610		
			26				PDO	59	194				
			26	I.-	W-beam guardrail,	W-beam guardrail,	Fat	2	1	0,443	0,443		
			26	II.BFG	steel post,	wooden post,	Inj	28	19	0,444	0,444		
			26		blocked	blocked	PDO	34	51				
			26	I.-	W-beam guardrail,	W-beam guardrail,	Fat	1	2	5,000	5,000		
			26	II.BFG	wooden post,	not blocked	Inj	19	6	0,927	0,927		
			26		blocked		PDO	51	22				

(*continued*)

(Appendix A continued)

Author	Year	Country	Design	Confounding variables	Type of object	Type of guardrail	Accident severity	Accidents before/without	Accidents after/with	Effect on accident severity-1	Effect on accident severity-2	Effect on accident rate-1	Effect on accident rate-2
			26	I.-	W-beam guardrail,	Wire,	Fat	2	0,5	0,233	0,233		
			26	II.BFG	not blocked	two or three	Inj	6	4	0,476	0,476		
			26				PDO	22	26				
Schandersson	1979	S	26	I.-	Rock side	W-beam	Fat	20	30	0,516	0,516		
			26	II.BFG			Inj	328	587	0,417	0,417		
			26				PDO	273	1160				
			26	I.-	Tree	W-beam	Fat	132	30	0,247	0,247		
			26	II.BFG			Inj	1092	587	0,352	0,352		
			26				PDO	809	1160				
			26	I.-	Utility pole	W-beam	Fat	38	30	0,692	0,692		
			26	II.BFG			Inj	486	587	1,061	1,061		
			26				PDO	1045	1160				
			26	I.-	Bridge pier	W-beam	Fat	5	30	0,543	0,543		
			26	II.BFG			Inj	33	587	1,750	1,750		
			26				PDO	125	1160				
			26	I.-	Highway sign	W-beam	Fat	1	30	16,245	16,245		
			26	II.BFG			Inj	123	587	3,530	3,530		
			26				PDO	823	1160				
Johnson	1980	GB	26	I.ABCDE	Median	W-beam	Fat	39	35	0,778	0,778		
			26	II.FG			Inj	334	359	0,860	0,860		
			26				PDO	360	442			1,140	1,140
Statens Vägverk	1980	S	37		Median	W-beam	Fat	39	2	0,783	0,783		
			37				Inj	406	14	0,416	0,416		
			37				PDO	602	52				
Hall	1982	USA	27	I.ABCD	Various	W-beam	Fat	19	12	0,530	0,530		
			27	II.FGI			Inj	65	71	0,584	0,584		
			27				PDO	13	22				

(continued)

(*Appendix A continued*)

Author	Year	Country	Design	Confounding variables	Type of object	Type of guardrail	Accident severity	Accidents before/without	Accidents after/with	Effect on accident severity-1	Effect on accident severity-2	Effect on accident rate-1	Effect on accident rate-2
Griffin	1984	USA	26	I.BCD	Bridge ends	Crash cushion	Fat	25	0,5	0,366	0,366		
			26	II.AFG			Inj	123	3	0,203	0,203		
			26				PDO	60	7				
			26	I.BCD	Bridge ends	Crash cushion	Fat	70	0,5	0,152	0,152		
			26	II.AFG			Inj	268	7	0,144	0,144		
			26				PDO	52	8				
			26	I.BCD	Bridge ends	Crash cushion	Fat	24	1	2,417	2,417		
			26	II.AFG			Inj	116	2	1,243	1,243		
			26				PDO	56	1				
			26	I.BCD	Bridge ends	Crash cushion	Fat	51	5	0,148	0,148		
			26	II.AFG			Inj	333	128	0,133	0,133		
			26				PDO	48	125				
			26	I.BCD	Bridge ends	Crash cushion	Fat	8	3	0,488	0,488		
			26	II.AFG			Inj	72	40	0,403	0,403		
			26				PDO	27	36				
			26	I.BCD	Bridge ends	Crash cushion	Fat	3	0,5	0,355	0,355		
			26	II.AFG			Inj	26	9	0,538	0,538		
			26				PDO	23	14				
			26	I.BCD	Bridge ends	Crash cushion	Fat	29	2	0,300	0,300		
			26	II.AFG			Inj	277	55	0,607	0,607		
			26				PDO	114	35				
			26	I.BCD	Bridge ends	Crash cushion	Fat	7	1	0,520	0,520		
			26	II.AFG			Inj	80	9	0,133	0,133		
			26				PDO	22	19				

(*continued*)

(*Appendix A continued*)

Author	Year	Country	Design	Confounding variables	Type of object	Type of guardrail	Accident severity	Accidents before/without	Accidents after/with	Effect on accident severity-1	Effect on accident severity-2	Effect on accident rate-1	Effect on accident rate-2
Kurucz	1984	USA	27	I.-	Fixed	W-beam	Inj	71	40	0,172	0,172		
			27	II.ACFG	objects (trees)		PDO	11	36				
Bryden,Fortuniewicz	1985	USA	26	I.-	W-beam, heavy post	W-beam, light post	Fat	18	23	0,544	0,544		
			26	II.FG			Inj	545	1132	0,695	0,695		
			26				PDO	248	732				
Houh,Epstein,Lee	1986	USA	31	I.ABCDE	Exit ramp	Crash cushion	Inj	61	11	1,307	1,307		
			31	II.FG			PDO	77	8			0,160	0,160
Schultz	1986	USA	26	I.-	Bridge ends	W-beam	Fat	56	276	0,299	0,299		
			26	II.FG			Inj	733	9168	0,482	0,482		
			26				PDO	351	8724				
			26	I.-	Tree	W-beam	Fat	366	276	0,650	0,650		
			26	II.FG			Inj	10348	9168	0,513	0,513		
			26				PDO	5076	8724				
			26	I.-	Utility pole	W-beam	Fat	318	276	1,309	1,309		
			26	II.FG			Inj	17559	9168	0,571	0,571		
			26				PDO	9435	8724				
			26	I.-	Ditch	W-beam	Fat	54	276	1,460	1,460		
			26	II.FG			Inj	3030	9168	0,730	0,730		
			26				PDO	2080	8724				
			26	I.-	Highway sign	W-beam	Fat	46	276	1,322	1,322		
			26	II.FG			Inj	2134	9168	0,982	0,982		
			26				PDO	1980	8724				
Schoon	1990	NL	26	I.-	Tree	Crash cushion	Fat	15	0,5	0,265	0,265		
			26	II.FG			Inj	97	6	0,372	0,372		
			26				PDO	205	32				

(*continued*)

(*Appendix A continued*)

Author	Year	Country	Design	Confounding variables	Type of object	Type of guardrail	Accident severity	Accidents before/without	Accidents after/with	Effect on accident severity-1	Effect on accident severity-2	Effect on accident rate-1	Effect on accident rate-2
Ray,Troxel,Carney	1991	USA	26	I.-	Tree	W-beam	Fat	784	70	0,228	0,228		
			26	II.AFGH			Inj	19665	3445	0,285	0,285		
			26				PDO	20867	12481				
			26	I.-	Utility pole	W-beam	Fat	434	70	0,360	0,360		
			26	II.AFGH			Inj	18593	3445	0,251	0,251		
			26				PDO	16969	12481				
			26	I.-	Bridge pier	W-beam	Fat	44	70	0,175	0,175		
			26	II.AFGH			Inj	344	3445	1,022	1,022		
			26				PDO	1408	12481				
			26	I.-	Highway sign	W-beam	Fat	12	70	2,544	2,544		
			26	II.AFGH			Inj	1935	3445	0,725	0,725		
			26				PDO	5011	12481				
			26	I.-	Embankment	W-beam	Fat	12	70	2,416	2,416		
			28	II.AFGH			Inj	2480	3445	0,465	0,465		
			26				PDO	4116	12481				
Hunter, Stewart,Council	1993	USA	25	I.-	Concrete	W-beam,	Fat	0,5	2	0,431	0,431		
			25	II.ABFGH		blocked	Inj	11	80	0,428	0,428		
			25				PDO	3	50				
			25	I.-	W-beam,	W-beam	Fat	2	0,5	0,602	0,602		
			25	II.ABFGH	blocked		Inj	80	27	0,621	0,621		
			25				PDO	50	27				
			25	I.-	Concrete	W-beam	Fat	2	1	1,803	1,803		
			25	II.ABFGH	(median barrier)	(median barrier)	Inj	101	28	1,014	1,014		
			25				PDO	36	10				
			25	I.-	W-beam	Wire	Fat	1	0,5	0,559	0,559		
			25	II.ABFGH	(median barrier)	(median barrier)	Inj	28	13	0,222	0,222		
			25				PDO	10	21				

APPENDIX B

Description of the Logodds Method of Meta-Analysis

Notation
The following notation will be adopted:

ACC	Number of accidents of all degrees of severity
VKM	Vehicle kilometres of travel
FAT	Number of fatal accidents
INJ	Number of injury accidents
PDO	Number of property-damage-only accidents
G	Median barrier, guardrail or crash cushion present
O_i	Odds ratio for result i
$\ln(O_i)$	Natural logarithm of odds ratio for result i
W	Statistical weight
W_i	Statistical weight for result i

Definition of each result and its statistical weight
Effect on accident rate is defined in terms of the following odds ratio:

$$O_{\text{rate}} = [ACC(G)/VKM(G)]/[ACC/VKM] \quad (1)$$

where (G) denotes the presence of a median barrier, guardrail, or crash cushion.

Effect on the probability of a fatal accident, given that an accident has occurred, is defined in terms of the following odds ratio:

$$O_{\text{fatal}} = [FAT(G)/\{INJ(G) + PDO(G)\}]/[FAT/\{INJ + PDO\}] \quad (2)$$

Effect on the probability of an injury accident, given that an accident has occurred, is defined in terms of the following odds ratio:

$$O_{\text{injury}} = [\{FAT(G) + INJ(G)\}/PDO(G)]/[\{FAT + INJ\}/PDO] \quad (3)$$

The statistical weight of each result (Fleiss, 1981) is equal to the reciprocal of the sum of the reciprocals of each of the accident figures that enter into calculation of the odds ratio. For accident rate, the statistical weight is defined as:

$$W_{\text{rate}} = 1/[ACC(G) + 1/ACC]. \quad (4)$$

For fatal accidents, the statistical weight of each result is:

$$W_{\text{fatal}} = 1/[FAT(G) + 1/\{INJ(G) + PDO(G)\} + 1/FAT + 1/\{INJ + PDO\}]. \quad (5)$$

For injury accidents, the statistical weight of each result is:

$$W_{\text{injury}} = 1/[1/\{FAT(G) + INJ(G)\} + 1/PDO(G) + 1/\{FAT + INJ\} + 1/PDO]. \quad (6)$$

In case an accident figure is 0, 0, 5 is added.

Definition of weighted mean result and variance
The weighted mean result of n studies, each of which is expressed in terms of an odds ratio, is defined as:

$$\bar{o} = \exp\left(\frac{\sum_{i=1}^{n} \ln(o_i) w_i}{\sum_{i=1}^{n} w_i}\right). \quad (7)$$

The total variance of the resuls around the weighted mean result is defined as:

$$\text{Var}(t) = \sum_{i=1}^{n} \left[(o_i - \bar{o})^2 \times \left(w_i \Big/ \sum_{i=1}^{n} w_i\right)\right]. \quad (8)$$

The variance of each result is defined as follows for accident rate, fatal accidents and injury accidents:

$$Var(O_{\text{rate}}) = (O_{\text{rate}} \times O_{\text{rate}}) \cdot [1/ACC(G) + 1/ACC]. \quad (9)$$

$$Var(O_{\text{fatal}}) = (O_{\text{fatal}} \times O_{\text{fatal}}) \cdot [1/FAT(G) + 1/\{INJ(G) + PDO(G)\} + 1/FAT + 1/\{INJ + PDO\}]. \quad (10)$$

$$Var(O_{\text{injury}}) = (O_{\text{injury}} \cdot O_{\text{injury}}) \cdot [1/\{FAT(G) + INJ(G)\} + 1/PDO(G) + 1/\{FAT + INJ\} + 1/PDO]. \quad (11)$$

The contribution of random variance to the total variance of all results is defined as:

$$Var(r) = \sum_{i=1}^{n} \left[Var_{o_i} \cdot \left[w_i \Big/ \sum_{i=1}^{n} w_i\right)\right] \quad (12)$$

Systematic variation in research results is defined as the difference between total variance and random variance:

$$Var(s) = Var(t) - Var(r) \quad (13)$$

Definition of confidence intervals for weighted mean result
The lower 95% confidence limit of the weighted mean result is defined as:

$$\bar{o}_{\text{lower}} = \exp\left(\frac{\sum_{i=1}^{n} \ln(o_i) w_i}{\sum_{i=1}^{n} w_i} - 1{,}96 \times 1 \Big/ \sqrt{\sum_{i=1}^{n} w_i}\right). \quad (14)$$

The upper 95% confidence limit of the weighted mean result is defined as:

$$\bar{o}_{\text{upper}} = \exp\left(\frac{\sum_{i=1}^{n} \ln(o_i) w_i}{\sum_{i=1}^{n} w_i} + 1{,}96 \times 1 \Big/ \sqrt{\sum_{i=1}^{n} w_i}\right). \quad (15)$$

Note that the confidence limits are not symmetrical around the weighted mean value.

Paper 2

Meta-Analysis of Evaluations of Public Lighting as Accident Countermeasure

RUNE ELVIK

A meta-analysis of 37 studies evaluating the safety effects of public lighting is reported. The 37 studies contain a total of 142 results. The studies included were reported from 1948 to 1989 in 11 different countries. The presence of publication bias was tested by the funnel graph method. It was concluded that there is no evidence of publication bias and that it makes sense to estimate a weighted mean safety effect of public lighting on the basis of the 142 individual results. This is done by the log-odds method of meta-analysis. The validity of the combined results was tested against a number of rival hypotheses. It was concluded that the results are unlikely to have been caused by regression-to-the-mean and secular accident trends. The results were robust with respect to research design, decade of study, country of study, and type of traffic environment studied. The safety effects of public lighting were, however, sensitive to accident severity and type of accident. It was concluded that the best current estimates of the safety effects of public lighting are, in rounded values, a 65 percent reduction in nighttime fatal accidents, a 30 percent reduction in nighttime injury accidents, and a 15 percent reduction in nighttime property-damage-only accidents.

Public lighting of roads is widely accepted as an effective road accident countermeasure. Numerous studies have been done to determine the effects of public lighting on the number of accidents. In a synthesis of safety research related to traffic control and roadway elements, Schwab et al. (*1*) summarized the results of research by stating that "night accidents can be substantially reduced in number and severity by the use of good road lighting." This interpretation of the evidence from evaluation studies is not accepted by Vincent (*2*). In a critical review of 29 publications on road lighting and accidents, he concludes that "All of the studies claiming statistically significant accident reductions resulting from road lighting are deficient in any or all of: site selection, types of comparison, accident measures, measures of lighting and statistical evaluation techniques."

In nonexperimental accident research numerous threats to the validity of results exist. It is rarely possible to deal with all of them in a fully satisfactory way. Most literature surveys do not discuss the threats to validity at all or treat them informally, as Vincent (*2*) did. This paper argues that some issues that arise in studies attempting to summarize and interpret evidence from a number of evaluation studies can be resolved by quantitative meta-analysis. Three issues lend themselves to treatment by quantitative meta-analysis:

1. Is it meaningful to summarize the results of a number of studies of the effects of a certain accident countermeasure into an estimate of the mean effect on safety of the countermeasure? If yes, what is the best estimate of mean safety effects?
2. Which are the most and least valid and reliable results of studies that have evaluated the effects of an accident countermeasure? How can the most valid results be identified?
3. Why do the results of different evaluation studies concerning the same countermeasure vary? What are the most important sources of variation in study results?

This paper reports the results of a quantitative meta-analysis of evaluation studies concerning the safety effects of public lighting. Those studies were designed to address the three issues raised in the preceding paragraph. The studies have evaluated the effects on safety of public lighting on any type of road, including residential streets, rural highways, and freeways and covered both rural and urban areas and lighting of intersections as well as continuous roadway segments.

EVALUATION STUDIES INCLUDED IN META-ANALYSIS

Thirty-seven studies evaluating the effects of public lighting on road safety are included in the meta-analysis. The 37 studies contained a total of 142 results concerning the effects of road lighting on road safety; these results were expressed in terms of either changes in the number of nighttime accidents or changes in the nighttime accident rate per million vehicle kilometers of travel. The studies were retrieved by a systematic literature survey. A detailed description of how the literature survey was conducted is given elsewhere (*3*). The final sample consisted of evaluation studies that satisfied the following requirements:

1. The study contained one or more numerical estimates of the effects of public road lighting on the number of accidents or the accident rate.
2. The study primarily assessed the effects of introducing lighting at unlit locations. Studies that primarily assessed the effects of changing the level of existing lighting were not included.
3. The study presented the number of accidents on which estimates of the effects of lighting were based. Studies giving only accident rates, without stating the number of accidents used to estimate those rates, were not included.
4. The study was published. Unpublished studies were not included.

In the meta-analysis each estimate of safety effect was used as the unit of analysis. A total of 142 results were included. The results that were included in the analysis are provided in a later section (see Table 4). For each result, data concerning the following variables were collected:

Institute of Transport Economics, P.O. Box 6110 Etterstad, N-0602 Oslo, Norway.

1. Author or authors of study,
2. Year of publication,
3. Country to which each result refers,
4. Study design (coded variable with seven categories),
5. Type of traffic environment studied (coded variable with three categories),
6. Type of accident studied (coded variable with five categories),
7. Accident severity (coded variable with four categories),
8. Number of nighttime accidents before or without lighting,
9. Number of nighttime accidents after or with lighting,
10. Number of daytime accidents before or without lighting,
11. Number of daytime accidents after or with lighting, and
12. Estimate of the effect of lighting on road safety.

Table 1 describes in more detail how the variables included in the analysis were coded.

In terms of study design a broad distinction can be made between various forms of before-and-after studies on the one hand and various forms of comparative studies on the other. Conforming to the language of epidemiology [see, e.g. Hennekens and Buring (*4*)], the comparative studies will be referred to as case–control studies, in which one or more lit locations constitute the cases, whereas one or more unlit locations constitute the controls. The two main groups of research design differ in terms of the criterion of safety (CS) effect generally adopted. In before-and-after studies the basic CS effect is the odds ratio, commonly defined as

CS effect = (no. of nighttime accidents after/no. of nighttime accidents before)
÷ (no. of daytime accidents after/no. of daytime accidents before)

If this ratio is less than 1.0 lighting reduces the number of nighttime accidents. If it is more than 1.0 lighting increases the number of nighttime accidents. In some before-and-after studies, as well as in all case–control studies, the odds ratio is expressed in terms of accident rates rather than the number of accidents. If the introduction of public lighting does not affect exposure, the odds ratio of accident rates will be identical to the odds ratio of accident frequencies. The comparability of the two measures of safety effect is discussed in a subsequent section of the paper.

TECHNIQUES OF META-ANALYSIS

Meta-analysis can be done by several techniques (*5–9*). The simplest kind of meta-analysis is the vote counting method, which consists of compiling a frequency distribution of results by safety effect. A vote count of the 142 results concerning the safety effects of road lighting included in the present study shows that 115 results (81 percent) indicate that safety has improved and 27 results (19 percent) indicate that safety has deteriorated. Since the majority of results indicate that safety has improved, it is concluded that road lighting is likely to improve safety in most cases.

TABLE 1 Variables Included in Meta-Analysis

Variable	*Categories of the variable*
Author	Listed alphabetically
Year of publication	1948 through 1989
Country of origin	11 different countries represented
Study design	(1) 22 = before-and-after study with nighttime accidents on unlighted road sections as comparison group (2) 23 = before-and-after study with daytime accidents as comparison group (3) 2223 = before-and-after study with daytime accidents as comparison group and an additional comparison group of unlighted road sections (4) 2331 = before-and-after study with daytime accidents as comparison group and data on traffic volume by time of day before and after lighting (5) 26 = case-control study where comparisons between cases and controls are stratified according to one or more confounding variables (6) 27 = case-control study where cases and control have been matched according to one or more confounding variables (7) 33 = simple case-control study; cases and controls are compared directly with no control for confounding variables
Traffic environment	(1) Urb = urban; (2) Rur = rural; (3) Mwy = Motorway (freeway)
Type of accident	(1) All = all accidents; (2) Ped = pedestrian accidents; (3) Veh = accidents involving just vehicles; (4) Junc = accidents at junctions; (5) Sec = accidents between junctions
Accident severity	(1) Du = Fatal accidents; (2) Psu = injury accidents, (3) Msu = property-damage-only accidents (4) All = accidents of unspecified severity; all accidents included
Number of accidents	Recorded directly, in the following four categories: (1) NL = nighttime, lit road; (2) = NU = nighttime, unlit road; (3) DL = daytime, lit road; (4) DU = daytime, unlit road
Effect of lighting	Defined in terms of the odds ratio = O = (NL/NU)/(DL/DU), which may be equivalently expressed in terms of accident rates (number of accidents per million vehicle kilometres of travel)

A simple vote count is, however, not very informative. A refinement of the vote counting method consists of grouping results according to their statistical significance. Applied to the 142 results concerning the safety effects of road lighting, this version of the vote counting method shows that 45 results indicated a statistically significant safety improvement at the 5 percent level of significance. Ninety-seven results did not show any statistically significant changes in safety at this level of significance (*5*). This result illustrates the point raised by Hauer (*10*) about the danger of relying on tests of statistical significance alone in summarizing the results of several evaluations of a safety measure. Evidence of safety effects typically comes in small doses that are not always statistically significant. When a large number of studies are put together, however, their combined evidence can be very strong indeed.

The basic idea in more sophisticated techniques of meta-analysis is to combine statistically the evidence from several studies by computing a weighted mean result. Weighting can be done by several techniques, depending on the statistical properties of the results that are combined. In the present study the log-odds method described by Fleiss (*5*) was used.

Once a method for combining the results of different studies has been chosen, it is possible to study the effects of several variables on the combined result of case studies. Does, for example, the combined safety effect of public lighting vary according to the research design used in different studies? In meta-analysis this question can be answered by defining a variable describing study design (Table 1), combining evidence from all studies that use the same design, and comparing the combined evidence from studies that use different designs. In this paper the effects of several variables on the results of evaluation studies have been analyzed in this manner.

IS THERE A GENERAL EFFECT OF PUBLIC LIGHTING ON ROAD SAFETY?

Vincent (*2*) argues that it does not make sense to estimate a mean safety effect of public lighting, because the locations studied have not been sampled at random from a known sampling frame. Besides, the safety effect of public lighting is likely to vary substantially from one case to another, depending, inter alia, on luminance levels, traffic environment, and predominant type of accident at the location. In meta-analysis three requirements must be fulfilled for a weighted mean estimate of safety effect to make sense: (a) there should not be publication bias, (b) the assumption that all results belong to a distribution having a well-defined mean value should be reasonably well supported, and (c) all studies should use comparable measures of safety effect.

Testing for Publication Bias

The term *publication bias* refers to the tendency not to publish results that are unwanted or believed not to be useful, for example, because they show an increase in accidents or because they are not statistically significant (*6*).

Light and Pillemer (*6*) have developed a graphical technique of testing for publication bias called the *funnel graph method*. It relies on visual inspection of a diagram in which each study result is plotted in a coordinate system. The horizontal axis shows each result. The vertical axis shows the sample size on which each result is based. The idea is that if there is no publication bias the scatter plot of study results should resemble the form of a funnel turned upside down. The dispersion of points in the diagram should narrow as sample size increases, since large sample sizes provide more precise estimates of effects than small sample sizes. If the tails of the scatter plot are symmetrical and the density of points is the same in all areas of the diagram, this indicates that there is no publication bias.

Figures 1 to 4 show funnel graph diagrams of study results for studies of the effects of public lighting on fatal accidents (Figure 1), injury accidents (Figure 2), property-damage-only accidents (Figure 3), and accidents of unspecified severity (Figure 4). The latter category presumably includes accidents at all levels of severity. Statistical weight is used as a measure of sample size. The statistical weight of a result is proportional to the inverse of the variance of that result. For example, for a result based on 45 (dark, before), 25 (dark, after), 90 (day, before) and 85 (day, after) accidents, the statistical weight is $1/(1/45 + 1/25 + 1/90 + 1/85)$. Accidents of different degrees of severity were treated separately, because both safety effects and sample sizes are likely to differ across severity levels.

Inspection of Figures 1 to 4 does not give any indication of a clear publication bias. There is, however, a considerable amount of spread in the results. This indicates that statistically aggregating the results in terms of a weighted mean estimate of safety effect may be problematic.

Is There a True Mean Safety Effect?

The shape of scatter plots in funnel graph diagrams indicates if it makes sense to estimate a weighted mean safety effect. If the funnel graph is bimodal (has two humps) or multimodal or if there is no clear pattern in the scatter plot, a weighted mean will not be very informative. If a funnel pattern is clearly visible, estimating a weighted mean safety effect will be informative and will indicate the size of the effect that studies tend to converge to as sample size increases. In Figures 1 to 4 the funnel pattern is visible and a weighted mean value of the safety effects of lighting has been estimated.

In addition, Fleiss (*5*) describes a formal test of the homogeneity of the results. This test indicates that the results referring to fatal accidents and property-damage-only accidents are homogeneous, whereas there is a statistically significant heterogeneity in the results referring to injury accidents and accidents of unspecified severity. It was nevertheless decided to combine evidence from the various studies referring to injury accidents and accidents of unspecified severity to explore some of the sources of heterogeneity in the results.

Comparability of Measures of Effect

As pointed out earlier two measures of safety effect have been used in studies evaluating the safety effects of public lighting: changes in the odds ratio based on the number of accidents and changes in the odds ratio based on accident rates. In the funnel graph diagrams these two measures of safety effect have been mixed, relying on the assumption that neither the total amount of exposure nor its distribution between daytime and nighttime is affected by road lighting.

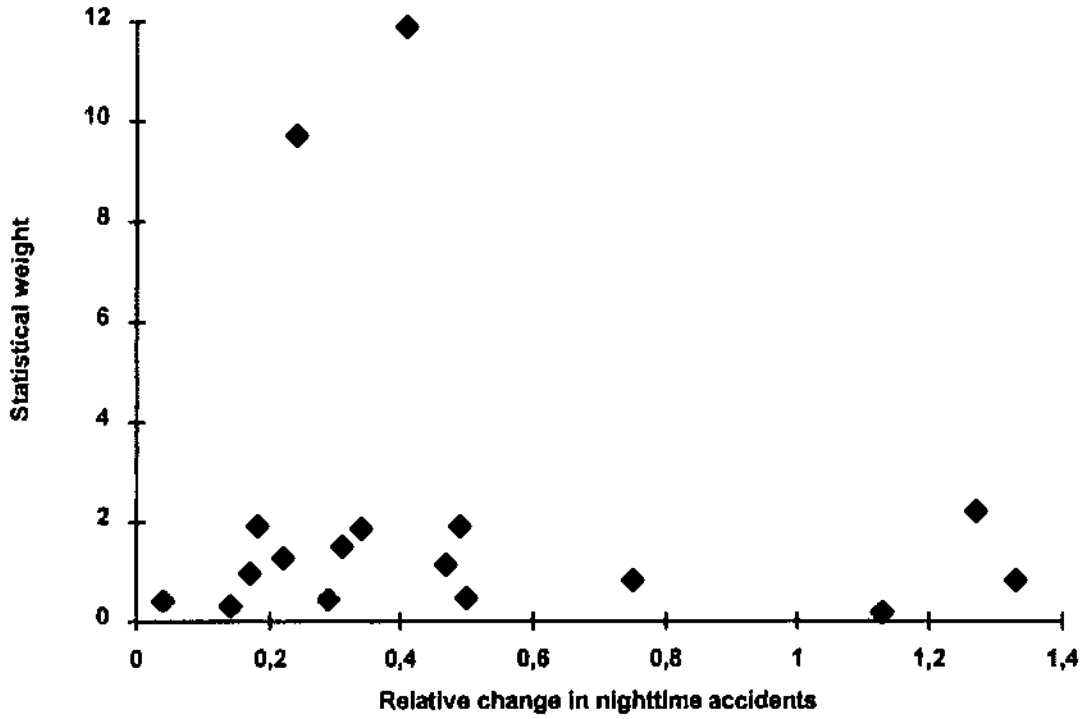


FIGURE 1 Funnel graph diagram for fatal accidents.

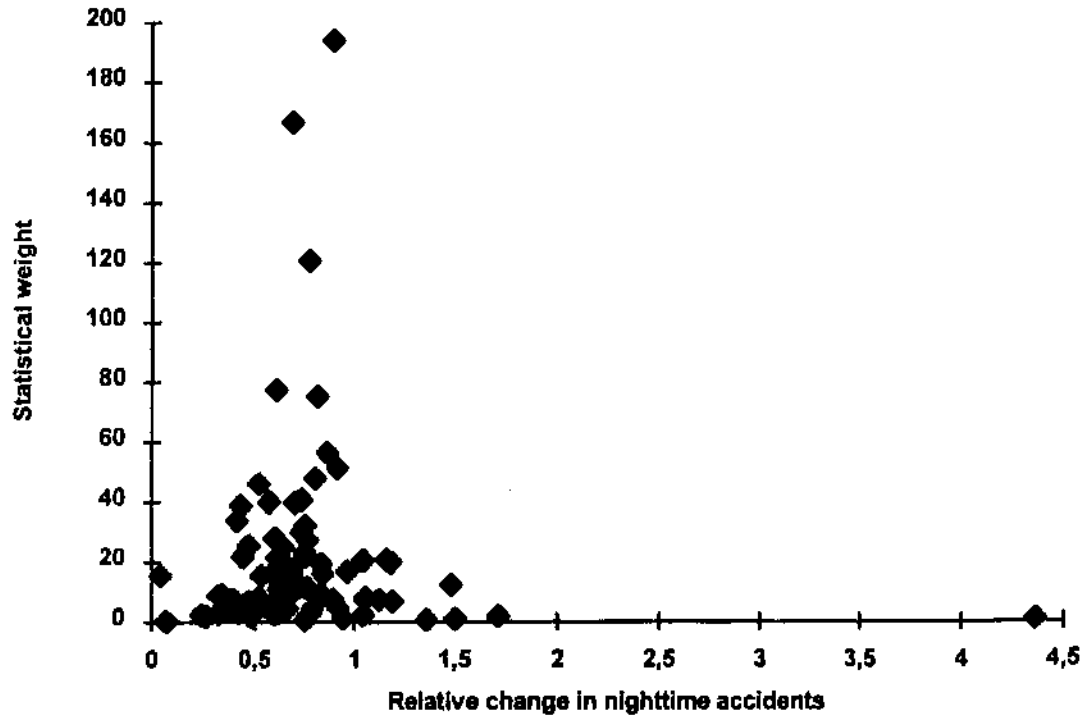


FIGURE 2 Funnel graph diagram for injury accidents.

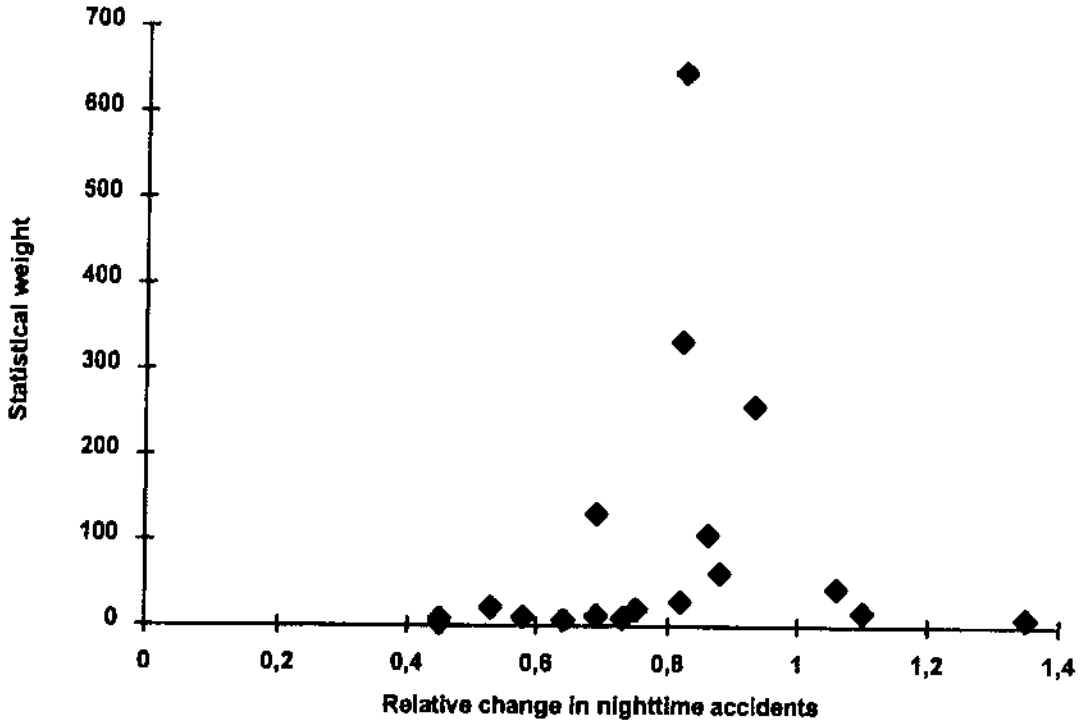


FIGURE 3 Funnel graph diagram for property-damage-only accidents.

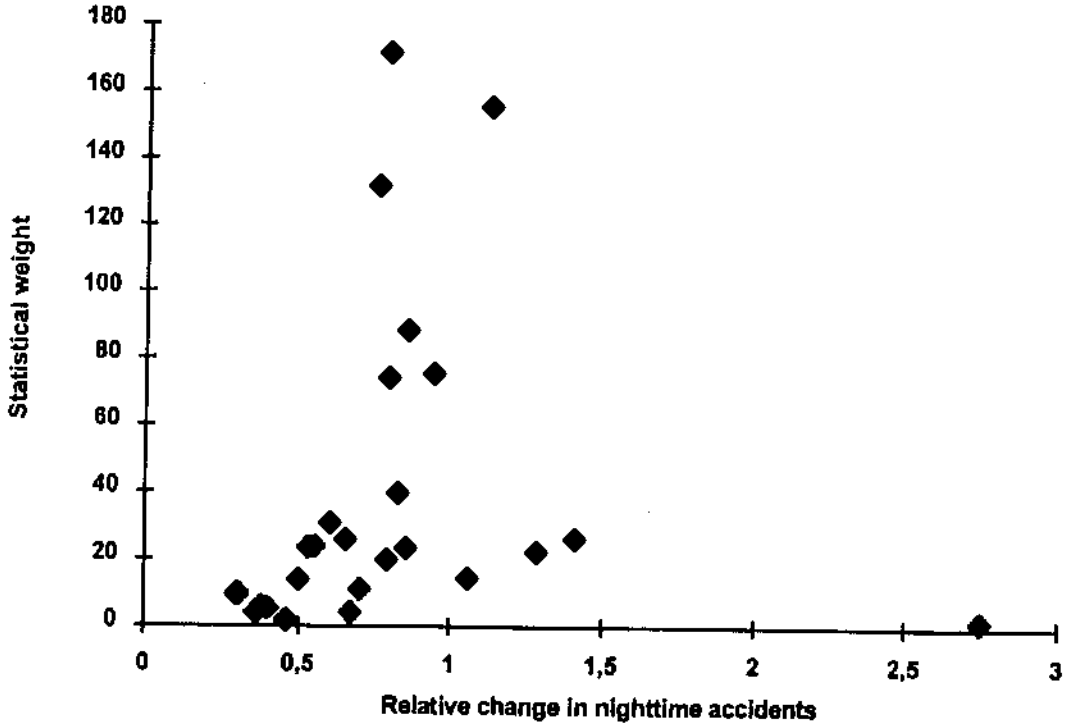


FIGURE 4 Funnel graph diagram for accidents of unspecified severity.

The validity of this assumption can be tested by relying on before-and-after studies in which both measures of safety effect can be estimated and compared. This can be done in all before-and-after studies in which exposure data are available for both the before and after periods. The studies of Tamburri, et al. (*11*), Box (*12*), Lipinski and Wortman (*13*), Walker and Roberts (*14*), Jørgensen (*15*), and Lamm et al. (*16*) allow this kind of comparison to be made. The combined estimate of the safety effect of lighting, based on these studies and measured by means of the number of accidents, is a 30 percent reduction in the number of nighttime accidents (lower 95 percent confidence limit, 21 percent reduction; upper limit, 38 percent reduction). If the safety effect is measured by means of accident rates, the combined estimate is a 33 percent reduction in nighttime accident rate (lower 95 percent confidence limit, 25 percent reduction; upper limit, 41 percent reduction). These values are very close to each other. It is concluded that changes in accident rates and changes in accident frequency can be interpreted as equivalent measures of the changes to be expected in the number of accidents with the introduction of road lighting.

VALIDITY OF EVALUATIONS OF PUBLIC LIGHTING

All of the evaluation studies included in this meta-analysis are nonexperimental. In this section, a number of threats to the validity of these studies will be discussed, including

1. Regression to the mean,
2. Secular accident trends, and
3. Contextual confounding variables.

Regression to the Mean

The most common research design in evaluation studies concerning the safety effects of public lighting is a before-and-after design, in which nighttime accidents form the experimental group and daytime accidents are used as a comparison group. In this kind of research design, regression to the mean (*17,18*) may jeopardize the validity of the results. In particular, if road lighting is introduced because of an abnormally high recorded number of accidents in the before period, a subsequent decline in the number of accidents must be expected even if lighting has no effect.

The use of daytime accidents as a comparison group in before-and-after studies will take care of the regression-to-the mean effect, provided that this effect affects daytime accidents to the same extent as nighttime accidents. This is not likely to be the case if road lighting was introduced because an abnormally high proportion of all accidents occurred in darkness. In that case one might expect the percent decline in nighttime accidents because of regression to the mean to be greater than the corresponding percent decline in daytime accidents, thus creating an apparent effect of road lighting.

On the other hand, a high percentage of nighttime accidents could indicate a real problem. In that case one would expect the true effect of road lighting to be greater when the percentage of nighttime accidents is high than when it is low. By juxtaposing the results of before-and-after studies and case–control studies made at locations with various percentages of nighttime accidents, it is possible to get an indication of whether a greater effect of road lighting at locations with a high percentage of nighttime accidents reflects regression to the mean or a genuine accident problem in the darkness.

If the regression-to-the-mean hypothesis is correct, one would expect the apparent effect of lighting to vary according to the percentage of all accidents occurring at night in before-and-after studies but not in case–control studies. If the real-darkness-problem hypothesis is correct, one would expect the effect of road lighting to vary according to the percentage of all accidents occurring at night in both before-and-after and case–control studies.

Table 2 presents data that are relevant for the two hypotheses. Study locations have been grouped according to the percentage of all accidents occurring at night (in the before period in before-and-after studies). In both before-and-after studies and case–control studies the effect of road lighting on the number of nighttime accidents is found to be greater at locations where more than 50 percent of all accidents occur at night than at locations where fewer than 50 percent of all accidents occur at night. This result weakens the regression-to-the-mean hypothesis and strengthens the real-darkness-problem hypothesis. However, the validity of the assumptions underlying the comparison cannot be tested directly. Hence, the comparison is just an indication, not a stringent test.

Secular Accident Trends

Over time the percentage of all accidents occurring at night may change. Changes in traffic distribution by hour of the day, improved vehicle headlights, and changes in the driver population are some of the factors that could generate such changes. In before-and-after studies with just one before period and just one after period and no comparison group consisting of locations where road lighting was

TABLE 2 Results of Before-and-After Studies and Case–Control Studies by Proportion of Nighttime Accidents: Weighted Mean Effect of Public Lighting on Nighttime Accidents

Study design	Percentage of accidents at night	Proportion of statistical weights	Per cent change in nighttime accidents		
			Lower 95%	Best estimate	Upper 95%
Before-and-after designs (cf table 1)	> 50%	0.089	-28	-35	-41
	33-50%	0.326	-17	-21	-25
	< 33%	0.231	-17	-22	-26
Case-control designs (cf table 1)	> 50%	0.071	-24	-32	-39
	33-50%	0.136	-7	-15	-21
	< 33%	0.147	-14	-21	-27
All designs	All	1.000	-20	-23	-25

TABLE 3 Weighted Mean Effect of Public Lighting on Nighttime Accidents According to Potential Confounding Variables

Variable	Category	Proportion of statistical weights	Per cent change in nighttime accidents		
			Lower 95%	Best estimate	Upper 95%
Accident severity	(1) Fatal	0.008	-52	-65	-75
	(2) Injury	0.387	-26	-29	-32
	(3) PDO	0.381	-13	-17	-21
	(4) Unspecified	0.224	-13	-18	-23
Study design (cf table 1 for fuller description)	(A) Fatal accs				
	(2) Design 23	0.798	-48	-63	-74
	(3) Design 2223	0.161	-40	-73	-88
	(5) Design 26	0.041	+95	-59	-91
	(B) Injury accs				
	(1) Design 22	0.036	-5	-26	-32
	(2) Design 23	0.526	-25	-30	-34
	(3) Design 2223	0.080	-16	-29	-39
	(4) Design 2331	0.007	-32	-60	-77
	(5) Design 26	0.154	-17	-26	-35
	(6) Design 27	0.044	-24	-39	-51
	(7) Design 33	0.153	-15	-25	-34
	(C) PDO accs				
	(2) Design 23	0.868	-11	-16	-20
	(4) Design 2331	0.008	+35	-19	-51
	(5) Design 26	0.038	+9	-15	-33
	(7) Design 33	0.086	-17	-30	-40
	(D) Unspec accs				
	(2) Design 23	0.024	-25	-50	-66
	(4) Design 2331	0.217	-18	-29	-37
	(5) Design 26	0.593	-1	-8	-15
	(6) Design 27	0.166	-17	-28	-38
Decade of publication	(1) 1940s	0.125	-8	-15	-22
	(2) 1950s	0.052	-21	-30	-39
	(3) 1960s	0.174	-14	-19	-25
	(4) 1970s	0.523	-19	-22	-26
	(5) 1980s	0.126	-25	-31	-37
Country	(1) Australia	0.198	-14	-19	-25
	(2) Denmark	0.024	-0	-17	-31
	(3) Finland	0.015	-1	-22	-38
	(4) France	0.017	-24	-39	-51
	(5) Germany	0.010	+1	-24	-43
	(6) Great Britain	0.123	-27	-32	-38
	(7) Israel	0.003	-8	-46	-68
	(8) Japan	0.005	-32	-56	-71
	(9) Sweden	0.063	-14	-24	-32
	(10) Switzerland	0.015	+0	-21	-38
	(11) United States	0.527	-17	-20	-23
Traffic environment	(1) Urban	0.593	-19	-22	-25
	(2) Rural	0.117	-19	-26	-32
	(3) Motorways	0.290	-20	-23	-25
Type of accident	(1) Not stated	0.478	-18	-21	-24
	(2) Pedestrian	0.045	-45	-52	-58
	(3) Vehicles only	0.312	-13	-17	-21
	(4) Junctions	0.125	-24	-30	-36
	(5) Midblocks	0.040	-0	-14	-25
All	All	1.000	-20	-23	-25

Note: The statistical weights sum to 1.000 for each variable (each severity level for the variable study design)

not introduced, the possibility that secular accident trends are confounded with the effects of road lighting cannot be ruled out. However, in all other research designs that have been used in evaluations of the safety effect of public lighting, this particular source of error can be ruled out.

Table 3 compares the results of evaluations that have relied on different research designs. With a few exceptions the weighted mean safety effect of lighting is virtually identical in all research designs. It is therefore highly unlikely that the results of before-and-after studies with only daytime accidents as a comparison group could be explained in terms of secular accident trends alone. The study results that were included in the analysis are listed in Table 4.

Contextual Confounding Variables

To what extent do variables related to study context affect the results of evaluations of the safety effects of public lighting? Table

TABLE 4 Data from 37 Studies of Safety Effects of Public Lighting

Study	Year	Country	Design	Environment	Type of accident	Accident severity	Night before/without	Night after/with	Day before/without	Day after/with	Effect
(19)	1948	USA	23	Urb	All	Du	3	1	3	2	0,500
			23	Urb	All	Psu	45	34	47	57	0,623
			23	Urb	All	Msu	201	200	324	365	0,883
			23	Urb	All	Du	17	5	10	6	0,490
			23	Urb	All	Psu	210	135	172	152	0,727
			23	Urb	All	Msu	828	789	1411	1443	0,932
			23	Urb	All	Du	8	4	3	2	0,750
			23	Urb	All	Psu	96	51	75	59	0,675
			23	Urb	All	Msu	323	340	547	672	0,857
			23	Urb	All	Psu	67	86	80	99	1,037
			23	Urb	All	Psu	173	82	126	99	0,603
			23	Urb	All	Psu	43	23	45	23	1,047
			23	Urb	All	Psu	72	28	31	36	0,335
(20)	1955	GB	23	Urb	Ped	Du	6	1	1	4	0,042
			23	Urb	Ped	Psu	31	19	73	71	0,630
			23	Urb	Kjt	Du	4	2	8	3	1,333
			23	Urb	Kjt	Psu	120	98	283	330	0,700
(21)	1958	CH	23	Urb	All	Psu	70	65	159	231	0,639
(22)	1958	GB	23	Urb	Ped	Du	15	6	5	11	0,182
			23	Urb	Ped	Psu	144	85	314	323	0,574
			23	Urb	Kjt	Du	13	9	11	6	1,269
			23	Urb	Kjt	Psu	333	303	918	1086	0,769
(23)	1960	USA	26	Mwy	All	All	52	168	71	177	1,291
(24)	1962	USA	26	Mwy	All	Psu	8	2	13	2	4,361
			26	Mwy	All	All	27	108	42	316	0,500
(25)	1962	GB	23	Mwy	All	Psu	8	7	13	19	0,599
			23	Mwy	All	Psu	41	1	71	22	0,236
(26)	1962	USA	26	Mwy	All	All	184	1004	172	997	0,943
			26	Mwy	All	All	401	1004	514	997	1,120
(27)	1965	S	23	Urb	All	Psu	14	13	41	69	0,552
			23	Urb	All	Msu	48	52	96	95	1,095
			23	Rur	All	Psu	23	15	35	42	0,543
			23	Rur	All	Msu	27	20	85	86	0,732
(28)	1966	GB	23	Urb	Ped	Psu	7	0,5	1	1	0,071
			23	Urb	Kjt	Psu	2	3	5	5	1,500
			23	Rur	All	Psu	40	22	37	39	0,522
			23	Mwy	All	Psu	82	54	123	132	0,614
(29)	1966	USA	33	Mwy	All	Psu	588	706	547	950	0,691
			33	Mwy	All	Msu	395	576	430	911	0,688
(11)	1968	USA	2331	Urb	June	All	75	27	39	39	0,304
			2331	Urb	June	All	25	11	31	34	0,396
			2331	Urb	June	All	33	13	31	34	0,377
			2331	Urb	June	All	37	15	12	12	0,355
			2331	Urb	June	All	11	5	7	8	0,455

* Number of nighttime accidents on unlit roads before and after.

(continued on next page)

3 presents results that shed light on this question for the variables (a) definition of accident severity, (b) study design, (c) decade of publication of study, (d) country where the study was performed, (e) traffic environment where the study was performed, (f) type of accident studied.

The effects of road lighting vary significantly with respect to accident severity. Nighttime fatal accidents are reduced by about 65 percent, nighttime injury accidents are reduced by about 30 percent, and nighttime property-damage-only accidents are reduced by about 15 percent. This means that studies that do not specify the severity of accidents are less informative than studies that specify accident severity. The observed weighted mean safety effect in studies of accidents of unspecified severity is an 18 percent reduction in nighttime accidents. This indicates that most of the accidents probably were property-damage-only accidents.

These results hold when controlling for study design. In general, study design appears to have a minor effect on study results. As argued earlier the robustness of the results with respect to study design indicates that the results are valid and not just the product of various confounding factors that are left uncontrolled by the various research designs. Different research designs take different confounding factors into account. Therefore, agreement of results across research designs indicates that uncontrolled confounding factors are not major sources of variation in the results of different studies.

The oldest study included was reported in 1948; the most recent was reported in 1989. Studies performed in different decades have yielded similar results. There is no indication that the safety effects of road lighting have diminished over time. Eleven different countries are represented in this analysis. Studies performed in different countries have also yielded similar results. It should be noted, however, that most studies have been performed in the United States, Great Britain, and Australia. Studies performed in other countries have been on a smaller scale, as indicated by their contribution to the statistical weights.

Three types of traffic environment have been identified: urban, rural, and freeways. The results of evaluation studies are the same for all three environments. This holds when controlling for accident severity. With respect to type of accident, studies can be divided into three groups. The first and largest group consists of studies that do not specify the types of accident studied. A second group consists of studies in which a distinction is made between pedestrian accidents and other accidents. A third group consists of studies in which a distinction is made between accidents at junctions (inter-

TABLE 4 (*continued*)

Study	Year	Country	Design	Environment	Type of accident	Accident severity	Night before/without	Night after/with	Day before/without	Day after/with	Effect
(30)	1969	USA	23	Urb	All	All	13	9	37	38	0,674
(31)	1969	USA	26	Urb	All	Du	4	14	2	15	0,468
			26	Urb	All	Psu	203	309	295	551	0,811
			26	Urb	All	Msu	83	240	220	592	1,062
(32)	1970	CH	23	Rur	All	Psu	64	77	92	94	1,178
			23	Mwy	All	Psu	10	5	25	12	1,042
			23	Mwy	All	Psu	18	5	41	27	0,422
			23	Mwy	All	Psu	4	6	25	22	1,705
			23	Urb	All	Psu	36	14	104	36	1,123
(33)	1971	AUS	23	Urb	Ped	Psu	10	6	23	18	0,767
			23	Urb	Ped	Psu	16	10	18	19	0,592
			23	Urb	Ped	Psu	15	6	16	20	0,320
			23	Urb	Ped	Psu	17	6	28	20	0,494
		USA	23	Urb	Ped	Psu	175	122	221	294	0,524
			23	Urb	Kjt	Psu	152	317	176	427	0,860
			23	Urb	Kjt	Msu	983	1674	1069	2215	0,822
			23	Urb	Ped	Du	84	22	42	46	0,239
			23	Urb	All	Du	60	38	40	62	0,409
			23	Urb	All	Psu	48	37	52	53	0,756
			23	Urb	All	All	38	30	62	70	0,699
(34)	1971	DK	2223	Urb	Ped	Psu	20	21	58	93	1,047
(12)	1972	USA	2331	Urb	All	Psu	23	10	15	17	0,384
			2331	Urb	All	Psu	52	20	30	30	0,385
			2331	Urb	All	Msu	23	4	25	12	0,448
			2331	Urb	All	Msu	53	33	75	49	0,687
			26	Mwy	Rear	All	176	198	614	863	0,794
			26	Mwy	Kjt	All	69	48	142	123	0,786
			26	Mwy	Ped	All	11	11	11	4	2,750
			26	Mwy	Off	All	102	102	132	92	1,410
			26	Mwy	All	All	356	697	888	2184	0,784
			26	Mwy	All	All	270	72	428	192	0,822
(35)	1972	GB	2223	Urb	All	Du	4	0,5	2	2	0,140
			2223	Urb	All	Psu	85	60	134	138	0,750
			2223	Rur	All	Du	11	3	5	6	0,220
			2223	Rur	All	Psu	73	56	121	145	0,620
			2223	Mwy	All	Du	8	2	4	6	0,170
			2223	Mwy	All	Psu	54	50	99	95	0,960
			2223	Urb	All	Du	1	1	0,5	0,5	1,130
			2223	Urb	All	Psu	18	9	37	32	0,660
			2223	Rur	All	Du	11	3	9	8	0,310
			2223	Rur	All	Psu	84	56	132	118	0,750
			2223	Mwy	All	Du	13	4	10	9	0,350
			2223	Mwy	All	Psu	110	75	186	175	0,730
(36)	1972	AUS	23	Urb	Ped	Psu	32	13	57	58	0,399

* Number of nighttime accidents on unlit roads before and after.

(*continued on next page*)

sections) and accidents at road sections (midblock accidents). On the basis of these classifications, road lighting appears to have a greater effect on pedestrian accidents than on other types of accidents and a greater effect at junctions than at other locations.

The general impression is that the contextual variables have a rather small impact on the results of evaluation studies. It is particularly reassuring that results are robust with respect to study design. Study decade, the country where the study was performed, and type of traffic environment hardly affect study results. On the other hand, accident severity and type of accident seem to be of some importance for study results. These variables are not directly related to study design. However, any good study should specify clearly the severity of the accidents that are studied and indicate clearly the types of accidents that are studied.

DISCUSSION OF RESULTS

The analysis presented here shows that the results of studies that have evaluated the effects of public lighting on road safety are quite robust with respect to a number of potentially confounding variables. These results cannot be dismissed as merely showing the vagaries of poor data, inadequate research design, or peculiarities of the locations that have been investigated. There is little to support the misgivings voiced by Vincent (*2*) with respect to these and related points.

On the other hand, the present analysis did not consider every conceivable source of error in previous studies. In particular, errors that may arise from an inappropriate choice of comparison groups in case–control studies or from the use of an inappropriate statistical technique in analyzing data were not considered. Most studies provide few details concerning the sampling of cases and controls. It is therefore difficult to know whether biased sampling is found and how it may have affected evaluation results. As far as statistical techniques for data analysis are concerned, most studies have relied on quite simple techniques, like estimating an odds ratio and testing it for statistical significance. More advanced multivariate analyses, in which the choice of statistical techniques is more important, are not found in this area.

The effect of public lighting on road safety was found to vary with respect to accident severity and type of accident. There are no doubt a large number of other variables with respect to which the effects of public lighting might be expected to vary. It would, for example, be of interest to know whether lighting satisfying current

TABLE 4 *Continued*

Study	Year	Country	Design	Environment	Type of accident	Accident severity	Night before/without	Night after/with	Day before/without	Day after/with	Effect
(37)	1971	GB	22	Urb	All	Psu	44	26	*12532	*8785	0,840
			22	Urb	All	Psu	23	16	*3924	*3286	0,830
			22	Rur	All	Psu	23	27	*3381	*2681	1,480
			22	Rur	All	Psu	93	35	*9027	*7245	0,470
(38)	1976	GB	23	Mwy	All	All	52	24	34	53	0,296
(13)	1976	USA	2331	Rur	June	All	356	438	656	1022	0,748
(14)	1976	USA	2331	Rur	June	All	90	46	225	207	0,551
(39)	1977	DK	33	Mwy	All	Psu	91	434	191	1006	0,905
			33	Mwy	All	Psu	91	289	191	759	0,799
(40)	1977	AUS	23	Urb	Ped	Psu	162	87	219	276	0,426
			23	Urb	Kjt	Psu	779	762	746	820	0,890
			23	Urb	Kjt	Msu	1908	1840	3854	4510	0,826
(41)	1977	JPN	23	Mwy	All	Psu	95	52	109	135	0,642
		USA	26	Mwy	All	Du	6	36	0,5	14	0,288
			26	Mwy	All	Psu	38	639	28	804	0,533
			26	Mwy	All	Msu	45	1372	41	2454	0,533
(42)	1978	SF	23	Urb	Sec	Psu	104	67	181	153	0,762
			23	Urb	Sec	Msu	112	75	187	153	0,818
			23	Urb	June	Psu	19	12	36	25	0,909
			23	Urb	June	Msu	26	15	43	39	0,636
(43)	1978	ISR	2223	Urb	Ped	Psu	79	34	77	61	0,623
(15)	1980	DK	2331	Urb	All	Psu	8	5	10	13	0,480
(44)	1981	S	26	Rur	June	Psu	58	11	90	36	0,474
			26	Rur	June	Psu	27	3	26	11	0,263
			26	Rur	June	Psu	153	34	306	82	0,829
			26	Rur	June	Psu	104	48	194	77	1,163
			26	Rur	June	Psu	19	20	58	69	0,885
			26	Rur	June	Psu	1	3	4	16	0,750
			26	Rur	June	Psu	31	13	102	36	1,188
			26	Rur	June	Psu	21	9	57	31	0,788
(1)	1982	D	23	Urb	Ped	Psu	51	19	44	51	0,321
		CH	23	Urb	Ped	Psu	34	15	52	60	0,382
		F	27	Urb	June	Psu	290	209	389	459	0,611
		USA	23	Mwy	June	Psu	76	43	83	80	0,587
(45)	1985	S	23	Rur	June	Psu	58	19	137	64	0,701
(16)	1985	D	2331	Mwy	All	All	30	77	61	148	1,062
			2331	Mwy	All	All	46	121	102	316	0,845
(46)	1986	S	27	Rur	June	All	114	63	258	236	0,604
			27	Rur	June	All	449	157	1256	517	0,849
			27	Rur	June	All	93	43	251	218	0,532
			27	Rur	June	All	219	43	390	218	0,646

* Number of nighttime accidents on unlit roads before and after.

Study	Year	Country	Design	Environment	Type of accident	Accident severity	Night before/without	Night after/with	Day before/without	Day after/with	Effect
(47)	1987	GB	26	Mwy	All	Psu	669	212	264	51	0,412
			26	Mwy	All	Psu	71	57	256	35	1,037
			26	Mwy	All	Psu	58	24	267	44	0,733
			26	Mwy	All	Psu	61	35	301	116	0,681
			26	Mwy	All	Psu	59	144	218	398	0,749
(48)	1989	USA	33	Urb	June	Psu	1	42	3	93	1,355
			33	Urb	June	Msu	15	160	19	447	0,453
			33	Urb	Sec	Psu	36	2	51	3	0,944
			33	Urb	Sec	Msu	133	19	218	23	1,356
			23	Urb	June	Psu	21	15	29	36	0,575
			23	Urb	June	Msu	51	57	117	174	0,752
			23	Urb	Sec	Psu	15	8	28	31	0,482
			23	Urb	Sec	Msu	29	23	84	114	0,584
							15879	18769	54272	57940,5	0,737

* Number of nighttime accidents on unlit roads before and after.

warrants is more effective than lighting not satisfying current warrants. However, few studies provide information concerning this. The availability of data limits the topics that can be included in a meta-analysis.

CONCLUSIONS

The following conclusions summarize the results of the research reported in this paper.

1. A meta-analysis of 37 evaluation studies of the safety effect of public lighting containing 142 results has been performed. The log-odds method was applied.

2. The presence of publication bias was tested. No evidence of publication bias was found.

3. Changes in accident rate were found to predict accurately changes in the number of accidents associated with the introduction of public lighting. These two measures of safety effect were therefore treated as equivalent in the meta-analysis.

4. The validity of research results was tested with respect to (a) regression to the mean; (b) secular accident trends; and (c) contextual confounding variables, including definition of accident severity, study design, decade of publication, country where the study was performed type of traffic environment, and type of accident studied. It was concluded that regression to the mean and secular accident trends are unlikely to have affected the results of evaluation studies materially. As far as confounding variables are concerned, accident severity and type of accident studied were found to affect study results. The other confounding variables did not affect study results.

5. The best current estimate of the safety effects of road lighting in rounded values is a 65 percent reduction in nighttime fatal accidents, a 30 percent reduction in nighttime injury accidents, and a 15 percent reduction in nighttime property-damage-only accidents.

REFERENCES

1. Schwab, R. N., N. E. Walton, J. M. Mounce, and M. J. Rosenbaum. Roadway Lighting. *Synthesis of Safety Research Related to Traffic Control and Roadway Elements*, Vol. 2. Report FHWA-TS-82-233. FHWA, U.S. Department of Transportation, 1982.
2. Vincent, T. Streetlighting and Accidents. Paper 17. In *Traffic Accident Evaluation* (D. C. Andreassend and P. G. Gipps, eds.). Papers presented at Esso-Monash Civil Engineering Workshop, Normanby House, Monash University, February 15 to 17, 1983. Department of Civil Engineering, Monash University, Australia, 1983.
3. Elvik, R. *Metaanalyse av Effektmålinger av Trafikksikkerhetstiltak*. TØI-Rapport 232. Transportøkonomisk Institute, Oslo, Norway 1994.
4. Hennekens, C. H., and J. E. Buring. *Epidemiology in Medicine*. Little, Brown & Co, Boston, 1987.
5. Fleiss, J. L. *Statistical Methods for Rates and Proportions*, 2nd ed. John Wiley and Sons, New York, 1981.
6. Light, R. J., and D. B. Pillemer. *Summing Up. The Science of Reviewing Research*. Harvard University Press, Cambridge, Mass., 1984.
7. Hedges, L. V., and I. Olkin. *Statistical Methods for Meta-Analysis*. Academic Press, San Diego, Calif., 1985.
8. Hunter, J. E., and F. L. Schmidt. *Methods of Meta-Analysis. Correcting Error and Bias in Research Findings*. Sage Publications, Newbury Park, Calif., 1990.
9. Rosenthal, R. M. Meta-Analytic Procedures for Social Research. *Applied Social Research Methods Series*, Vol. 6. Sage Publications, Newbnry Park, Calif., 1991.
10. Hauer, E. Should Stop Yield? Matters of Method in Safety Research. *ITE-Journal*, Sept. 1991, pp. 25–31.
11. Tamburri, T. N., C. J. Hammer, J. C. Glennon, and A. Lew. Evaluation of Minor Improvements. In *Highway Research Record 257*, HRB, National Research Council, Washington, D.C., 1968, pp. 34–79.
12. Box, P. C. Freeway Accidents and Illumination. In *Highway Research Record 416*, HRB, National Research Council, Washington, D.C., 1972, pp. 10–20.
13. Lipinski, M. E., and R. H. Wortman. Effect of Illumination on Rural At-Grade Intersection Accidents. In *Transportation Research Record 611*, TRB, National Research Council, Washington, D.C., 1976, pp. 25–27.
14. Walker, F. W., and S. E. Roberts. Influence of Lighting on Accident Frequency at Highway Intersections. In *Transportation Research Record 562*, TRB, National Research Council, Washington, D.C., 1976, pp. 73–78.
15. Jørgensen, E. *Eksempler på Effektstudier fra SSV*. Vejdirektoratet, Sekretariatet for Sikkerhedsfremmende Vejforanstaltninger (SSV), Næstved, 1980.
16. Lamm, R., J. H. Klöckner, and E. M. Choueiri. Freeway Lighting and Traffic Safety—A Long-Term Investigation. In *Transportation Research Record 1027*, TRB, National Research Council, Washington, D.C., 1985, pp. 57–63.
17. Hauer, E. Bias-by-Selection: Overestimation of the Effectiveness of Safety Countermeasures Caused by the Process of Selection for Treatment. *Accident Analysis and Prevention*, Vol. 12, 1980, pp. 113–117.
18. Hauer, E. On the Estimation of the Expected Number of Accidents. *Accident Analysis and Prevention*, Vol. 18, 1986, pp. 1–12.
19. Seburn, T. C. Relighting A City. *Proc., Institute of Traffic Engineers Nineteenth Annual Meeting*, 1948, pp. 58–72.
20. Tanner, J. C., and A. W. Christie. Street Lighting and Accidents—A Study of Some New Installations in the London Area. *Light and Lighting*, Vol. 48, 1955, pp. 395–397.
21. Borel, P. Accident Prevention and Public Lighting. *Bulletin des Schweitzerisches Elektrotechnisches Verbands*, Vol. 49, No. 1, 1958, pp. 8–11.
22. Tanner, J. C. Reduction of Accidents by Improved Street Lighting. *Light and Lighting*, Vol. 51, 1958, pp. 353–355.
23. Taragin, A., and B. M. Rudy. Traffic Operations as Related to Highway Illumination and Delineation. *Bulletin 255* HRB, National Research Council, Washington, D.C., 1960, pp. 1–22.
24. Billion, C. E., and N. C. Parsons. Median Accident Study—Long Island, New York. *Bulletin 308*, HRB, National Research Council, Washington, D.C., 1962, pp. 64–79.
25. Christie, A. W. Some Investigations Concerning the Lighting of Traffic Routes. *Public Lighting*, Vol. 27, 1962, pp. 189–204.
26. Ives, H. S. Does Highway Illumination Affect Accident Occurrence? *Traffic Quarterly*, Vol. 16, 1962, pp. 229–241.
27. *Väg-och Gatubelysningens Inverkan på Trafik-Säkerheten*. Meddelande 60. Transportforskningskommissionen, Stockholm, Sweden, 1965.
28. Christie, A. W. Street Lighting and Road Safety. *Traffic Engineering and Control*, Vol. 7, 1966, pp. 229–231.
29. Institute of Traffic Engineers and Illuminating Engineering Society. Joint Committee of Public Lighting. *Public Lighting Needs*. Special Report to U.S. Senate, 1966.
30. Cleveland, D. E. Illumination. *Traffic Control and Roadway Elements—Their Relationship to Highway Safety*, Revised. Automotive Safety Foundation, Washington, D.C. 1969, Chapt. 3.
31. Tennessee Valley Authority. A Study of the Benefits of Suburban Highway Lighting. *Illuminating Engineering*, April 1969, pp. 359–363.
32. Walthert, R., F. Mäder, and P. Hehlen. *Données Statistiques sur la Proportion des Accidents le Jour et la Nuit, leurs Causes et Conséquences*. La Conduite de Nuit, Automobil Club de Suisse, 1970.
33. Fisher, A. J. *A Review of Street Lighting in Relation to Road Safety*. Report 18. Australian Department of Transport, Australian Government Publishing Service, Canberra, 1971.
34. Jørgensen, N. O., and Z. Rabani. *Fodgængeres Sikkerhed i og ved Fodgænger-Overgange*. RfT-Rapport 7. Rådet for Trafiksikkerhedsforskning, Copenhagen, Denmark, 1971.
35. Cornwell, P. R., and G. M. Mackay. Lighting and Road Traffic. Part 1. Public Lighting and Road Accidents. *Traffic Engineering and Control*, Vol. 13, 1972, pp. 142–144.
36. Pegrum, B. V. The Application of Certain Traffic Management Techniques and Their Effect on Road Safety. National Road Safety Symposium, Department of Transport, Canberra, Australia 1972.
37. Sabey, B. E., and H. D. Johnson. *Road Lighting and Accidents: Before and After Studies on Trunk Road Sites*. TRRL Report LR 586. Transport and Road Research Laboratory, Crowthorne, Berkshire, United Kingdom, 1973.
38. Austin, B. R. Public Lighting—The Deadly Reckoning. *Traffic Engineering and Control*, Vol. 17, 1976, pp. 262–263.
39. Andersen, K. B. *Uheldsmønsteret på Almindelige 4-Sporede Veje*. RfT-Rapport 20. Rådet for Trafiksikkerhedsforskning, Copenhagen, Denmark, 1977.
40. Fisher, A. J. Road Lighting as an Accident Countermeasure. *Australian Road Research*, Vol. 7, No. 4, 1977, pp. 3–15.
41. Ketvirtis, A. *Road Illumination and Traffic Safety*. Road and Motor Vehicle Traffic Safety Branch, Transport Canada, Ottawa, Ontario, Canada, 1977.
42. National Board of Public Roads and Waterways. *Traffic Safety Effects of Road Lights*. Väg-och Vattenbyggnadsstyrelsen, Helsinki, Finland, 1978.
43. Polus, A., and A. Katz. An Analysis of Nighttime Pedestrian Accidents at Specially Illuminated Crosswalks. *Accident Analysis and Prevention*, Vol. 10, 1978, pp. 223–228.
44. Brüde, U., and J. Larsson. *Vägkorsningar på Landsbygd inom Huvudvägnätet. Olycksanalys*. VTI-Rapport 233. Statens Väg-och Trafikinstitut, Linköping, Sweden, 1981.
45. Brüde, U., and J. Larsson. *Korsningsåtgärder Vidtagna inom vägförvaltningarnas Trafiksäkerhetsarbete. Regressions-och åtgärdseffekter*.

VTI-Rapport 292. Statens Väg-och Trafikinstitut, Linköping, Sweden, 1985.
46. Brüde, U., and J. Larsson. *Trafiksäkerhetseffekter av Korsningsåtgärder*. VTI-Rapport 310. Statens Väg-och Trafikinstitut, Linköping, Sweden, 1986.
47. Cobb, J. Light on Motorway Accident Rates. *The Journal of the Institution of Highways and Transportation*, Oct. 1987, pp. 29–33.
48. Box, P. C. Major Road Accident Reduction by Illumination. In *Transportation Research Record 1247*, TRB, National Research Council, Washington, D.C., 1989, pp. 32–38.

Publication of this paper sponsored by Committee on Methodology for Evaluating Highway Improvements.

Paper 3

Pergamon

Accid. Anal. and Prev., Vol. 28, No. 3, pp. 339–347, 1996

0001-4575/96 $15.00 + 0.00

0001-4575(95)00073-9

DOES PRIOR KNOWLEDGE OF SAFETY EFFECT HELP TO PREDICT HOW EFFECTIVE A MEASURE WILL BE?

RUNE ELVIK

Institute of Transport Economics, PO Box 6110 Etterstad, N-0602 Oslo, Norway

(Received 28 June 1995)

Abstract—Studies evaluating the effects of traffic safety measures are often done for the purpose of predicting the effects of future applications of the measures. The predictive value of evaluation studies is unknown. Some general arguments for and against attributing a general predictive value to the results of evaluation studies are discussed. Predictability is shown to depend on many factors. Meta-analyses of evidence from evaluation studies can be used as a basis for testing the predictive performance of such studies. The predictive performance of studies that have evaluated the safety effects of road lighting and traffic separation is tested. Predictive performance is found to depend mainly on whether the results of evaluation studies are stable over time or exhibit a trend. In the latter case, predictions based on evidence accumulated before the trend became apparent can be very erroneous. It is shown that increasing the amount of evidence that predictions are based on does not necessarily make the predictions more accurate. More research does not always improve predictive performance. Copyright © 1996 Elsevier Science Ltd

Keywords—Safety measure, Evaluation, Prediction, Meta-analysis, Testing

INTRODUCTION

Evaluations of the effects of traffic safety measures and syntheses of such evaluations are normally intended for use in predicting the effects of future applications of the measures. In fact, were it not for the possible use of the results of evaluation studies for prediction, there would not be much point in doing such studies. Although it is always nice to know history, road safety research is an applied field, where knowledge is produced mainly for its presumed practical usefulness. It is therefore somewhat surprising to find that the performance of evaluation studies in predicting the safety effects of traffic safety measures has never been tested. Such tests are, admittedly, difficult. Evaluations of the effects of traffic safety measures are not done on a routine basis; nowhere is the use of the results of evaluation studies for prediction of safety effects monitored systematically; safety measures may undergo technical innovations over time and accident reporting is incomplete and may change over time. These are just a few of the problems facing anyone who wants to test the predictive performance of evaluation studies.

This paper presents an attempt to test the predictability of the effects of two widely used traffic safety measures on safety: road lighting and traffic separation. It just scratches the surface of a vastly complex problem and is nothing more than a first attempt to test the performance of evaluation studies in predicting safety effects. Before presenting the evidence, some general arguments for and against attributing predictive value to the results of studies that have evaluated the safety effects of measures will be discussed.

THE PREDICTIVE VALUE OF EVALUATION STUDIES: GENERAL DISCUSSION

Does more and better evaluation research improve the ability to predict safety effects of future applications of the measures that have been evaluated? Intuition suggests that the answer to this question is yes; surely, improving knowledge should improve the ability to predict the effects of safety measures. However, a closer consideration of some arguments for and against attributing predictability to the results of evaluation studies suggests that matters are more complex.

Some arguments supporting the idea that the results of evaluation studies have predictive value, and particularly the idea that more and better evaluation research leads to improved prediction, include:

P1: Predictability improves as estimates of safety effects become more precise; more research

leads to more precise estimates of safety effects.

P2: Predictability is related to explicability; more and better research leads to better explanations of previous research findings.

P3: The development over time of motorization rate, accident rate and the number of accidents in the highly motorized countries is strikingly similar, suggesting that the major factors affecting safety, including traffic safety measures, have similar effects everywhere.

P4: Many traffic safety measures are highly standardized with respect to both design and implementation. This is particularly true of vehicle safety measures, like seat belts or headlights, and also to a large extent traffic control measures, like traffic signals.

P5: Many traffic safety measures influence risk factors whose effects on accident occurrence are likely to be quite universal, like driving speed, driver age, visibility (especially at night), and the influence of alcohol. Their effects are therefore likely to be similar everywhere.

The validity of these arguments must be determined by means of empirical research. Some arguments against expecting evaluation studies to have predictive value include:

C1: Evaluation research is atheoretical. When an iron rod is heated to a certain temperature, it will expand. The amount of expansion can be measured quite precisely. If the experiment is repeated with the same iron rod under identical conditions, the results will be identical to within a small measuring error. Similar predictions based on controlled experiments cannot be made with respect to road safety evaluation studies. Road accidents are the result of an incompletely understood stochastic process and very few results can be ruled out on theoretical grounds.

C2: Evaluation studies are not done on a routine basis or according to a standardized sampling plan. The extent to which reported results are representative of the conditions characterizing future applications of a certain measure is often unknown.

C3: Evaluation studies that are published may be a biased sample of those that are made. In particular, studies showing no effect or an increase in the number of accidents may end up in file drawers more often than studies showing accident reductions.

C4: Most evaluation studies are non-experimental and rely on official accident data known to be incompletely reported. Unreliable data and flawed research designs make the results of such studies highly unreliable. The fact that the research designs employed in evaluation studies tend to change over time compounds this difficulty.

C5: Evidence from evaluation studies is often conflicting. It is not always possible to identify any subset of the evidence as more valid and reliable than other subsets or find good explanations of the conflicting results.

C6: Safety measures may undergo technical innovations or changes over time that invalidate the results of evaluation studies. Road user behavioural adaptation to safety measures may also change over time.

C7: The effectiveness of safety measures is likely to tend to be reduced in accordance with the law of diminishing marginal returns. In particular, highway design and traffic control measures are likely to be carried out at the worst blackspots before they are introduced at other locations. Results of evaluation studies referring to blackspots do not necessarily apply to other locations.

C8: The traffic system changes over time. The vehicle fleet, the population of drivers, the road network, etc are not the same today as twenty years ago. In the meantime, a broad range of safety measures has been introduced. It cannot be taken for granted that the effect of a certain safety measure is the same today as it was in the past.

Again, the validity of these arguments is an empirical question. The lists of arguments for and against attaching predictive value to the results of evaluation studies show that predictability is affected by a large number of factors. The lists of arguments are by no means exhaustive. The present study cannot test the validity of all the arguments, but will concentrate on just a few of them.

TESTING THE PREDICTIVE VALUE OF EVALUATION STUDIES: DATA AND RESEARCH APPROACH

Test cases

In order to test the predictive value of the results of evaluation studies, studies of two safety measures are used as cases. The two cases are road lighting

and traffic separation. A meta-analysis of evaluation studies concerning the effects of public road lighting on number of accidents has been reported previously (Elvik 1995a). The analysis presented here is based on that analysis. A list of the studies included in the meta-analysis is available upon request.

Traffic separation consists of physical measures intended to separate pedestrians and/or cyclists from motor vehicles. Three types of solutions have been included in this study: (1) sidewalks, separated from motor vehicles by means of kerbstone and intended for use by pedestrians and cyclists travelling in both directions, (2) cycle paths, intended for cyclists only, separated from motor vehicles by means of a dividing area (usually a 3 m wide grass covered area with V-profile) or kerbstone and separated from sidewalks by means of kerbstone or road markings, (3) pedestrian and cyclist tracks, separated from motor vehicles by means of a dividing area, supplemented with guardrails at locations where the dividing area is narrow.

A meta-analysis of evidence from 26 evaluation studies containing a total of 151 results concerning traffic separation was made (Elvik, 1995b). Reference to the studies included in this meta-analysis is available upon request. A detailed presentation of the results of this meta-analysis is beyond the scope of this paper.

Approach to testing predictive value

Figure 1 shows the approach adopted to test the predictive value of the evaluation studies concerning road lighting and traffic separation.

Studies were arranged in chronological order and the evidence of each study quantified in terms of the statistical weight contributed by it. For each result, the statistical weight is proportional to the inverse of the variance of that result. For example, a result from a before-and-after study based on 145 accidents before and 97 accidents after has a statistical weight of $1/(1/145+1/97)$. For further details, see Fleiss (1981) and Elvik (1995c). A moving sum of statistical weights was formed. The first stage of analysis was to partition the evidence from all evaluation studies into quintiles (shares of 20% each), based on statistical weights.

A weighted mean safety effect was estimated on the basis of the first 20% of the evidence from evaluation studies. This estimate was then treated as a prediction of the effect estimated on the basis of the next 20% of the evidence (stage 2 in Fig. 1). Predicted and actual values were compared. This process was repeated, using the first 40%, the first 60% and the first 80% of the evidence from evaluation studies to predict results estimated from, respectively, the third, fourth and fifth 20% of the evidence. At each stage the predicted values were compared to the actual values.

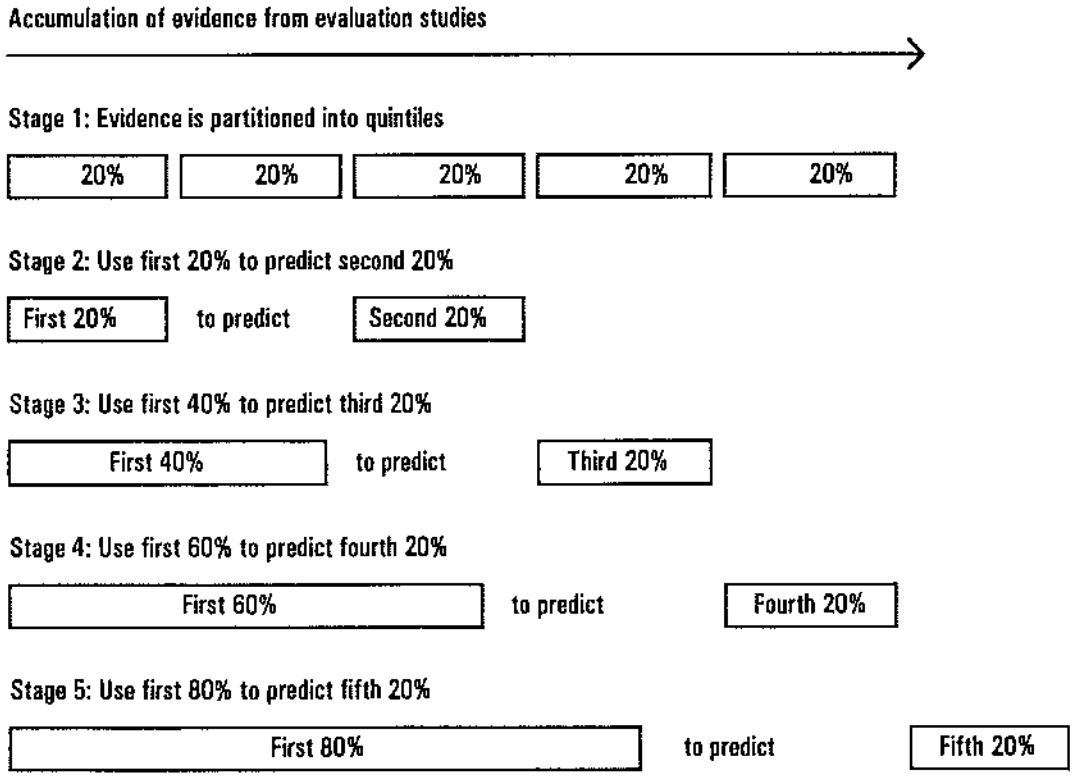


Fig. 1. Design of the test for the predictive performance of road safety evaluation studies.

Measures of predictive performance

The predictive performance of evaluation studies was measured in two ways. First, by comparing the 95% confidence interval of the predicted values (the weighted mean safety effects estimated on the basis of the first 20, 40, 60 or 80% of evidence) to the 95% confidence interval of the actual values (the weighted mean safety effects estimated on the basis of the second, third, fourth and fifth 20% of evidence). If the smaller confidence interval is entirely contained within the larger, the prediction is classified as a "perfect hit". If the confidence intervals overlap partly, the prediction is classified as "partly a hit, partly a miss". If the confidence intervals do not overlap at all, the prediction is classified as a "complete miss".

The second measure of predictive performance is the weighted mean percent prediction error. For each estimate of safety effect, the accuracy of prediction is described by means of the following measure:

$$\text{Accuracy of prediction} = \frac{\text{Actual safety effect}}{\text{Predicted safety effect}}$$

where both the actual and the predicted safety effect is measured in terms of the relative change in the number of accidents. A value of, for example, 0.85 indicates a 15% reduction in the number of accidents. The value of the accuracy measure is 1.0 if the prediction is perfectly correct, less than 1.0 if the actual safety effect is more beneficial (i.e. greater accident reduction or smaller increase) than the predicted effect and greater than 1.0 if the actual safety effect is less beneficial (i.e. greater increase or smaller reduction) than the predicted effect. A value of, for example, 1.05 indicates a prediction error of 5%. The prediction error estimated for each result was weighted by means of the statistical weight of each result and a weighted mean prediction error estimated by means of the following formula:

$$\text{Weighted error} = \exp\left[\sum_{i=1}^{n} \ln(\text{actual}_i/\text{predicted}_i)\cdot\text{weight}_i \Big/ \sum_{i=1}^{n} \text{weight}_i\right]$$

where actual_i denotes the actual safety effect of result i, predicted_i the predicted safety effect, weight_i the statistical weight of result i, ln the natural logarithm and exp the exponential function. This definition of the weighted mean prediction error is strictly analogous to the definition of the weighted mean safety effect according to the log odds method of meta-analysis (Fleiss, 1981).

THE PREDICTIVE VALUE OF EVALUATION STUDIES: ROAD LIGHTING*

A previous meta-analysis (Elvik 1995a) indicated that the results of studies that have evaluated the effects of public road lighting on the number of nighttime accidents are very robust with respect to study design, decade of publication and country where the study was made. The safety effects of road lighting appear to be very stable over time. The possible presence of publication bias was tested by means of the funnel graph method (Light and Pillemer 1984). This method relies on visual inspection of results plotted in a coordinate system. The abscissa shows the relative change in the number of accidents according to each result. The ordinate shows the statistical weight of each result. The idea is that if there is no publication bias, the scatter plot of results should resemble the form of a funnel turned upside down. The dispersion of the data points should narrow as sample size (statistical weight) increases, since large samples give more precise estimates of effects than small samples. If the tails of the scatter plot are symmetrical and the density of data points nearly the same in all areas of the diagram, this indicates that there is no publication bias. No evidence of publication bias was found for road lighting. The effects of public lighting were found to vary according to accident severity.

Table 1 presents the results of a test of the predictive performance of studies that have evaluated the safety effects of public lighting. Twelve tests are contained in the table. In terms of the classification explained above, three predictions were perfect hits, eight were partly hits, partly misses and one was a complete miss. The degree of overlap between the confidence intervals for the partly correct predictions was, in general, quite extensive.

The results of evaluation studies display a remarkable stability over time for fatal and injury accidents. For property-damage-only accidents there is evidence that the effects of road lighting have increased over time. The reasons for this are unknown. Improved quality of lighting is one possible reason. Also, the reporting of property-damage-only accidents is likely to be less complete and more unreliable than the reporting of fatal and injury accidents and may have declined over time.

The accuracy of predictions is shown in Table 2. The weighted mean prediction error is, in most cases, less than 10%, which must be regarded as quite small.

*A list of the studies referred to in the meta-analyses of road lighting and traffic separation used in this paper is available from the author upon request.

Table 1. Predicted effects of road lighting on the number of accidents in darkness by quintiles. Upper and lower 95% confidence limits

		Percent change in the number of accidents in darkness					
		Predicted values			Actual values		
Prediction based on	Prediction referring to	Lower 95%	Best estimate	Upper 95%	Lower 95%	Best estimate	Upper 95%
			All accidents				
First 20%	Second 20%	−24	−19	−13	−27	−22	−16
First 40%	Third 20%	−24	−20	−16	−28	−23	−18
First 60%	Fourth 20%	−24	−21	−18	−24	−19	−14
First 80%	Fifth 20%	−25	−23	−20	−37	−32	−27
			Fatal and injury accidents				
First 20%	Second 20%	−37	−30	−23	−35	−28	−20
First 40%	Third 20%	−34	−29	−24	−39	−32	−24
First 60%	Fourth 20%	−34	−30	−26	−31	−24	−15
First 80%	Fifth 20%	−32	−28	−25	−41	−34	−27
			Property-damage-only accidents				
First 20%	Second 20%	−16	−9	−1	−24	−18	−12
First 40%	Third 20%	−19	−14	−9	−26	−20	−12
First 60%	Fourth 20%	−20	−16	−12	−25	−19	−12
First 80%	Fifth 20%	−20	−17	−13	−34	−27	−18

Table 2. Mean percent prediction error for predicted effects of road lighting on the number of accidents in darkness

Accident severity	Prediction based on	Prediction referring to	Weighted mean percentage prediction error	Direction of prediction error
All accidents	First 20%	Second 20%	4.0%	Effect underpredicted
	First 40%	Third 20%	4.1%	Effect underpredicted
	First 60%	Fourth 20%	2.1%	Effect overpredicted
	First 80%	Fifth 20%	14.1%	Effect underpredicted
Fatal and injury	First 20%	Second 20%	3.8%	Effect overpredicted
	First 40%	Third 20%	6.6%	Effect underpredicted
	First 60%	Fourth 20%	9.4%	Effect overpredicted
	First 80%	Fifth 20%	7.3%	Effect underpredicted
Property-damage-only	First 20%	Second 20%	7.6%	Effect underpredicted
	First 40%	Third 20%	8.5%	Effect underpredicted
	First 60%	Fourth 20%	3.6%	Effect underpredicted
	First 80%	Fifth 20%	11.9%	Effect underpredicted

The effect of road lighting on property-damage-only accidents is consistently underpredicted (i.e. actual effects were greater than the predicted effects). For fatal and injury accidents, prediction errors are not consistently in one direction.

The mean percent prediction error does not decline as the amount of evidence the predictions are based on, increases. Predictions based on 80% of the evidence are not more accurate than predictions based on 20% of the evidence. At first glance, these results may appear somewhat counter intuitive. After all, the estimates of mean safety effect become more precise as the amount of evidence (the number of studies and their statistical weights) increases.

It is, however, a logical fallacy to think that predictions based on a highly precise estimate of an effect will necessarily be more precise than predictions based on a less precise estimate of an effect. This can perhaps be seen by comparing, Figures 2 and 3. Figure 2 shows the weighted mean safety effect of road lighting on all accidents. The thick line shows the best estimate of the weighted mean safety effect, and the dashed lines are the upper and lower 95% confidence limits of the weighted mean safety effect.

It is readily seen that the weighted mean safety effect stabilizes quite quickly and remains remarkably stable when the cumulative statistical weights have reached about 1500. The confidence interval narrows just as quickly and is very small beyond a cumulative statistical weight of about 2500.

Figure 3 shows each of the 142 estimates of the safety effects of road lighting, arranged in chronological order. The distance between data points along the abscissa shows the statistical weight of each result. The statistical weights are seen to vary substantially, from a minimum of less than 10 to a maximum of about 650 (for the data point located at about 3800 on the horizontal scale).

There is large variation in the individual results, ranging from an outlier showing more than a 4-fold

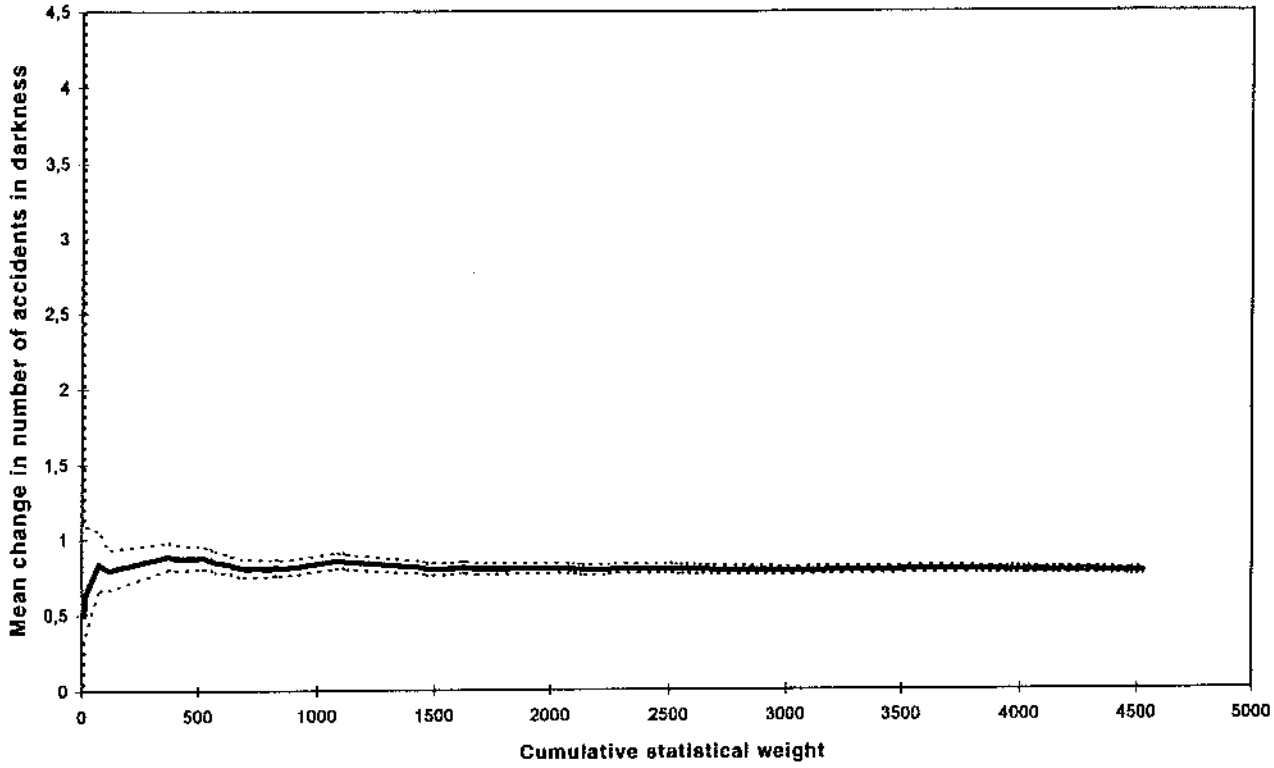


Fig. 2. Weighted mean safety effect of road lighting with upper and lower 95% confidence limits. Values < 1.0 = reduced number of accidents.

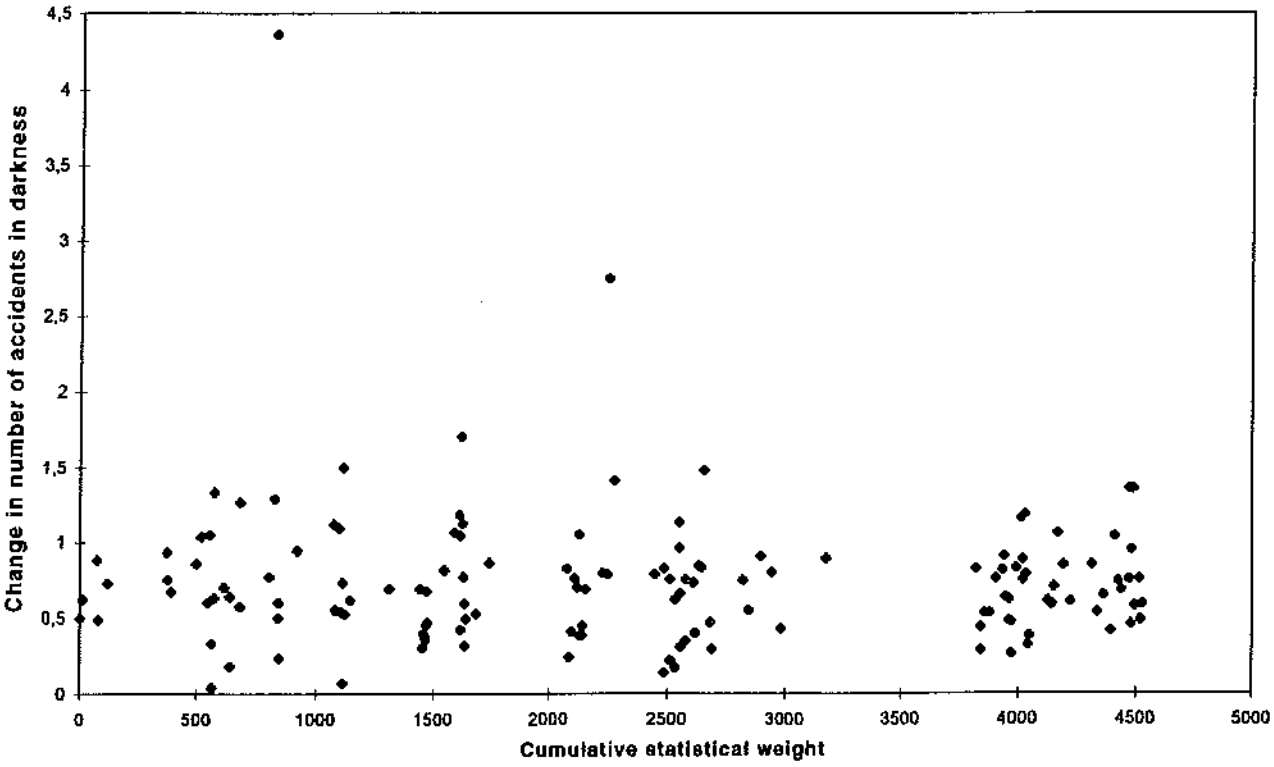


Fig. 3. 142 estimates of the safety effect of road lighting in chronological order. Values < 1.0 = reduced number of accidents.

increase in the number of accidents to a few data points indicating a reduction in the number of accidents of more than 90%. Most of this variation is attributable to random variation in the number of accidents (Elvik 1995a). The variation in individual results is quite stable over time. The mean prediction error depends on the amount of variation in the individual estimates of safety effects around the predicted mean safety effect, not on the precision of the estimate of the mean effect itself. There is no reason to believe that the variability of individual estimates has become smaller over time. Such a tendency would be expected only if each new estimate was based on a larger accident sample than the previous estimates. This does not appear to be the case for road lighting. Evidence of safety effect comes in small doses and the doses have not become larger over time.

THE PREDICTIVE VALUE OF EVALUATION STUDIES: TRAFFIC SEPARATION

Table 3 shows the performance of evaluation studies in predicting the safety effects of traffic separation, based on a meta-analysis reported elsewhere (Elvik, 1995b). A distinction has been made between accidents involving pedestrians or cyclists, which are generally taken as the target accidents for traffic separation, and accidents involving motor vehicles exclusively, which are generally presumed not to be affected by traffic separation.

The predictions for accidents involving pedestrians or cyclists are seen to be rather poor. There are two complete misses, one partly correct prediction and one perfect hit. For accidents involving motor vehicles exclusively, there are three partly correct predictions and one complete miss. Table 4 shows the mean percentage prediction errors.

For accidents involving pedestrians or cyclists, the first two predictions (based on the first 20% and first 40% of evidence) were erroneous by a margin of 20–30%. Subsequent predictions were more accurate, as the weight of evidence from studies showing no effect of traffic separation was incorporated into the basis for the predictions. There is a clear tendency towards more accurate predictions as the amount of evidence predictions were based on, increases. A similar tendency is found for predictions of effects on the number of accidents involving motor vehicles exclusively.

It would be interesting to know why the effect of traffic separation on accidents involving pedestrians and cyclists has diminished over time. There are at least two possible explanations for this tendency, one

Table 3. Predicted effects of traffic separation on the number of accidents by quintiles. Upper and lower 95% confidence limits

		Percent change in the number of injury accidents					
		Predicted values			Actual values		
Prediction based on	Prediction referring to	Lower 95%	Best estimate	Upper 95%	Lower 95%	Best estimate	Upper 95%
		Pedestrian and cycle accidents					
First 20%	Second 20%	−31	−25	−18	−15	−8	+1
First 40%	Third 20%	−22	−17	−11	+1	+10	+21
First 60%	Fourth 20%	−13	−8	−4	−11	−4	+5
First 80%	Fifth 20%	−10	−6	−2	−14	−5	+4
		Accidents involving motor vehicles exclusively					
First 20%	Second 20%	−7	+1	+9	−17	−10	−3
First 40%	Third 20%	−10	−5	+0	+3	+10	+18
First 60%	Fourth 20%	−4	+1	+5	−11	−4	+3
First 80%	Fifth 20%	−4	−1	+3	−14	−7	−0

Table 4. Mean percent prediction error for predicted effects of traffic separation on the number of accidents

Accident severity	Prediction based on	Prediction referring to	Weighted mean percentage prediction error	Direction of prediction error
Pedestrian and cycle accidents	First 20%	Second 20%	22.7%	Effect overpredicted
	First 40%	Third 20%	32.1%	Effect overpredicted
	First 60%	Fourth 20%	5.4%	Effect overpredicted
	First 80%	Fifth 20%	0.9%	Effect overpredicted
Motor vehicle accidents	First 20%	Second 20%	10.7%	Effect underpredicted
	First 40%	Third 20%	16.1%	Effect overpredicted
	First 60%	Fourth 20%	4.9%	Effect underpredicted
	First 80%	Fifth 20%	5.4%	Effect underpredicted

methodological, the other substantive. The methodological explanation suggests that the quality of evaluation studies has improved over time, and that studies employing a superior design tend to find smaller effects of the safety measure than inferior studies do. This has been referred to as the "Iron Law" of evaluation studies (Rossi and Freeman 1985). For example, recent before-and-after studies take account of the regression-to-the-mean effect and of secular accident trends, whereas older studies did not take account of these confounding factors. The substantive explanation suggests that the true safety effect of traffic separation has become smaller over time, due, for example, to increased driving speeds (among cyclists as well as motorists) or a declining use of the facilities caused, for example, by inadequate maintenance.

There is no direct way of testing these explanations; the required data are not available. An indirect test is attempted in Table 5, where the weighted mean effects on the number of accidents involving pedestrians and cyclists is shown by type of solution and study design. The idea is that, if the substantive explanation is true, one would find a declining safety effect over time, going from the first to fifth quintile, when controlling for type of solution and study design.

Table 5 clearly shows the difficulties of such a test. The various study designs have not been used regularly. None of them are represented in all five quintiles of the data set. Besides, this detailed partitioning of the data greatly reduces the sample size of each data point and thus enlarges the contribution of random variation to the results. Not all of the figures given in Table 5 are statistically significant, but confidence intervals have been left out in order not to overcrowd the table with numbers. A tendency towards smaller effects on safety going from the first to the fifth quintile is not found in Table 5. This does not support the substantive explanation, but the data are too limited to support a firm conclusion.

In general, it is not obvious that providing a more complete explanation of research findings would improve the ability to predict future effects. Effects of safety measures are known only from evaluation studies that are made. In that sense, it is only the results of future evaluation studies that can be predicted. If the findings of evaluation studies vary according to study design, then predicting future results would involve predicting the designs that will be adopted in future research. There is no obvious basis for making such a prediction.

Examples can be given both of very erroneous accident predictions based on models explaining more than 98% of the variation in the number of accidents (Partyka 1991) and of highly successful accident predictions based on models explaining no more than 10–20% of the variation in the number of accidents (Brüde and Larsson 1993). It would therefore seem to be a logical fallacy to presume that an improved understanding of why certain events occurred in the past entails an improved ability to predict future events.

DISCUSSION

Studies evaluating the effects of traffic safety measures are made mainly for use in planning future use of the measures, and not just for historical record. But it is not always correct to assume that the future will be like the past, and that the results of evaluation studies will correctly predict the future effects of the measures that have been evaluated. Arguments can be given both for and against attributing a general predictive value to the results of evaluation studies.

Table 5. Predicted effects of traffic separation on the number of accidents by quintiles, type of separation and study design

		Quintiles of research evidence				
Type of solution	Study design	First	Second	Third	Fourth	Fifth
Sidewalk	Case–control, stratified by confounders			+54	−8	+25
	Case–control, matched groups	−31				
	Simple before-and-after					−32
Cycle path	Case–control, stratified by confounders	−12	−4	−4		
	Before-and-after with general comparison group		+32	+26	−1	−3
	Before-and-after with matched comparison group			+62		
Pedestrian and cycle track	Case–control, stratified by confounders			+15	−14	−23
	Simple before-and-after		−85	−23		−72
	Before-and-after with general comparison group	−44	−38	−16	−84	
	Before-and-after, taking account of regression-to-the-mean			−0	−17	

The predictive value of evaluation studies is essentially an empirical question which cannot be settled by means of theoretical arguments.

Testing the predictive performance of evaluation studies rigorously is very complex. The efforts reported in this paper are just a first approach to the problem. A more rigorous test would take better account of potential confounding factors than the tests reported in this paper. If, for example, it was known that new road lighting was of a better technical quality than older installations, and if the relationship between lighting quality and safety effect was known, a prediction could take account of this knowledge. A prediction taking account of these facts would perhaps not be a simple extrapolation of evidence from past studies, but might predict a different effect from that obtained in earlier evaluations. Similar comments apply to the effects of study design on the results of evaluation studies.

The knowledge required to make these more sophisticated predictions and more rigorous tests of them is, however, not presently available, nor likely to become available in the near future. The important role of randomness in accident counts must not be forgotten. There is no way of eliminating, much less predicting, random variation. The contribution of random variation can be reduced by relying on larger accident samples. Again, however, there is no way of predicting the sample sizes of future evaluation studies.

The case studies presented in this paper show that predictions can be very inaccurate when a trend is present in the data set. In these cases, predictions based on the most recent studies may be more accurate than predictions relying on all available evidence from past evaluation studies. Increasing the amount of evidence that predictions are based on, does not necessarily improve their accuracy.

CONCLUSIONS

The main results of the research reported in this paper can be summarized as follows:

1. Studies evaluating the effects of traffic safety measures are almost always intended for use in predicting future effects of the measures. Despite this, no study has been performed to test the predictive performance of evaluation studies.
2. A number of arguments can be given both for and against assigning a general predictive value to evaluation studies. These arguments merely show that predictability depends on many factors. It is not possible to assess the importance of the various arguments on theoretical grounds.
3. Meta-analyses of evaluation studies offer the possibility for testing the predictive performance of such studies. In this paper, simple tests were made for two safety measures for which meta-analyses of evaluation studies have been performed: road lighting and traffic separation. The tests involved partitioning the evidence from evaluation studies into quintiles and using the first 20%, the first 40%, the first 60% and the first 80% of evidence to predict, respectively, the second, third, fourth and fifth 20%.
4. The accuracy of predictions was found to depend mainly on whether the safety effect is stable over time or exhibits a trend. When a trend is found, predictions based on evidence from studies made before the trend became apparent, can be very misleading. Examples of this were found both for road lighting and traffic separation. When the effect is stable, predictions were reasonably accurate. Accuracy was not found to improve when the amount of evidence serving as the basis for predictions increased.
5. Trying to explain variation in the results of past evaluation studies will not necessarily improve the ability to predict the results of future studies. In general, the data required for explaining variation in the results of past research are not available.

REFERENCES

Brüde, U.; Larsson, J. Models for predicting accidents at junctions where pedestrians and cyclists are involved. How well do they fit? Accid. Anal. Prev. 25: 499–509; 1993.

Elvik, R. A meta-analysis of evaluations of public lighting as an accident countermeasure. Transportation Research Record, 1485, 112–124; 1995a.

Elvik, R. Virkninger av fysiske trafikksikkerhetstiltak for fotgjengere og syklister. Arbeidsdokument TST/0669/95. Oslo: Transportøkonomisk institutt; 1995b (In Norwegian).

Elvik, R. The safety value of guardrails and crash cushions: a meta-aualysis of evidence from evaluation studies. Accid. Anal. Prev. 27: 523–549; 1995C.

Fleiss, J. L. Statistical methods for rates and proportions, 2nd edition. New York: Wiley; 1981.

Light, R. J.; Pillemer, D. B. Summing up. The science of reviewing research. Cambridge, MA: Harvard University Press; 1984.

Partyka, S. C. Simple models of fatality trends revisited seven years later. Accid. Anal. Prev. 23: 423–430; 1991.

Rossi, P. H.; Freeman, H. E. Evaluation. A systematic approach, 3rd edition. Beverly Hills, CA: Sage; 1985.

Paper 4

Pergamon

Accid. Anal. and Prev., Vol. 28, No. 6, pp. 685–694, 1996

0001-4575/96 $15.00 + 0.00

PII: S0001-4575(96)00041-3

A META-ANALYSIS OF STUDIES CONCERNING THE SAFETY EFFECTS OF DAYTIME RUNNING LIGHTS ON CARS

RUNE ELVIK

Institute of Transport Economics, PO Box 6110 Etterstad, N-0602 Oslo, Norway

(Received 25 January 1996)

Abstract—A meta-analysis of 17 studies that have evaluated the effects on traffic safety of using daytime running lights (DRL) on cars is presented. A distinction is made between studies that have evaluated the effects of DRL on the accident rates of each car using it and studies that have evaluated changes in the total number of accidents in a country following the introduction of mandatory use of DRL. Three different definitions of the measure of safety effects are compared and their validity discussed. It is concluded that the nse of DRL on cars reduces the number of multi-party daytime accidents by about 10–15% for cars using DRL. The estimated effects on the total number of accidents of introducing DRL laws are somewhat smaller, 3–12% reduction in multi-party daytime accidents, and are likely to contain uncontrolled confounding effects. There is no evidence to indicate that DRL affects types of accident other than multi-party daytime accidents. Copyright © 1996 Elsevier Science Ltd

Keywords—Daytime running lights, Evaluation studies, Meta-analysis, Safety effects

INTRODUCTION

The mandatory use of daytime running lights (DRL) as a road safety measure has become more widespread in recent years. Countries requiring cars to turn on their headlights at all times now include Canada (for cars from model year 1990), Denmark, Finland, Hungary, Norway and Sweden. Most of the studies that have evaluated the safety effects of daytime running lights conclude that it is effective in reducing the number of daytime accidents involving more than one party (multi-party daytime accidents). But the estimates of safety effects vary and critics have pointed out flaws in many evaluation studies (Elvik 1993). In a recent paper, Theeuwes and Riemersma (1995) argue that the odds ratio method used in evaluations of the safety effects of DRL laws in Finland (Andersson et al. 1976), Sweden (Andersson and Nilsson 1981), Norway (Vaaje 1986; Elvik 1993), Canada (Arora et al. 1994) and Hungary (Hollo 1995) is highly unreliable. The odds ratio method makes the estimate of the effect of DRL very sensitive to changes in the number of accidents that are assumed not to be affected by the use of DRL (single vehicle daytime accidents and all nighttime accidents). Reanalyzing data from the evaluation of the DRL law in Sweden (Andersson and Nilsson 1981), Theeuwes and Riemersma find that the effect attributed to DRL was largely due to an unexplained increase in the number of single vehicle daytime accidents in the first year after the DRL law. According to their analysis, the DRL law in Sweden did not reduce multi-party daytime accidents as a proportion of all accidents, as one would expect if DRL affected just this type of accident. Hauer has questioned the validity of the assumption made in the odds ratio method that single vehicle daytime accidents aud nighttime accidents are not affected by DRL (Hauer 1995).

The purpose of this paper is to try to sort out some of the issues raised in debates concerning DRL evaluations. The following main problems are discussed:

1. How do the different ways of defining the variable intended to measure the safety effect of DRL (the dependeut variable) affect estimates of that effect?
2. What are the best current estimates of the safety effects of DRL according to different definitions of the dependent variable?
3. Are the results of different evaluation studies consistent or are there large, unaccounted for, variations in the results of these studies?

In order to shed light on these questions, a meta-analysis of evidence from 17 studies that have evaluated the safety effects of DRL has been performed.

Before presenting the results of the analysis, some of its elements will be explained.

META-ANALYSIS OF EVALUATION STUDIES

Retrieval of studies

The studies included in the meta-analysis were retrieved by means of literature surveys that were part of previous evaluation studies (Elvik 1993) and by scanning recent reports (Arora et al. 1994; Hansen 1993,1995; Hollo 1995). A total of 17 studies were included. The studies that were included are listed in Table 1. The studies contain a total of 60 estimates of the effects of DRL on accident occurrence. Studies not reporting the number of accidents on which their results were based could not be included in the meta-analysis. The results of studies not included in the meta-analysis are listed in Table 2. The studies of Allen and Clark (1964) and Attwood (1981) contain both results that were included and results that could not be included in the meta-analysis and are therefore listed in both Tables 1 and 2.

Characteristics of evaluation studies

Each evaluation study included in the meta-analysis was categorized with respect to study design and the level of safety effects studied. A distinction was made between three types of study design: (1) experimental studies, in which cars are randomly assigned to either a DRL-condition or a no DRL-condition, (2) before-and-after studies with a comparison group, in which accident records for cars which had DRL installed are compared to those of cars where DRL was not installed, (3) simple before-and-after studies, in which accident records before and after cars had DRL installed, or before and after a DRL law was passed, are compared.

A distinction was made between two levels of safety effects: (1) the effects of DRL for each car using it. These effects are referred to as the intrinsic effects of DRL. (2) The effects on the total number of accidents in a country having a DRL law, or a campaign designed to promote the use of DRL. These effects are referred to as the aggregate effects of DRL. Seven studies with a total of 23 results refer to the intrinsic effects of DRL, 10 studies with a total of 37 results refer to the aggregate effects of DRL.

Accident typology

Figure 1 shows the classification of accidents used in studies of the effects of DRL. The four basic categories are: (1) single vehicle daytime accidents (SD), (2) multi-party daytime accidents (MD), (3) single vehicle nighttime accidents (SN) and (4) multi-party nighttime accidents (MN). Accidents in twilight have been omitted from most studies. Twilight accidents were classified as nighttime accidents in the

Table 1. List of studies included in meta-analysis

Authors and year	Country	Study design	Level of effects
Allen and Clark 1964	United States	Simple before-after	Aggregate
Cantilli 1965	United States	Experiment	Individual
Cantilli 1970	United States	Experiment	Individual
Andersson et al. 1976	Finland	Simple before-after	Aggregate
Andersson and Nilsson 1981	Sweden	Simple before-after	Aggregate
Attwood 1981	Canada	Experiment	Individual
Stein 1985	United States	Experiment	Individual
Vaaje 1986	Norway	Simple before-after	Aggregate
Sparks et al. 1989	Canada	Simple before-after	Individual
Hocherman and Hakkert 1991	Israel	Simple before-after	Aggregate
Elvik 1993	Norway	Simple before-after	Aggregate
Hansen 1993	Denmark	Simple before-after	Aggregate
Kuratorium für Verkehrssicherheit 1993	Austria	Before-after with comparison	Individual
		Simple before-after	Individual
Sparks et al. 1993	Canada	Before-after with comparison	Individual
Arora et al. 1994	Canada	Simple before-after	Aggregate
Hansen 1995	Denmark	Simple before-after	Aggregate
Hollo 1995	Hungary	Simple before-after	Aggregate

Table 2. Results of studies not included in meta-analysis

Authors and year	Country	Context of study	DRL effect	Type of accident
Allen and Clark 1964	United States	Greyhound Bus company	−11%	Daytime accidents
Allen 1965	United States	Questionnaire to companies	−39%	Not stated
Allen 1979	United States	Checker Cab company	−7%	Not stated
Attwood 1981	United States	ATT Long Lines Division	−32%	Not stated

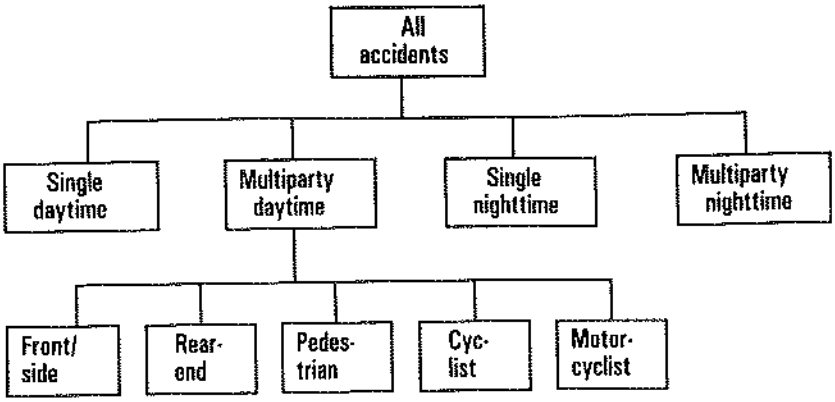


Fig. 1. Classification of accidents in studies of the effects of DRL on accident occurrence.

studies of Andersson et al. (1976) and Andersson and Nilsson (1981).

Some studies have analysed the effects of DRL on specific types of multi-party daytime accidents. The types that have been studied most frequently are listed in Fig. 1. Front or side impacts and rearend collisions generally involve motor vehicles exclusively. The other categories are collisions between cars and pedestrians, cars and cyclists and cars and motorcyclists. Cars include passenger cars, vans, trucks and buses. The DRL laws or campaigns included in this paper mostly refer to cars, but have sometimes required all motor vehicles to use DRL.

Measures of safety effect

Three different definitions of the variable intended to measure the safety effects of DRL are compared. The three definitions are referred to as (1) effects on accident rate ('accident rate' for short), (2) effects on the proportion of daytime multi-party accidents ('simple odds' for short) and (3) the odds ratio measure of effects ('odds ratio' for short). All three measures of effect refer to the effects of DRL on multi-party daytime accidents. Applying the notation introduced above and subscripting 'before or without DRL' with b and 'after or with DRL' with a, the three measures of effect are defined as follows:

Effect on accident rate
$$=(MD_a/KMT_a)/(MD_b/KMT_b)$$

Effect on simple odds
$$=[MD_a/(MN_a+SD_a+SN_a)]/[MD_b/(MN_b+SD_b+SN_b)]$$

Effect on odds ratio
$$=[(MD_a/SD_a)/(MN_a/SN_a)]/[MD_b/SD_b)/(MN_b/SN_b)]$$

where KMT denotes vehicle kilometres of travel. All three measures of effect take on values less than 1.0 if DRL reduces the number of multi-party daytime accidents.

The accident rate for multi-party daytime accidents (accidents per million vehicle kilometres of travel) is the most common measure of effect in fleet studies of the effects of DRL. The odds ratio measure of effect has been the most common measure of effect in studies that have evaluated the effects of DRL laws. The simple odds measure has been proposed by Theeuwes and Riemersma (1995). According to the simple odds measure of effect, changes in the number of multi-party daytime accidents associated with the use of DRL are compared to changes in the number of all other types of accident combined.

The studies evaluating the effects of DRL laws in Sweden (Andersson and Nilsson 1981) and Norway (Vaaje 1986; Elvik 1993) did not report data on vehicle km of travel needed to estimate the accident rate. Data on vehicle kilometres of travel for these two countries were obtained from other sources (OECD 1994; Rideng 1995).

Statistical weighting of results

Each result is assigned a statistical weight that is proportional to the inverse of the variance of that result (Fleiss 1981). The variance of each result is determined by the number of accidents it is based on. The statistical weight of each result was estimated for the different measures of effect according to the following definitions:

Weight of accident rate
$$=1/(1/MD_a+1/MD_b)$$

Weight of simple odds
$$=1/[1/MD_a+1/MD_b+1/(MN_a+SD_a+SN_a)+1/(MN_b+SD_b+SN_b)]$$

Weight of odds ratio
$$=1/(1/MD_a+1/SD_a+1/MN_a+1/SN_a+1/MD_b+1/SD_b+1/MN_b+1/SN_b)$$

A weighted mean safety effect was estimated by means of the logodds method. The weighted mean safety effect was defined as follows:

Weighted mean safety effect =

$$\exp[\{\Sigma \ln(\theta_j) \cdot W(\theta_j)\}/\Sigma W(\theta_j)]$$

where exp denotes the exponential function, θ_j each estimate of effect and $W(\theta_j)$ the statistical weight of each estimate of effect. The statistical significance of the weighted mean safety effect was assessed by estimating a 95% confidence interval. A more detailed description of the logodds method of meta-analysis can be found in Fleiss (1981).

Testing for publication bias

The term, publication bias, denotes a tendency not to publish the results of evaluation studies that are, for some reason, believed not be useful. It may be the case, for example, that studies of DRL showing an increase in the number of accidents or no statistically significant change, are less likely to be published than those showing accident reductions. To test for publication bias, the funnel graph method of Light and Pillemer(1984) was used. This method is not a formal test in a strict sense, but does give some clues as to the possible presence of publication bias. It relies on visual inspection of a diagram in which results are plotted against sample size (statistical weight).

Figure 2 shows a funnel graph diagram for results showing the intrinsic effects of DRL on the accident rates of cars using it. Values to the left of 1.0 on the abscissa show reductions in accident rate, values to the right of 1.0 show increases in accident rate. Statistical weight is used as an indicator of sample size.

Each data point in Fig. 2 is the resnlt of an evaluation study. The basic assumption of the funnel graph method, is that if the dispersion of results mainly reflects random variation around a certain mean value, and if there is no publication bias, the dispersion of the data points should resemble the shape of a funnel turned upside down. If, on the other hand, one of the tails of the funnel is missing, this indicates publication bias.

The results in Fig. 2 are predominantly based on small accident samples. One result, with a statistical weight of almost 3,000, alone represents more than half of the sum of statistical weights for all data points in Fig. 2. There is a concentration of data points around the value of 0.8 on the abscissa, indicating a 20% reduction in accident rate. However, several data points are found both to the left and to the right of this value. It is concluded that Fig. 2 gives no clear indication of any publication bias.

It was not possible to test for publication bias in results referring to the aggregate effects of DRL. These results are based on highly varying changes in the use of DRL, and the aggregate effects of DRL can be expected to vary accordingly. A funnel graph diagram for results referring to aggregate effects would therefore violate the assumption made in the funnel graph technique that the dispersion of results should mainly reflect random variation around the weighted mean effect.

Relating aggregate effects to intrinsic effects

When the use of DRL is made mandatory, the aggregate safety effects will depend on: (1) the intrinsic effects of DRL, (2) the initial rate of DRL use, (3) the size of the increase in DRL use, and (4) the accident involvement rates of those who start using DRL compared to those who continue to drive without DRL. The relationship between the intrinsic and aggregate effects of DRL is complex.

Koornstra (1993,1995) has derived two mathematical functions to relate the aggregate effects of DRL to the intrinsic effects. According to these functions, an implicit intrinsic effect of DRL is estimated as a function of the aggregate effect and the percentage of cars using DRL before and after a law was introduced or a campaign conducted. One function applies to accidents involving cars and road users not using DRL (pedestrians and cyclists), the other applies to accidents involving only cars or other potential users of DRL (motorcycles). The functions are as follows:

Implicit intrinsic effect for pedestrians and cyclists

$$= E/[L_a - L_b \cdot (1-E)]$$

Implicit intrinsic effect for motor vehicles

$$= E/[(2L_a - L_a^2) - (2L_b - L_b^2) \cdot (1-E)]$$

where E denotes the aggregate effect of DRL, L the proportion using DRL, a the after period and b the before period. According to the functions, the implicit intrinsic effect of DRL will depend on (1) the initial usage rate for DRL, (2) the size of the increase in DRL use when a law is introduced and (3) the estimated aggregate effect of DRL. The validity of the estimated intrinsic effects of DRL according to these functions will therefore depend on the validity of the estimated aggregate effects of DRL.

RESULTS

Intrinsic effects of DRL

Table 3 shows the weighted mean percent change in the number of multi-party daytime accidents associated with the use of DRL by definition of the measure of effect and study design.

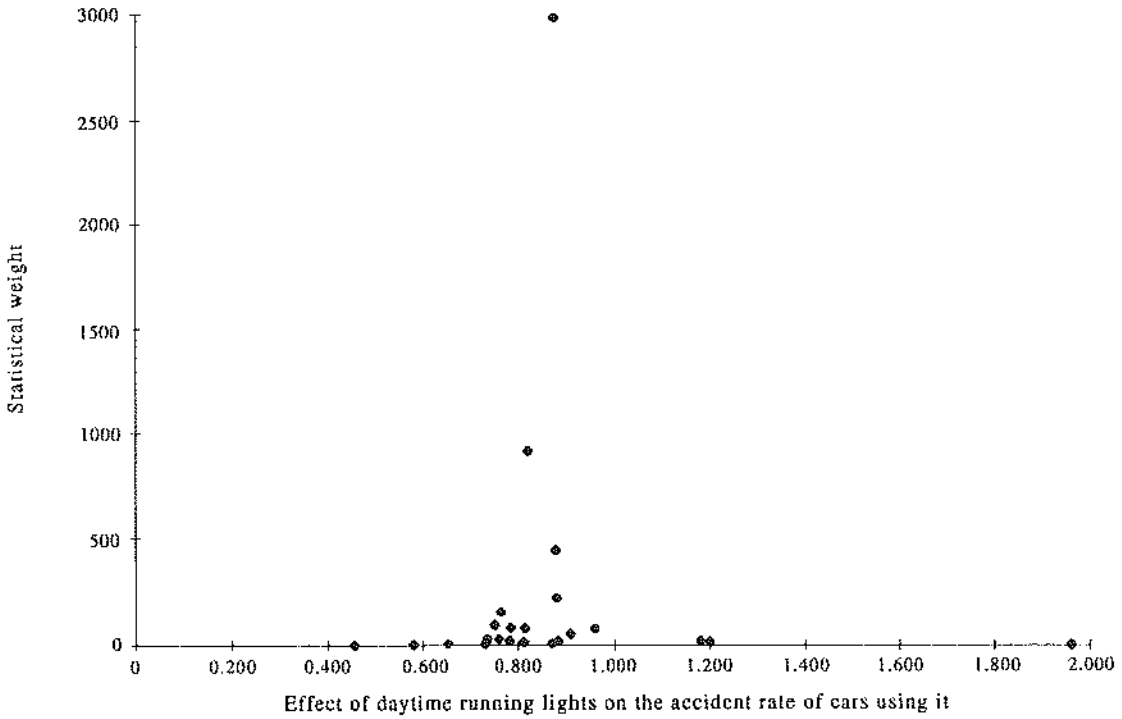


Fig. 2. Funnel graph diagram for effects of DRL on the individual accident rates of cars.

The results are very consistent across study designs. It has been suggested that simple before-and-after studies of vehicle fleets are suspect, because of the possible presence in these studies of uncontrolled regressiou-to-the-mean effects (Elvik 1993). If such effects were present, one would expect the estimated effect of DRL to be greater in simple before-and-after studies, than in experiments with random assignment (which guarantees against selecting just cars with bad accident records for DRL installation). This is not the case.

Moreover, the effects of DRL are highly consistent for the different definitions of measure of effect. The effects of DRL are slightly smaller accordiug to the simple odds measure of effect than according to the other two measures of effect, but the differences are not statistically significant and all estimates point in the same direction. Based on Table 3, the best estimate of the intrinsic effect of DRL on multi-party daytime accidents is a reduction of about 10–15%.

Table 4 shows the intrinsic effect of DRL with respect to various types of accident. In Table 4, the results of studies using different study designs have been merged. This was regarded as appropriate in view of the rather small differences in results between study designs according to Table 3.

According to Table 4, DRL is effective in reducing all types of multi party daytime accident. There is no consistent pattern in the variation of the effects of DRL between accident types. A quite large reduction in the number of pedestrian accidents is found for the accident rate measure of effect. This finding

Table 3. Intrinsic effects of daytime running lights by study design and definition of measure of effect

	Percent change in the number of multiparty daytime accidents								
	Effect on accident rate			Effect on simple odds			Effect on odds ratio		
Study design	Lower 95%	Best estimate	Upper 95%	Lower 95%	Best estimate	Upper 95%	Lower 95%	Best estimate	Upper 95%
Experiments with random assignment	−32	−15	+6	−39	−13	+24			
Before-and-after with comparison group	−28	−18	−5	−31	−6	+29			
Simple before-and-after design	−16	−14	−12	−15	−11	−8	−22	−14	−5

Table 4. Intrinsic effects of daytime running lights by type of accident and definition of measure of effect

Type of accident	Percent change in the number of multiparty daytime accidents								
	Effect on accident rate			Effect on simple odds			Effect on odds ratio		
	Lower 95%	Best estimate	Upper 95%	Lower 95%	Best estimate	Upper 95%	Lower 95%	Best estimate	Upper 95%
Front or side impacts	−16	−13	−10	−16	−11	−6	−25	−15	−3
Rearend collisions	−21	−16	−11	−19	−13	−6	−25	−12	+3
Pedestrian accidents	−38	−25	−8	−46	−15	+35	−87	+130	+4064
Accident type not specified	−20	−14	−7	−18	−3	+15	−66	−29	+49
Mean effect for all types of accident	−16	−14	−12	−15	−11	−8	−22	−14	−5

is reversed for the odds ratio measure of effect. However, the estimate based on the odds ratio measure of effect is far from statistical significance at the 5% level. The effects of DRL with respect to front or side impacts and rearend collisions are very consistent across the different measures of effect. In general, the findings are quite robust with respect to the definition of the measure of effect.

Aggregate effects of DRL

Table 5 shows the aggregate effects of DRL laws or campaigns to increase the use of DRL on various types of multi-party daytime accidents according to the three different measures of effect. All the estimates reported in Table 5 are based on simple before-and-after studies.

The weighted mean effect of DRL laws or campaigns on all multi-party daytime accidents is a 3–12% reduction, depending on the measure of safety effect. The accident reduction is statistically significant at the 5% level for all measures of safety effect. The estimated effect of DRL is, however, smaller for the simple odds number measure of effect than for the other two measures of effect.

In general, the aggregate effects of DRL are consistent with the intrinsic effects, at least as far the direction of the effect is concerned. The aggregate effect of DRL for rearend collisions is, however, inconsistent with the intrinsic effect. The intrinsic effect according to Table 4 is a 12–16% accident reduction, depending on the measure of effect used. The aggregate effect according to Table 5 is an increase of 3–20% in the number of rearend collisions, depending on the measure of effect used. It is only for the simple odds measure of effect that the increase found in Table 5 is statistically significant at the 5% level. The reasons for this inconsistency between intrinsic and aggregate effects are unknown. A possible explanation is related to changes in how drivers react to cars with the rear lights on as a consequence of an increase in DRL use. When only a few cars use DRL, a driver may take a lit rear light to mean that the car is braking. When more cars drive with lit rear lights, this reaction is less natural. Brake lights may become masked by rear lights, making it more difficult to detect when a car is braking.

Dose–response relationship for effects of DRL laws

The introduction of DRL laws has led to increased rates of DRL use in the countries where the laws have been introduced. At any given initial rate of DRL use, a large increase in DRL use is expected to have a greater impact on the number of accidents than a small increase in DRL use. The results presented in Table 6 test if such a dose–response relationship is found in studies that have evaluated the aggregate effects of DRL laws.

The accident rate measure of effect does not

Table 5. Aggregate effects of daytime running lights by type of accident and definition of measure of effect

Type of accident	Percent change in the number of multiparty daytime accidents								
	Effect on accident rate			Effect on simple odds			Effect on odds ratio		
	Lower 95%	Best estimate	Upper 95%	Lower 95%	Best estimate	Upper 95%	Lower 95%	Best estimate	Upper 95%
Front or side impacts	−14	−13	−12	−6	−5	−3	−15	−12	−8
Rearend collisions	−0	+3	+6	+16	+20	+24	−3	+4	+11
Pedestrian accidents	−22	−20	−18	−15	−13	−10	−14	−10	−5
Cyclist accidents	−9	−6	−4	−5	−2	+2	−25	−19	−12
Motorcycle accideuts	−23	−20	−18	−5	−0	+5	−10	+4	+20
Accident type not specified	−9	−7	−4	−9	−5	−1	−18	−11	−3
Mean effect for all types of accident	−13	−12	−11	−4	−3	−2	−11	−9	−7

Table 6. Aggregate effects of daytime running lights by percentage increase in use of daytime running lights and definition of measure of effect

		Percent change in the number of multiparty daytime accidents								
		Effect on accident rate			Effect on simple odds			Effect on odds ratio		
Use of DRL in before-period	Use of DRL in after-period	Lower 95%	Best estimate	Upper 95%	Lower 95%	Best estimate	Upper 95%	Lower 95%	Best estimate	Upper 95%
About 30%	About 60%	−16	−14	−12	−6	−4	−2	−12	−8	−3
About 40%	About 90%	−17	−11	−6	−2	+6	+14	−18	−5	+9
About 30%	About 90%	−10	−9	−8	−3	−1	+1	−15	−12	−9
About 50%	About 80%	−20	−16	−13	−13	−9	−4	−21	−12	−3
About 50%	About 95%	−20	−18	−16	−7	−5	−2	−12	−8	−3

show any clear dose–response pattern. Increasing DRL use from about 30 to 60% is found to reduce accident rates more than increasing DRL use from about 30 to 90%. The effect on accident rates of increasing DRL use from about 50 to 80% is almost the same as the effect of increasing it from 50 to 95%. No clear dose–response pattern is evident for the simple odds measure of effect, either. As far as the odds ratio measure of effect is concerned, the effect of increasing DRL use from 30 to 90% is nominally greater than the effect of increasing DRL use from 30 to 60%, but the difference is not statistically significant. On the other hand, the effect of increasing DRL use from 50 to 95% is smaller than the effect of increasing DRL use from 50 to 80%.

The absence of a clear dose–response pattern in the studies that have evaluated the effects of DRL laws suggests that the effects attributed to DRL in these studies are confounded with the effects of other variables. Examples of such variables include general trends in accident rates and measures that affect, selectively, the types of accident that serve as comparison accidents (presumably not affected by DRL) in evaluations of DRL laws. This does not necessarily mean that the DRL laws were not effective in reducing the number of accidents, but that the numerical estimates of those effects are highly uncertain. Moreover, the quality of data on the level of DRL use is quite poor in some studies (Elvik 1993; Hansen 1993, 1995).

Relationship between latitude and effects of DRL

Koornstra (1993, 1995) argues that there is a relationship between the geographical latitude of a country and the effects of DRL in that country. The further away from the equator one moves, the longer become the periods of dusk and dawn, and 'low sun' (the sun located just above the horizon), when DRL has the largest effect on vehicle conspicuity. On this basis Koornstra proposes that DRL has a greater effect on accidents the further away from the equator one moves. In order to test Koornstra's hypothesis, an implicit intrinsic effect of DRL was estimated for each country included in this study, applying the formulas Koornstra has derived for relating the aggregate effects of DRL to the intrinsic effects. The estimate of the DRL effect for each country is a weighted mean effect based on all evaluation studies reported in that country. The following mean latitudes of the countries included were used: Israel: 33; United States: 39; Hungary: 46; Austria: 47; Canada: 48; Denmark: 55; Sweden: 58; Norway: 62 and Finland: 63. The results of the test are given in Fig. 3.

The accident rate measure of effect does indicate that the effects of DRL vary according to latitude. The estimated, implicit intrinsic effect is a 9% reduction of the accident rate for multi-party daytime accidents in Israel and a 60% reduction of the accident rate for multi-party daytime accidents in Finland. Although the relationship is a bit noisy, its direction is clear. For the simple odds measure of effect, on the other hand, no clear relationship is found between latitude and DRL effect. The odds ratio measure of effect gives only a weak hint of a relationship between latitude and DRL effect. However, the direction of the association between latitude and DRL effect is the same for the odds ratio measure of effect as for the accident rate measure of effect, although the slope is much smaller for the former measure of effect than for the latter. On balance, the data presented in Fig. 3 indicate that there is a relationship between latitude and DRL effect in the direction predicted by Koornstra, but the relationship does not seem to fit very well the specific mathematical function Koornstra (1993, 1995) has proposed to describe it (the function is shown by the solid line in Fig. 3).

DISCUSSION

Does the use of DRL reduce the number of accidents during daytime in which more than one party is involved? On the basis of the evidence examined in this paper, an affirmative answer can be given to this question. The intrinsic effects of DRL

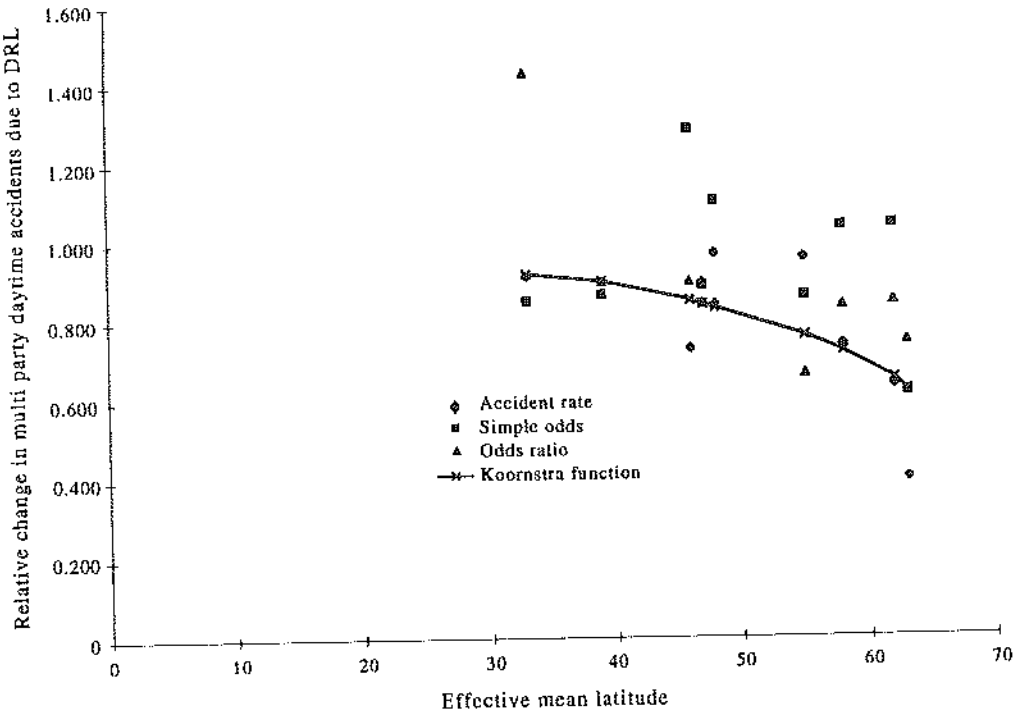


Fig. 3. Relationship between effective mean latitude and effects of DRL on multi-party daytime accidents assuming 100% use of DRL.

are quite consistent with respect to both study design and the definition of the variable intended to measure the effect. The intrinsic effects of DRL that have been found in vehicle fleet studies have been reproduced in studies of the effects of DRL laws. The intrinsic effects of DRL are very robust with respect both to study design and the definition of measure of effect. It is does not seem appropriate to dismiss the evidence of these studies as merely reflecting the effects of poor research design or a dubious definition of the variable intended to measure the effect of DRL.

The evidence from studies that have evaluated the aggregate effects of DRL laws or campaigns designed to increase the use of DRL is less convincing. In the first place, the weighted mean best estimate of the effect of DRL varies substantially according to the definition of measure of effect, from 3% for the simple odds number measure of effect to 12% for the accident rate measure of effect. These differences are statistically significant and could be of considerable practical importance in countries recording some 10–20,000 injury accidents per year (as the case is for Denmark, Finland, Norway and Sweden).

In the second place, the evidence concerning the aggregate effects of DRL comes from simple before-and-after studies only. These studies lack the experimental control of confounding factors provided by some of the vehicle fleet studies. The estimates of DRL effects based on before-and-after studies of DRL laws made, for example, in the Scandinavian countries are most likely to contain effects of uncontrolled confounding factors. The fact that no clear dose–response relationship was found between the size of the increase in DRL use and the size of the effects of DRL on accidents lends further support to this point of view.

Theeuwes and Riemersma (1995) have shown how sensitive the odds ratio measure of effect is to changes in the number of accidents that are supposed to be unaffected by the use of DRL, for example, single vehicle daytime accidents. One way of avoiding the problem of basing the estimate of the effect of DRL on the frequency of various types of accident that are presumed to be unaffected by DRL is to use the accident rate for multi-party daytime accidents only as the measure of effect. This solution, however, generates its own problems. Hansen (1993,1995), in his analyses of the effects of the Danish DRL law, uses accident statistics going back 10 years to show that there is a long term trend of decline in the multi-party daytime accident rate. This trend was evident long before the use of DRL was made mandatory in Denmark. Similar long term trends in accident rate have been found in both Norway and Sweden. In simple before-and-after studies, with just one before-period and just one after-period, no account is taken

of such a long term trend. It is likely that such studies overstate the effects of DRL on multi-party daytime accident rate.

This paper has compared different estimates of the effects of DRL on multi-party daytime accidents only. The point raised by Hauer (1995), that DRL may have an effect on other types of accident as well, has not been investigated. Hauer rejects the assumption made in the odds ratio method that single vehicle daytime accidents are unaffected by DRL. Implicit in his argument is the suggestion that when DRL is used, vehicles on a collision course detect each other earlier and more frequently take evasive action that results in a single vehicle accident. Hauer further suggests that using DRL leads to more burned out light bulbs, which may in turn affect the number of accidents at night. Both hypotheses could have been tested in experimental vehicle fleet studies. Unfortunately none of these studies include such a test. Most of the vehicle fleet studies do not contain any information at all concerning nighttime accidents and some not even concerning single vehicle daytime accidents.

An Austrian before-and-after study on various vehicle fleets (Kuratorium für Verkehrssicherheit 1993) included accidents at night as well as daytime accidents. The results of this study do not indicate that the use of DRL lead to more single vehicle accidents or more accidents at night. There was a small decline in single vehicle daytime accidents and nighttime accidents, but a greater decline in multi-party daytime accidents. This study, although non-experimental, does not support the hypothesis that DRL affects all types of accident and not just multi-party daytime accidents.

One final point that deserves brief mention concerns the possibility that the results of the meta-analysis presented in this paper are biased because the meta-analysis does not include all studies that have evaluated the effects of DRL. Table 2 lists four results that could not be included in the meta-analysis. All four results refer to the intrinsic effect of DRL. The results are difficult to interpret, since three of the four cases do not state clearly what types of accident they refer to. The results range from a 7 to a 39% accident reduction, with an unweighted mean of 22% accident reduction. This is consistent with the results of the meta-analysis.

CONCLUSIONS

The major conclusions of the research reported in this paper are:

1. A meta-analysis has been made of 17 studies that have evaluated the effects on accidents of using daytime running lights (DRL) on cars. The logodds method of meta-analysis was applied to estimate a weighted mean effect of DRL on multi party daytime accidents. The sensitivity of the estimated effect was tested with respect to (i) study design, (ii) definition of the variable intended to measure the effect of DRL (the dependent variable) and (iii) whether the estimate of the DRL effect referred to each car (intrinsic effect) or to the total number of accidents in a country (aggregate effect).
2. The intrinsic effect of DRL was found to be very robust with respect to both study design and definition of the dependent variable. The best estimate of the intrinsic effect of DRL on cars is a 10–15% reduction in the number of multi-party daytime accidents.
3. All studies of the aggregate effect of DRL are non-experimental before-and-after studies. This study design does not take account of all confounding factors that are likely to be present. The aggregate effects of DRL were more sensitive to the definition of the dependent variable than the intrinsic effects. They were also smaller, ranging from 3 to 12% reduction in the number of multi-party daytime accidents.
4. There was no evidence of a dose–reponse relationship in the effects of DRL laws, in the sense that large increases in DRL use lead to greater reductions in the number of accidents than small increases in DRL use. There is probably a relationship between the latitude of a country and the effects of DRL, but the exact shape of this relationship cannot be inferred from currently available evidence.
5. It has been suggested that DRL may affect not just multi-party daytime accidents, but single vehicle daytime accidents and nighttime accidents as well. No stringent test has been made to determine what types of accident are affected by DRL, but evidence from a non-experimental fleet study suggests that DRL does not affect single vehicle accidents or nighttime accidents.

REFERENCES

Allen M.J. Running light questionnaire. Am. J. Optom. Arch. Am Acad. Optom. 42 (0):164–167; 1965.

Allen M.J. The current status of automobile running lights. J. Am. Optom. Assoc. 50:179–180; 1979.

Allen M.J.; Clark J.R. Automobile running lights — a research report. Am. J. Optom. Arch. Am. Acad. Optom. 41:293–315; 1964.

Andersson, K.; Nilsson, G. The effects on accidents of compulsory use of running lights during daylight in Sweden.

VTI-report 208A. Linköping, Sweden, National Road and Traffic Research Institute; 1981.

Andersson, K.; Nilsson, G.; Salusjärvi, M. Effekt på trafikolyckor av rekommenderad och påkallad användning av varselljus i Finland. VTI-rapport 102. Linköping, Sweden, Statens väg- och trafikinstitut; 1976.

Arora, H.; Collard, D.; Robbins, G.; Welbourne, E. R.; White, J. G. Effectiveness of daytime running lights in Canada. Report TP 12298 (E). Ottawa, Canada, Transport Canada; 1994.

Attwood, D. A. The potential of daytime running lights as a vehicle collision countermeasure. SAE Technical Paper 810190. Warrendale, PA, Society of Automotive Engineers; 1981.

Cantilli, E. J. Daylight 'lights-on' plan by Port of New York Authority. Traffic Engineeering, 17, December; 1965.

Cantilli E.J. Accident experience with parking lights as running lights. Highway Res. Rec. 332:1–13; 1970.

Elvik R. The effects on accidents of compulsory use of daytime running lights for cars in Norway. Accid. Anal. Prev. 25:383–398; 1993.

Fleiss, J. L. Statistical methods for rates and proportions. Revised edition. New York, John Wiley and Sons; 1981.

Hansen, L. K. Kørelys i Danmark. Effektvurdering af påbudt kørelys i dagtimerne. Notat 2/1993. København, Denmark, Rådet for Trafiksikkerhedsforskning; 1993.

Hansen, L. K. Kørelys. Effektvurdering baseret på uheldstal efter knap 3 års erfaring med kørelys. Arbejdsrapport 1/1995. København, Denmark, Rådet for Trafiksikkerhedsforskning; 1995.

Hauer, E. Before-and-after studies in road safety. Estimating the effect of highway and traffic engineering measures on road safety. Lecture notes, March 1995. Department of Civil Engineering, University of Toronto, Ontario, Canada; 1995.

Hocherman, I.; Hakkert, A. S. The use of daytime running lights during the winter months in Israel — evaluation of a campaign. Proceedings of the third workshop of ICTCT in Cracow, Poland, November 1990, 123–131. Bulletin 94, University of Lund, Sweden, Lund Institute of Technology, Department of Traffic Planning and Engineering; 1991.

Hollo, P. Changes of the DRL-regulations and their effect on traffic safety in Hungary. Paper presented at the conference Strategic Highway Safety Program and Traffic Safety, Prague, The Czech Republic, September 20–22, 1995. Preprint for sessions on September 21; 1995.

Koornstra, M. J. Daytime running lights: its safety revisited. SWOV Report D-93-25. Leidschendam, The Netherlands, SWOV Institute for Road Safety Research; 1993.

Koornstra, M. J. Annotated review of recent DRL results since 1991. Unpublished manuscript dated July 19, 1995. Leidschendam, The Netherlands, SWOV Institute for Road Safety Research; 1995.

Kuratorium für Verkehrssicherheit, Institut für Verkehrstechnik und Unfallstatistik. Fahren mit Licht — auch am Tag. Analyse der verkehrsunfälle beim Kraftwagendienst der Österreichischen Bundesbahnen und bei der Österreichischen Post- und Telegraphenverwaltung nach Einführung der Verwendung des Abblendlichtes auch am Tag. Wien, Austria, August; 1993.

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

OECD — Scientific Expert Group. Targeted Road Safety Programmes. Paris, OECD; 1994.

Rideng, A. Transportytelser i Norge 1946–1994. TØI-rapport 303. Oslo, Norway, Transportøkonomisk institutt; 1995.

Sparks, G. A.; Neudorf, R. D.; Smith, A. E. An analysis of the use of daytime running lights in the CVA fleet in Saskatchewan. Traffic Safety Services Department, SaskAuto, Saskatoon, Saskatchewan; 1989.

Sparks G.A.; Neudorf R.D.; Smith A.E.; Wapnuan K.R.; Zador P.L. The effect of daytime running lights on crashes between two vehicles in Saskatchewan: a study of a government fleet. Accid. Anal. Prev. 25:619–625; 1993.

Stein, H. Fleet Experience with Daytime Running Lights in the United States. SAE Technical Paper 851239. Warrendale, PA, Society of Automotive Engineers; 1985.

Theeuwes J.; Riemersma J. Daytime running lights as a vehicle collision countermeasure: the Swedish evidence reconsidered. Accid. Anal. Prev. 27:633–642; 1995.

Vaaje, T. Kjørelys om dagen reduserer ulykkestallene. Arbeidsdokument av 15.8.1986, Q-38 CRASH. Oslo, Norway, Transportøkonomisk institutt; 1986.

Paper 5

Pergamon

Accid. Anal. and Prev., Vol. 29, No. 2, pp. 191–199, 1997

0001-4575/97 $17.00 + 0.00

PII: S0001-4575(96)00070-X

EVALUATIONS OF ROAD ACCIDENT BLACKSPOT TREATMENT: A CASE OF THE IRON LAW OF EVALUATION STUDIES?

RUNE ELVIK

Institute of Transport Economics, PO Box 6110, Etterstad N-0602, Oslo , Norway

(Received 24 April 1996; in revised form 25 August 1996)

Abstract—Numerous evaluation studies have reported large accident reductions when road accident blackspots are treated. A critical examination of these studies reveals that many of them do not account for the effects of well known confounding factors, like the regression-to-the-mean effect that is likely to occur at road accident blackspots. This paper shows that the more confounding factors evaluation studies account for, the smaller becomes the accident reduction attributed to blackspot treatment. Studies that account for both regression-to-the-mean and a possible accident migration to neighbouring untreated sites do not show any net accident reduction at all. This tendency conforms to the so called Iron Law of evaluation studies, which states that the more confounding factors an evaluation study accounts for, the less likely it is to show beneficial effects of the programme evaluated. Possible explanations of accident migration are discussed in the paper. © 1997 Elsevier Science Ltd. All rights reserved

Keywords—Road accident, Blackspot, Treatment, Evaluation study, Meta-analysis

INTRODUCTION

The identification, analysis and treatment of road accident blackspots is widely regarded as one of the most effective approaches to road accident prevention. In its *Guidelines for Accident Reduction and Prevention*, the Institution of Highways and Transportation (1990) states (p. 2):

> It is well established that considerable safety benefits may accrue from application of appropriate road engineering or traffic management measures at hazardous road locations. Results from such applications at "blackspots" demonstrating high returns from relatively low cost measures have been reported worldwide.

It is correct that a number of studies from different parts of the world have reported large reductions in the number of accidents when safety measures were introduced at road accident blackspots. Many of these studies are, however, simple before-and-after studies that do not take account of any confounding factors that might affect the number of accidents. In particular, it is known that an abnormally high recorded number of accidents at a certain location can result from random fluctuation in the number of accidents. To the extent that an abnormally high number of accidents, or an abnormally high accident rate, is the result of random fluctuations, a subsequent decline in the number of accidents (or the accident rate) must be expected even if no safety treatment is applied. This phenomenon is known as regression to the mean and has been found in several studies (see, for example, Forbes, 1939; Brüde and Larsson, 1982; Hauer and Persaud, 1983).

This source of confounding is particularly important in evaluations of road accident blackspot treatment. Rossi and Freeman (1985) have proposed what they term "The Iron Law of Evaluation Studies" in these terms (p. 391): "The better an evaluation study is technically, the less likely it is to show positive program effects". The purpose of this paper is to investigate whether the Iron Law of Evaluation Studies applies to studies that have evaluated the effects on safety of road accident blackspot treatment. To what extent do the effects on accidents attributed to blackspot treatment disappear as more confounding factors are controlled in evaluation studies? In order to shed light on this question, a meta-analysis has been made of 36 studies that have evaluated the effects on accidents of road accident blackspot treatment.

DATA AND METHOD

Evaluation studies included

A total of 36 evaluation studies are included. The studies were retrieved by means of a systematic

literature survey. The literature survey consisted of scanning peer reviewed journals like *Accident Analysis and Prevention*, *ITE-Journal*, *Journal of Safety Research*, *Traffic Engineering and Control* and *Transportation Research Record*. In addition, publications issued by highway agencies and research institutes in the Nordic Countries were included, as well as publications of highway agencies and major institutions in Australia, Great Britain and the United States.

Studies were included if: (1) they stated that the treatment evaluated was applied at an 'accident blackspot' or because of a 'bad accident record' or an 'abnormal accident experience', (2) they reported the number of accidents their results were based on and (3) the research design was described in sufficient detail to determine which confounding factors a study controlled for. A number of different formal, statistical definitions of a road accident blackspot have been proposed (Hauer, 1996). However, most evaluation studies describe the selection of locations for treatment only in general terms and do not state explicitly if a formal, statistical blackspot definition was applied. It was therefore not possible to confine the analysis to studies relying on a formal blackspot concept. Studies included are listed in Appendix A.

Statistical weighting of results

Each of the studies included contains one or more results of an evaluation of the effects on safety of one or several treatments carried out at one or several locations. All studies are non-experimental before-and-after studies. Some of the studies included comparison groups in addition to the treated sites. Weighted mean results were estimated by means of the logodds method of meta-analysis (Fleiss, 1981). Each result was assigned a statistical weight inversely proportional to the variance of the logodds of the estimated effect:

$$W_i = 1/(1/B_i + 1/A_i)$$

where B_i denotes the number of accidents at treated sites in the before-period for result i and A_i denotes the corresponding number of accidents in the after-period. This choice of weights for each result minimizes the variance of the weighted mean. In studies using comparison sites, the variance of the estimated effect of treatment depends on the number of accidents at both the treatment and comparison sites. However, many evaluation studies do not state the number of accidents recorded at comparison sites. Hence, the contribution of fluctuations in comparison group accidents to the variance of the estimated effects of treatment had to be ignored. This raises the value of the statistical weights assigned to results of studies using comparison groups. For example, the statistical weight of a result based on 38 accidents before and 22 accidents after in the treatment group, and 245 accidents before and 218 accidents after in the comparison group is 13.9, if accidents in the comparison group are ignored, but 12.4 if they are included when calculating the statistical weight.

In order to test if the method of estimating statistical weights might introduce bias in the weighted mean results, the weighted results were compared to simple unweighted mean results. The weighted and unweighted results were very similar and no systematic bias in any direction was found. Only the weighted mean results are presented in this paper, as they are statistically more precise than unweighted results. Weighted mean safety effects for groups of evaluation studies were estimated according to the following formula (Fleiss, 1981):

$$\text{Weighted mean safety effect} = \exp[(\Sigma \ln(\theta_i) \cdot W_i)/\Sigma W_i]$$

where exp denotes the exponential function, ln the natural logarithm, θ_i each estimate of treatment effect and W_i the statistical weight of each estimate of treatment effect. A 95% confidence interval for the weighted mean safety effect was estimated by applying methods described by Fleiss (1981).

Controlling for confounding factors

Confounding factors are all factors that weaken the basis for inferring a causal relationship between blackspot treatment and changes in road safety. Confounding factors represent alternative interpretations to the findings and ought ideally to be eliminated. Complete control of confounding factors is possible only by using an experimental research design, involving the random assignment of study units to a treatment or non-treatment condition. In non-experimental research, control of confounding factors will always be incomplete and imperfect. But the more known confounding factors a study controls for, the better becomes the basis for concluding that observed changes in road safety were caused by the treatment rather than the confounders. The confounding factors considered in this study are:

(1) Changes in traffic volume
(2) General trends in the number of accidents
(3) Regression to the mean
(4) Accident migration

These are some of the most important known confounding factors present in non-experimental before-and-after studies of road accident blackspot treatment.

Changes in traffic volume are usually controlled for by estimating accident rates (accidents per million vehicle kilometers or per million passing or entering

vehicles) and using changes in these as the measure of effect iu evaluation studies. It is normally assumed that the number of accidents is a linear function of traffic volume (Hauer, 1995). This assumptiou is not always correct. Hence, the use of changes in accident rates as the measure of effect in before-and-after studies does not necessarily remove the effects of chauges in traffic volume on the number of accidents. In this paper, however, evaluation studies using changes in accident rates as the measure of effect have been classified as controlling for chauges in traffic volume.

The presence of general trends in the uumber of accidents is usually controlled for by using a comparison group, often consisting of the total number of accidents in a country or in the area where the treated blackspots are located. The use of a comparison group relies on the assumption that changes in the number of accidents in the comparison group correctly predicts the changes that would have occurred at the treated sites in the absence of treatment. As shown by Hauer (1991), this assnmption will not always be correct. On the other hand, this assumption has traditionally been accepted, at least as approximately correct. Hence, studies using comparison gronps have been classified as taking account of general trends in the number of accidents, except when the comparison group consisted of untreated blackspots exclusively (see comment below).

Two methods have been used to control for regression to the mean in studies evaluating blackspot treatments. One method is to nse a comparison group of untreated blackspots. Changes in the number of accidents at nntreated blackspots are assumed to reflect mainly regression to the mean, rather than general trends. This interpretation is accepted in this paper. The other method of controlling for regression to the mean is to estimate this effect by means of a statistical model (Brüde and Larsson, 1982; Hauer, 1980, 1986, 1992). There are several models that differ in both assumptions and estimatiou techniques. A detailed discussion of these differences is beyond the scope of this paper. In this paper, all studies using one of the two methods for removing regression to the mean have heen classified as controlling for this confounding factor.

Accident migration denotes the transfer of accidents from the blackspots to surrounding locations as a result of blackspot treatment. The usual way of controlling for accident migration is to include the surrounding locations to which accidents are supposed to migrate in the treated group. Changes in the number of accidents for the enlarged group of locations will then reflect both the treatment effect at the treated sites and the accident migration effect at the surrounding sites. Some studies in addition estimate regression to the mean at both treated and surrounding sites by means of a statistical model, while other studies accept the recorded number of accident at treated and surrounding sites as unbiased estimates of the expected number of accidents. Studies using either of these designs have been classified as controlling for accident migration.

Design of analysis

Figure 1 shows the design of analysis used in the present study.

Blackspots were classified as road sections, junctions (intersections) and unspecified types of locations. For each type of blackspot, a distinction was made between injury accidents, accidents involving property damage only (PDO-accidents) and accidents of unspecified severity (generally including both injury and PDO-accidents in unknown proportions). For each type of blackspot and level of accident severity, the results of evaluation studies were compared with respect to which of the confounding factors, or combination of confonnding factors, that were controlled.

RESULTS

All types of treatment combined

Table 1 shows the weighted mean results of studies that have evaluated the safety effects of road accident blackspot treatment, expressed in terms of percent change in the number of accidents attributed to the treatment. In Table 1 all types of treatment have been combined.

The results presented in Table 1 show that the size of the effect attributed to blackspot treatment in evaluation studies varies substantially depending on which confounding variables are controlled. This is seeu by comparing the results printed in boldface italics in Table 1. In general, stndies that do not control for any confounding factors find the largest effects of treatment. Studies that control simultaneously for general trends in the number of accidents, regression to the mean and accident migratiou do not find any statistically significant changes in the number of accidents due to blackspot treatment. The more confounding factors accounted for, the smaller the effect attributed to blackspot treatment becomes. This finding applies both to junctions and other locations and both to injury accidents and PDO-accidents. Most of the evidence refers to iujury accidents. The results for PDO-accidents are more uncertain.

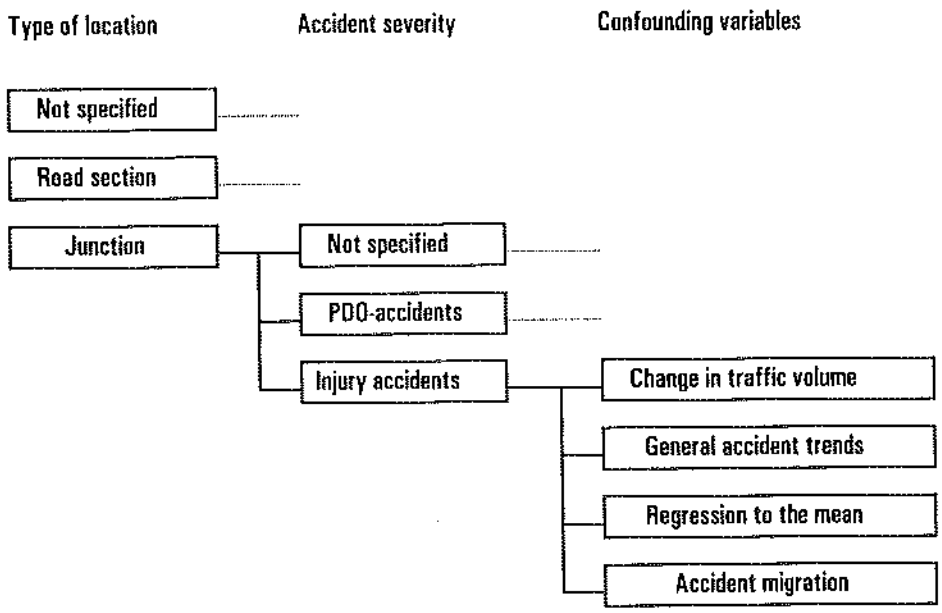


Fig. 1. Classification of types of location, accident severity and confounding variables controlled.

Some of the results are based on just one or two studies. In order to increase sample size, results that refer to injury accidents were combined for all types of location. The combined results are shown in the bottom of Table 1 (the results that refer to all types of location). There is a very clear tendency for the effect attributed to treatment to become smaller as more confounding factors are controlled. Studies that account for accident migration are, however, the only category which do not show statistically significant accident reductions following blackspot treatment.

Results for different kinds of treatment

An objection to this analysis is that different kinds of treatment are likely to have different effects; hence it does not make sense to estimate the weighted mean safety effects of different treatments combined. Estimates of effects ought to be made for each kind of treatment by itself. Table 2 presents an analysis of five common safety treatments at junctions, for studies with different degrees of control of confounding factors.

The tendency found when all treatments were combined is reproduced when different treatments are studied by themselves. In general, the more confounding factors studies account for, the smaller are the effects attributed to the treatment. This pattern is evident for all five treatments included in Table 2. Once again, however, some of the estimates are based on just one of two studies. There were too few studies to do a similar analysis of different treatments applied to road sections.

DISCUSSION

Road accident blackspot treatment has for a long time been accepted as an effective way of preventing road accidents. The results presented in this paper, if taken at face value, indicate that this belief is unfounded. The belief that blackspot treatment is particularly effective seems to have rested on an uncritical acceptance of the results of simple before and after studies that fail to account for confounding factors that may explain the observed reductions in the number of accidents or the accident rate.

Today, most researchers accept that in non-experimental before-and-after studies of treatments at locations that were selected for treatment because of their bad accident record, it is necessary as a minimum to remove the effects of changes in traffic volume, general trends in the number of accidents and regression to the mean before anything can be concluded with respect to the effects of the treatment. Some researchers were aware of the need to remove the effects of regression to the mean as early as 1968. Thus, Tamburri et al. (1968)(p. 38):

> The possibility always exists that an improvement project may have been initiated because of an unusually high accident experience which was merely a reflection of a temporary condition in the before period. In such cases, even if nothing had been done, an accident reduction would probably have been observed in the after period (regression to the mean theory). The possibility of such an influence was investigated.

Tamburri et al. (1968) go on to state that it was found that some locations had a permanent high level

Table 1. Weighted mean effects of blackspot treatment on the number of accidents by type of location, accident severity and confounding variables controlled

Type of location	Accident severity	Confounding variables controlled	Number of studies	Proportion of statistical weights	Percent change in accidents: Lower 95%	Best estimate	Upper 95%
Junction	Injury accidents	None	6	0.048	−66	*−60*	−54
		Traffic volume	5	0.093	−49	*−43*	−37
		Trend	4	0.359	−36	*−33*	−29
		Regression to mean	1	0.028	−44	*−31*	−16
		Trend, regression to mean	2	0.023	−31	*−14*	+7
		Trend, accident migration	1	0.449	−8	*−4*	+1
			19	1.000	−26	−24	−21
Junction	PDO-accidents	None	5	0.405	−46	*−37*	−25
		Traffic volume	3	0.483	−51	*−42*	−33
		Trend, regression to mean	1	0.112	−27	*+0*	+38
			9	1.000	−43	−36	−29
Junction	Not specified	None	3	0.265	−48	*−42*	−36
		Traffic volume	1	0.032	−60	*−46*	−29
		Trend, regression to mean	3	0.490	−46	*−42*	−38
		Trend, regression to mean, accident migration	1	0.213	−12	*−2*	+9
			8	1.000	−39	−36	−32
Road section	Injury accidents	None	3	0.123	−57	*−51*	−43
		Traffic volume	2	0.030	−23	*+3*	+37
		Trend	6	0.332	−19	*−12*	−3
		Trend, traffic volume	1	0.127	−42	*−33*	−23
		Trend, regression to mean	1	0.018	−61	*−44*	−18
		Trend, accident migration	1	0.370	−6	*+2*	+11
			14	1.000	−21	−16	−12
Road section	PDO-accidents	None	2	0.031	−95	*−92*	−86
		Traffic volume	1	0.084	−50	*−29*	−0
		Trend, traffic volume	1	0.787	−36	*−29*	−20
		Trend, regression to mean	1	0.098	−39	*−16*	+15
			5	1.000	−39	−32	−25
Not stated	Injury accidents	Traffic volume, regression to mean	2	0.103	−34	*−24*	−12
		Trend, regression to mean	4	0.392	−22	*−16*	−10
		Trend, regression to mean, accident migration	3	0.505	−7	*+0*	+7
			9	1.000	−13	−9	−5
All types	Injury accidents	None	8	0.052	−60	*−55*	−50
		Traffic volume	5	0.054	−45	*−39*	−32
		Trend	6	0.259	−30	*−28*	−24
		Regression to mean	3	0.041	−34	*−26*	−17
		Trend, traffic volume	1	0.029	−42	*−33*	−23
		Trend, regression to mean	7	0.119	−22	*−17*	−11
		Trend, accident migration	1	0.313	−6	*−2*	+2
		Trend, regression to mean, accident migration	3	0.133	−6	*+0*	+7
			34	1.000	−20	−18	−16

of accident experience, not just during the few years that were the before period in their study. For other locations, planning took so long that the number of accidents had already regressed to a more normal level when the safety treatment was carried out. In general, prolonging the before and after periods will water down the regression to the mean effect, but not remove it altogether (Nicholsou, 1988). On the other hand, long before and after periods enlarge the influence of general trends in accidents on the results of a study.

The need to control for regression to the mean in before-and-after studies of safety measures introduced at high accident locations can be deduced from elementary statistical theory. Despite this fact, studies that do not remove this important source of bias are still published (see, for example, the papers by Wong, 1990 and Proctor, 1995).

The possibility of accident migration, and the consequent need to control for it, was first raised by Boyle and Wright (1984). Their paper was criticized for not controlling for regression to the mean (McGuigan, 1985). Subsequent papers by Maher (1987, 1990) suggested that accident migration is a statistical artefact, generated mainly by a combination of regression to the mean downwards of abnormally high accident counts at treated sites and regression to the mean upwards of abnormally low

Table 2. Weighted mean safety effects of some common blackspot treatments in junctions by confounding variables controlled

Type of location	Treatment	Confounding variables controlled	Number of studies	Proportion of statistical weights	Percent change in accidents: Lower 95%	Best estimate	Upper 95%
Junction	Channelization	None	3	0.324	−58	*−52*	−45
		Traffic volume	3	0.138	−52	*−40*	−25
		Trend	1	0.193	−50	*−40*	−28
		Trend, regression to mean	1	0.077	−24	*+2*	+37
		Trend, accident migration	4	0.268	−12	*+2*	+20
			12	1.000	−38	−32	−27
Junction	Four way stop	Traffic volume	1	0.020	−85	*−76*	−64
		Trend, regression to mean	2	0.669	−50	*−46*	−41
		Trend, regression to mean, accident migration	1	0.311	−12	*−2*	+9
			4	1.000	−40	−36	−32
Junction	Traffic signals	None	1	0.025	−84	*−70*	−44
		Trend	5	0.443	−56	*−49*	−40
		Trend, regression to mean	2	0.073	−22	*+12*	+62
		Trend, accident migration	1	0.459	−20	*−7*	+8
			9	1.000	−36	−29	−22
Junction	Traffic signal improvements	None	2	0.140	−50	*−44*	−37
		Traffic volume	1	0.024	−60	*−47*	−29
		Trend	4	0.386	−31	*−26*	−21
		Trend, regression to mean	1	0.008	−26	*+21*	+98
		Trend, accident migration	1	0.442	−9	*−3*	+3
			9	1.000	−24	−20	−17
Junction	Surface friction improvement	None	1	0.630	−44	*−35*	−26
		Trend	1	0.075	−79	*−68*	−54
		Traffic volume, regression to mean	2	0.295	−44	*−31*	−16
			4	1.000	−44	−38	−31

accident counts at surrounding sites. The studies of Persaud (1987), Mountain and Fawaz (1989, 1992) and Mountain et al. (1992, 1994) have, however, controlled for regression to the mean, but nevertheless find some support for a hypothesis of accident migration. This raises the question of whether plausible explanations of accident migration are known or can be imagined.

Boyle and Wright (1984) proposed the following explanation: "It can be hypothesized that where an accident blackspot is treated, drivers will be subjected to fewer "near-misses" at that site, and consequently will be less aware of the need for caution. This reduced awareness may persist for some distance downstream, and consequently the risk of an accident in the area surrounding the blackspot may be increased." They do not produce any evidence to support this hypothesis. Several considerations suggest that the hypothesis is not a very plausible explanation for accident migration.

There seems to be an element of logical inconsistency in the hypothesis. If it is true that exposure to near-misses induces driver caution, and if, as the hypothesis seems to assume, the number of accidents is positively related to the number of near-misses, it is difficult to see how an accident blackspot could arise in the first place. If drivers experienced more near-misses before the blackspot was treated, their level of caution at that site, ought, according to Boyle and Wright, to have been higher before treatment than after. This makes it difficult to understand how treating a blackspot could really reduce the number of accidents at the blackspot itself. Boyle and Wright suggest that a reduced level of caution persists 'some distance' downstream. Why should this be the case, if drivers continuously adapt their level of caution to the number of near-misses they experience at any site?

The mechanism suggested by Boyle and Wright rests on the assumption that the number of accidents is related to the level of driver caution. The number of near-misses is obviously one of the factors that may influence the level of driver caution, but it is unlikely to be the only factor, and perhaps not even a very important one. In a study in Uppsala in Sweden, Johansson and Naeslund (1986) found that there was no correlation at all between the subjective hazard ratings drivers gave to specific locations in the city and the accident experience at those sites. The worst blackspots were not rated by drivers as particularly hazardous; perhaps that is one the reasons why these sites developed into blackspots. At sites that were perceived as hazardous, there were few accidents because drivers were careful. The perception of a site as hazardous was related to

sight distance, traffic volume and driving speed. Unfortunately, the study did not examine the influence of near-misses on subjective hazard ratings.

Persaud (1987) suggests that changes in driver expectancy may explain accident migration, when most intersections in Philadelphia were converted to four way stop control. Once four way stop control became the norm, drivers started to expect drivers entering from the major road in intersections with two way stop control to stop as well. Persaud does not produce direct evidence of such changes in driver expectancy, but the changes observed in accident counts for intersections with different types of traffic control (four way stop, two way stop, traffic signals) support the hypothesis.

It is not known if the mechanism suggested by Persaud applies to blackspot treatment in general. It does not seem likely that every kind of treatment will lead to similar changes, or any changes at all, in driver expectancy. The signing of hazardous curves may be a case in point. If hazard warning signs are put up in almost every curve, two things may happen. One, drivers will not take the signs seriously and two, the few curves where no hazard warning sign has been put up will become more surprising and therefore perhaps more prone to accidents. But if the use of hazard warning signs at curves is more restrictive, such adaptations seem less likely to occur.

More research is clearly needed to establish more firmly how real and widespread accident migration is. The changes in driver perception, expectancy or behaviour that may lead to accident migration have to be studied more in detail before it can be concluded that accident migration is a real phenomenon that will occur often or whenever accident blackspots are treated. The evidence presented in this paper is inconclusive.

CONCLUSIONS

The main conclusions of the research reported in this paper are:

(1) Based on before-and-after studies reporting large reductions in the number of accidents following road accident blackspot treatment, this is widely believed to be a particularly effective approach to road accident prevention. Some of these studies are simple before-and-after studies that do not account for any of the confounding factors known to affect the results of such studies.

(2) A meta-analysis of 36 before-and-after studies of road accident blackspot treatment was performed in order to determine how the degree of control for known confounding factors affected the results of those studies. Four known confounding factors were considered: (i) changes in traffic volume, (ii) general trends in the number of accidents, (iii) regression to the mean and (iv) accident migration. The logodds method of meta-analysis was used.

(3) It was found that the results of before-and-after studies of road accident blackspot treatment depend strongly on which of the confounding factors studies control for. Large reductions in the number of accidents, generally in the order of 50–90%, were found in studies not controlling for any confounding factors. The more confounding factors studies controlled for, the smaller were the effects attributed to blackspot treatment. Studies simultaneously controlling for general trends, regression to the mean and accident migration did not find any statistically reliable effect of blackspot treatment on the number of accidents.

(4) The need to control for changes in traffic volume, general trends in accident occurrence and regression to the mean in before-and-after studies of blackspot treatment is accepted by most researchers. Accident migration is a more controversial phenomenon. More research is needed to determine how widespread accident migration is and the mechanisms explaining it.

REFERENCES

Boyle, A. J. and Wright, C. C. (1984) Accident "migration" after remedial treatment at accident blackspots. *Traffic Engineering and Control* **25**, 260–267.

Brüde, U. and Larsson, J. (1982) Regressionseffekt. Några empriska exempel baserade på olyckor i vägkorsningar. (VTI-rapport 240). Statens väg-och trafikinstitut (VTI), Linköping, Sweden.

Christensen, P. (1988) Utbedringer av ulykkespunkter på riksveger og kommunale veger i perioden 1976–1983. Erfaringsrapport. (TØI-rapport 0009). Transportøkonomisk institutt, Oslo.

Corben, B. F., Ambrose, C. and Wai, F. C. (1990) Evaluation of accident black spot treatments. (Report 11).: Accident Research Centre, Monash University, Melbourne, Australia.

Dearinger, J. A. and Hutchinson, J. W. (1970) Cross section and pavement surface. In *Traffic Control and Roadway Elements—Their Relationship to Highway Safety*, Chapter 7. Revised Edition. Highway Users Federation for Safety and Mobility, Washington, DC.

Duff, J. T. (1971) The effect of small road improvements on accidents. *Traffic Engineering and Control* **12**, 244–245.

Elvik, R. (1985) Regresjonseffekt i ulykkespunkter. En empirisk undersøkelse på riksveger i Vest-Agder. (Arbeidsdokument av 9.9.1985 (prosjekt O-1146)).: Transportøkonomisk institutt, Oslo.

Exnicios, J. F. (1967) Accident reduction through channelization of complex intersections. In *Improved Street Utilization Through Traffic Engineering*, pp. 160–165.

Tamburri, T.N., Hammer, C. J., Glennon, J. C. and Lew, A. (1968) Evaluation of minor improvements. *Highway Research Record* **257**, 34–79.

Vodahl, S. B. and Johannessen, S. (1977) Ulykkesfrekvenser i kryss. Arbeidsnotat nr 7. Resultater av før/etterundersøkelsen. (Oppdragsrapport 178). Norges Tekniske Høgskole, Forskningsgruppen, Institutt for samferdselsteknikk, Trondheim.

Værø, H. (1992a) *Effekt af sortpletbekæmpelse i Hillerod.* København, Vejdirektoratet, Trafiksikkerhedsafdelingen.

Værø, H. (1992b) *Effekt af sortpletbekæmpelse i Nyborg.* København, Vejdirektoratet, Trafiksikkerhedsafdelingen.

Værø, H. (1992c) *Effekt af sortplethekæmpelse i Silkeborg.* København, Vejdirektoratet, Trafiksikkerhedsafdelingen.

Værø, H. (1992d) *Effekt af sortplethekæmpelse i Skælskor.* København, Vejdirektoratet, Trafiksikkerhedsafdelingen.

Wilson, J. E. (1967) Simple types of intersection improvements. In *Improved Street Utilization Through Traffic Engineering*, pp. 144–159. (Special Report 93). Highway Research Board, Washington, DC.

Wong, S-Y. (1990) Effectiveness of pavement grooving in accident reduction. *ITE Journal* **60**(6), 34–37.

APPENDIX A

List of studies included in meta-analysis and characteristics of each study.

Authors (Year)	Country	Types of locations studies	Types of treatment specified	Confounding variables controlled
Exnicios (1967)	United States	Junctions	Yes	None
Malo (1967)	United States	Junctions	Yes	None; traffic volume
Wilson (1967)	United States	Junctions	Yes	Traffic volume
Tamburri et al. (1968)	United States	Junctions	Yes	Traffic volume
Hammer (1969)	United States	Junctions; sections	Yes	None; traffic volume
Dearinger and Hutchinson (1970)	United States	Junctions; sections	Yes	None
Duff (1971)	Great Britain	Junctions; sections	Yes	Noue
Hatherly and Lamb (1971)	Great Britain	Junctions	Yes	None
Karr (1972)	United States	Sections	Yes	Treud; traffic volume
Hvoslef (1974)	Norway	Junctions; sections	Yes	Trend
OECD (1976)	France	Junctions	Yes	Trend
Hatherly and Young (1977)	Great Britain	Junctinns	Yes	Regression
Vodahl and Jnhannessen (1977)	Norway	Junctions	Yes	Trend; regression
Jørgensen (1979)	Denmark	Junctions; sections	Yes	Trend
Statens vegvesen (1983)	Norway	Sections	Yes	Trend
Boyle and Wright (1984)	Great Britain	Junctions; sections	Yes	Trend; migration
Elvik (1985)	Norway	Not stated	No	Regression
Lovell and Hauer (1986)	United States	Junctions	Yes	Trend; regression
Persaud (1987)	United States	Junctions	Yes	Trend; regression; migration
Christensen (1988)	Norway	Not stated	No	Regression
Mountain and Fawaz (1989)	Great Britain	Not stated	No	Trend; regression; migration
Corben et al. (1990)	Australia	Jnnctions; sections	Yes	Trend
Flagstad (1990)	Norway	Junctions; sections	No	Traffic volume
Woug (1990)	United States	Sections	Yes	Noue
Lalani (1991)	Uuited States	Jnnctions	Yes	Trend
Retting (1991)	United States	Sections	No	Nnne
Sørensen (1991)	Denmark	Junctions	Yes	None
Kolster Pedersen et al. (1992)	Denmark	Junctions; sectinns	Yes	Nonc
Mountain and Fawaz (1992)	Great Britaiu	Not stated	No	Trend; regressiou; migration
Mountain et al. (1992)	Great Britain	Not stated	No	Trend; regression
Værø (1992a,b,c,d)	Denmark	Junctions; sections	Yes	Trend; regression
Holmskov and Lahrmaun (1993)	Denmark	Junctions; sections	Yes	Trend; regression
Gregory and Jarrett (1994)	Great Britain	Not stated	No	Trend; regression
Mountain et al. (1994)	Great Britain	Not stated	No	Trend; regression; migration
Legassick (1995)	Great Britain	Sections	Yes	Trend
Proctor (1995)	Great Britain	Sections	Yes	None

(Special Report 93).: Highway Research Board, Washington, DC.
Flagstad, K. (1990) Før-etter analyse av trafikksikkerhetstiltak i Bergen. Hovedoppgave i samferdselsteknikk. Trondheim. Norges Tekniske Høgskole, Institutt for samferdselsteknikk.
Fleiss, J. L. (1981) *Statistical Methods for Rates and Proportions*. 2nd edn. John Wiley and Sons, New York.
Forbes, T.W. (1939) The normal automobile driver as a traffic problem. *Journal of General Psychology* **20**, 471–474.
Gregory, M. and Jarrett, D. F. (1994) The long-term analysis of accident remedial measures at high-risk sites in Essex. *Traffic Engineering and Control* **35**, 8–11.
Hanmer, C.G. (1969) Evaluation of minor improvements. *Highway Research Record* **286**, 33–45.
Hatherly, L.W. and Lamb, D. R. (1971) Accident prevention in London by road surface improvements. *Traffic Engineering and Control* **12**, 524–529.
Hatherly, L.W. and Young, A.E. (1977) The location and treatment of urban skidding hazard sites. *Transportation Research Record* **623**, 21–28.
Hauer, E. (1980) Bias-by selection: Overestimation of the effectiveness of safety countermeasures caused by the process of selection for treatment. *Accident Analysis and Prevention* **12**, 113–117.
Hauer, E. (1986) On the estimation of the expected number of accidents. *Accident Analysis and Prevention* **18**, 1–12.
Hauer, E. (1991) Comparison groups in road safety studies: An analysis. *Accident Analysis and Prevention* **23**, 609–622.
Hauer, E. (1992) Empirical Bayes approach to the estimation of "unsafety": The multivariate regression approach. *Accident Analysis and Prevention* **24**, 457–477.
Hauer, E. (1995) On exposure and accident rate. *Traffic Engineering and Control* **36**, 134–138.
Hauer, E. (1996) Identification of "sites with promise". (Paper 960995).: Transportatinn Research Board, 75th Annual Meeting, Washington, DC.
Hauer, E. and Persaud, B. N. (1983) Common bias in before-and-after accident comparisons and its elimination. *Transportation Research Record* **905**, 164–174.
Holmskov, O. and Lahrmann, H. (1993) Er sortpletbekæmpelse vejen frem? *Dansk Vejtidsskrift* **2**, 3–9.
Hvoslef, H. (1974) *Trafikksikkerhet i Oslo. Problemstilling, analyse og løsninger*. Oslo veivesen, Oslo.
Institution of Highways and Transportation (1990) *Guidelines for Accident Reduction and Prevention*. International Edition. London.
Johansson, R. and Naeslund, A-L. (1986) Upplevd och verklig olycksrisk––möjligheter till påverkan. (TFB-rapport 1986:18). Transportforskningsberedningen, Stockholm, Sweden.
Jørgensen, E. (1979) Sikkerhedsmæssig effekt af mindre anlægsarbejder. Effektstudie. Næstved, Vejdirektoratet, Sekretariatet for Sikkerhedsfremmende Vejforanstaltninger (SSV).
Karr, J. I. (1972) Evaluation of minor improvements—part 8, grooved pavements. Final Report. (Report CA-HY-TR-2151-4-71-00). California Division of Highways, Sacramento, CA.
Kolster Pedersen, S., Knlmala, R., Elvestad, B., Ivarsson, D. and Thuresson, L. (1992) Trafiksäkerhetsåtgärder i Väg-och Gatumiljö. Exempel hämtade från de nordiska länderna under 1980-talet. Nordiske Seminar-og Arbejdsrapporter 1992:607. København, Nordisk Ministerråd.
Lalani, N. (1991) Comprehensive safety program produces dramatic resnlts. *ITE-Journal* **61** (10), 31–34.
Legassick, R. (1995) The case for route studies in road traffic accident analysis investigations. Paper presented at the *Conference on Strategic Highway Research Program and Traffic Safety*, Pragne, The Czech Republic. Preprint for Sessions 21/9.
Lovell, J. and Hauer, E. (1986) The safety effect of conversion to all-way stop control. *Transportation Research Record* **1068**, 103–107.
Maher, M.J. (1987) Accident migration—a statistical explanation. *Traffic Engineering and Control* **28**, 480–483.
Maher, M.J. (1990) A bivariate negative binomial model to explain traffic accident migration. *Accident Analysis and Prevention* **22**, 487–498.
Malo, A. F. (1967) Signal modernization. In *Improved Street Utilization Through Traffic Engineering*, pp. 96–113. (Special Report 93). Highway Research Board, Washington, DC.
McGuigan, D.R.D. (1985) Accident "migration"—or a flight of fancy? *Traffic Engineering and Control* **26**, 229–233.
Mountain, L. and Fawaz, B. (1989) The area-wide effects of engineering measures on road accident occurrence. *Traffic Engineering and Control* **30**, 355–360.
Mountain, L. and Fawaz, B. (1992) The effects of engineering measures on safety at adjacent sites. *Traffic Engineering and Control* **33**, 15–22.
Mountain, L., Fawaz, B. and Sineng, L. (1992) The assessment of changes in accident frequencies on link segments: a comparison of four methods. *Traffic Engineering and Control* **33**, 429–431.
Mountain, L., Fawaz, B., Wright, C., Jarrett, D. and Lupton, K. (1994) Highway improvements and maintenance: their effects on road accidents. Paper presented at the *22nd PTRC Summar Annual Meeting*, Proceedings of Seminar J, pp. 151–161.
Nicholson, A. (1988) Accident count analysis: the classical and alternative approaches. *Proceedings of Session 2, Models for Evaluation, Traffic Safety Theory and Research Methods*, Amsterdam, The Netherlands. SWOV Institute for Road Safety Research.
OECD Road Research Group (1976) *Hazardous Road Locations. Identification and Countermeasures*. OECD, Paris.
Persaud, B.N. (1987) "Migration" of accident risk after remedial blackspot treatment. *Traffic Engineering and Control* **28**, 23–26.
Proctor, S. (1995) An independent review of 3M "Road Safety" products. Paper presented at the *Conference on Strategic Highway Research Program and Traffic Safety*, Prague, The Czech Republic. Preprint for Sessions 22/9.
Retting, R. A. (1991) *Improving Urban Traffic Safety: A Multidisciplinary Approach. Experiences From New York City 1983–1989*. Prepared in conjunction with the Volvo Traffic Safety Award 1991. Thompson Printing, Belleville, NJ.
Rossi, P. H. and Freeman, H. E. (1985) *Evaluation. A Systematic Approach*, 3rd edn. Sage Publications, Beverley Hills, CA.
Statens vegvesen (1983) *Veiledning. Håndbok 115. Analyse av ulykkessteder*. Statens vegvesen, Oslo.
Sørensen, M. (1991) Forsøg med særlig afmærkning af uheldskryds. *Dansk Vejtidsskrift* **5**, 17–19.

Paper 6

Pergamon

Accid. Anal. and Prev., Vol. 30, No. 2, pp. 255–266, 1998

0001-4575/98 $19.00 + 0.00

PII: S0001-4575(97)00076-6

EVALUATING THE STATISTICAL CONCLUSION VALIDITY OF WEIGHTED MEAN RESULTS IN META-ANALYSIS BY ANALYSING FUNNEL GRAPH DIAGRAMS

RUNE ELVIK*

Institute of Transport Economics, P.O. Box 6110, Etterstad, 0602 Oslo, Norway

(Received 2 January 1997; in revised form 25 July 1997)

Abstract—The validity of weighted mean results estimated in meta-analysis has been criticized. This paper presents a set of simple statistical and graphical techniques that can be used in meta-analysis to evaluate common points of criticism. The graphical techniques are based on funnel graph diagrams. Problems and techniqnes for dealing with them that are discussed include: (1) the so-called 'apples and oranges' problem, stating that mean results in meta-analysis tend to gloss over important differences that should be highlighted. A test of the homogeneity of results is described for testing the presence of this problem. If results are highly heterogeneous, a random effects model of meta-analysis is more appropriate than the fixed effects model of analysis. (2) The possible presence of skewness in a sample of results. This can be tested by comparing the mode, median and mean of the results in the sample. (3) The possible presence of more than one mode in a sample of results. This can be tested by forming a frequency distribution of the results and examining the shape of this distribution. (4) The sensitivity of the mean to the possible presence of atypical results (outliers) can be tested by comparing the overall mean to the mean of all results except the one snspected of being atypical. (5) The possible presence of pnblication bias can be tested by visual inspection of funnel graph diagrams in which data points have been sorted according to statistical significance and direction of effect. (6) The possibility of underestimating the standard error of the mean in meta-analyses by using multiple, correlated results from the same study as the unit of analysis can be addressed by using the jack-knife technique for estimating the uncertainty of the meau. Brief examples, taken from road safety research, are given of all these techniques.

Keywords—Meta-analysis, Weighted mean, Funuel graph, Validity, Evaluation stndy

INTRODUCTION

Meta-analysis is increasingly applied to summarize evidence from evaluation studies in road safety. Recent applications include a meta-analysis of road safety mass media campaigns (Elliott, 1993), a meta-analysis of methods used in studies of efforts to control drinking and driving (Wagenaar et al., 1995) and meta-analyses of studies that have evaluated the safety effects of guardrails (Elvik, 1995) and daytime running lights (Elvik, 1996). All these studies include weighted or unweighted estimates of the mean effect on safcty of the interventions studied. A common objection to analyses estimating mean results from several studies, is that these mean results tend to inappropriately mix results that are systematically different and should be kept apart. This argument is known as "the apples and oranges argument" (Glass et al., 1981).

Other objections to meta-analysis include:

(1) the publication bias argument, stating that most meta-analyses rely on published studies only and are therefore vulnerable to publication bias;
(2) the outlier bias argument, stating that estimates of mean effects in meta-analyses are sensitive to outliers;
(3) the inflated sample size argument (Bangert-Drowns, 1986), stating that meta-analyses based on multiple results from each study tend to overstate true sample size and nnderstate the uncertainty of the mean result.

All these criticisms cast doubt on the statistical conclusion validity of mean results in meta-analysis. The statistical conclusion validity of a mean resnlt denotes

*Tel: 00 47 2257 3800; Fax: 0047 2257 0290; e-mail: rune.elvik@toi.no

the extent to which it is unbiased and representative of the sample of results it applies to (Cook and Campbell, 1979).

This paper presents a set of simple statistical and graphical techniques that can be used in meta-analysis to assess the strength of the various threats to statistical conclusion validity discussed above. The techniques are illustrated by means of a sample of road safety evaluation studies relying on the logodds method. These studies have evaluated the effects of headrests in cars on the probability of injury in rearend crashes (O'Neill et al., 1972; States et al., 1972; McLean, 1974; Cameron and Wessels, 1979; Kahane, 1982; Nygren, 1984; Nygren et al., 1985). The results of the studies have been synthesized by means of the logodds method of meta-analysis. For an introduction to the logodds method of meta-analysis, the reader is referred to Fleiss (1981), Fleiss and Gross (1991) and Shadish and Haddock (1994). See also Appendix A.

THE ANALYSIS OF FUNNEL GRAPH DIAGRAMS

It is instructive to start any meta-analysis by preparing a funnel graph diagram (Light and Pillemer, 1984). A funnel graph diagram is a scatter plot of results. The abscissa measures the value of each result, in terms of the size of the change in the dependent variable, for example, the percentage change in the number of accidents or injuries. The ordinate measures the sample size each result is based on, for example, the statistical weight of each result in meta-analyses using the logodds method. Figure 1 shows a funnel graph plot of 30 results of the studies that will be used in this paper to illustrate the statistical techniques that are presented. The number of results are greater than the number of studies, as some studies contain multiple results.

Even a cursory visual inspection of a funnel graph diagram can give useful information. Looking at the diagram presented in Fig. 1, it is apparent that there is greater variation in estimates of effects based on small accident samples (as measured by statistical weight) than in estimates based on larger accident samples. The estimates based on large accident samples (statistical weight of >400) all lie between the values of 0.7 (i.e. 30% reduction in the probability of injury) and 1.0 (i.e. no change in the probability of injury). A weighted mean estimate of the effect of headrests on injury probability would therefore seem to make sense.

To evaluate the criticisms presented above, however, formal analyses are needed. More specifically the following techniques can be applied;

(1) weighting results according to a fixed effects model or a random effects model, depending on the outcome of a statistical test of the homogeneity of results;
(2) assessing skewness in a sample of results by comparing the weighted mean, the median and the modal value of the results;
(3) assessing the modality of a sample of results by means of identifying the general shape of the distribution of results based on a funnel graph plot;
(4) assessing the sensitivity of results to outlier bias by means of identifying the contribution of outlying data points to the estimated weighted mean;
(5) assessing publication bias by examining the contribution to a funnel graph plot of significant and non-significant data points in different directions;
(6) assessing the uncertainty of the weighted mean by means of the jack-knife technique for eliminating multiple results of the same study (Mosteller and Tukey, 1968).

TESTING THE HOMOGENEITY OF RESULTS

As noted in Section 1, one of the most common objections to meta-analysis is that it produces nonsensical mean effects based on highly heterogeneous samples of results. An easy way of evaluating the relevance of this argument, is to produce a graph of results, derived from a funnel graph, in which results are listed chronologically on the ordinate, and the 95% confidence interval of each result is displayed. Figure 2 shows such a diagram for the 30 results that are represented in Fig. 1.

In Fig. 2, one may assess the amount of systematic variation in study results by looking at the degree of overlap between the confidence intervals. It is readily seen that there are systematic differences. For example, the confidence intervals of results number 11 and 13 do not overlap. Results number 13 and 22, both of which have very small confidence intervals, also differ significantly from each other.

A formal test of the homogeneity of results is presented in Appendix A. The use of this test is illustrated in Table 1. Table 1 shows a χ^2 analysis of the homogeneity of the results with respect to effects of headrests on the probability of injury in rearend collisions.

Table 1 shows that there is a significant amount of heterogeneity in the 30 results. The overall effect of headrests is statistically significant at 5% level. The first partitioning of the results refers to the type of effect studied. Most results, 28 of a total of 30, refer

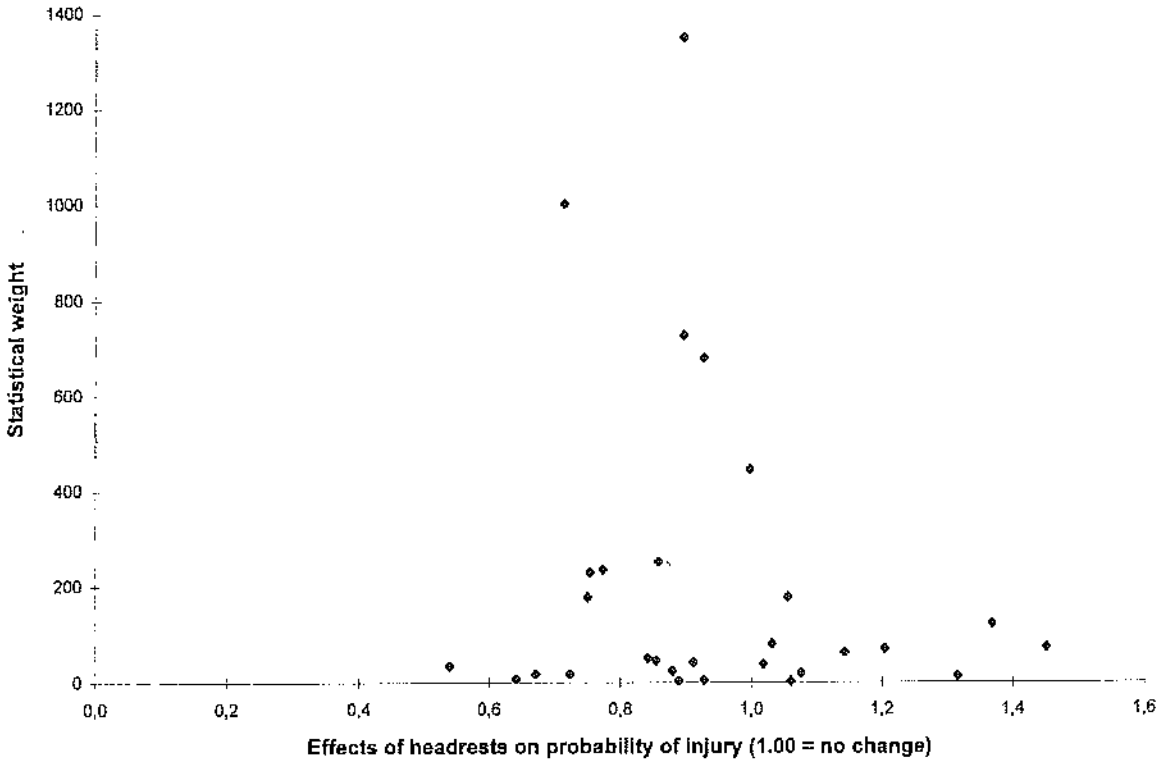


Fig. 1. Funnel graph diagram for results of studies that have evaluated the effects of headrests in cars on the probability of injury.

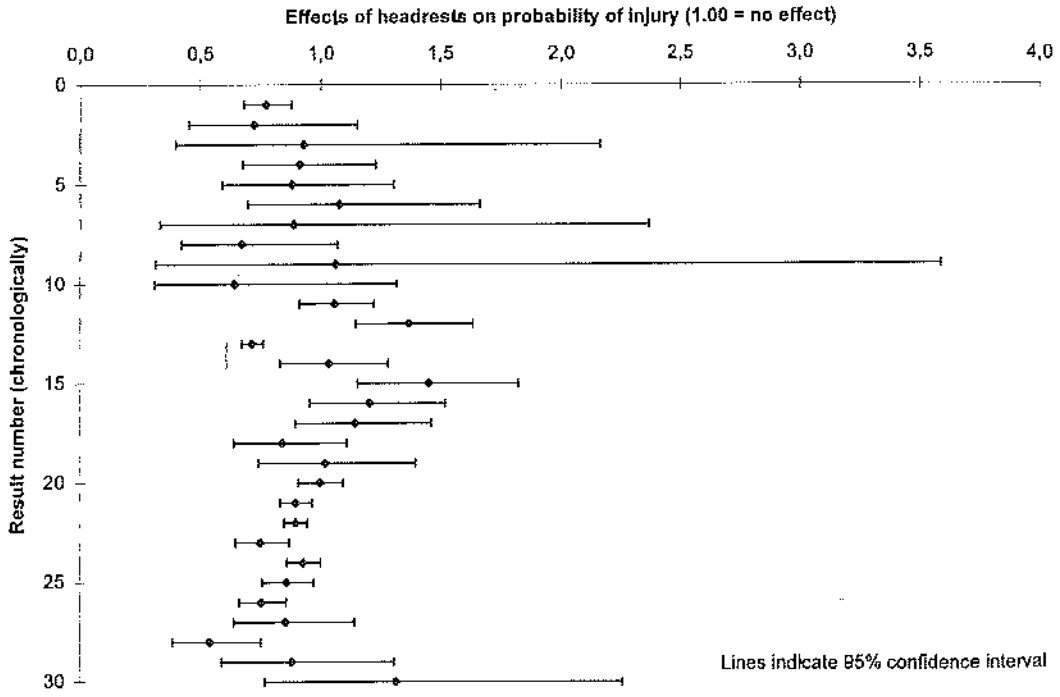


Fig. 2. Plot of results of studies that have evalnated the effects of headrests in cars on injury probability and the confidence interval of each data point.

to the effect of having a headrest in the car versus not having one. Two results compare adjustable and fixed headrests. Although these two results are significantly heterogeneous, further partitioning of the data is not feasible, since the χ^2 for homogeneity is not defined for a single study.

Table 1. χ^2 test of homogeneity of results of studies that have evaluated the effects of headrests on the probability of injury

Sample stratification	Values of stratifying variable	Number of results	χ^2 for homogeneity			χ^2 for association		
			Value	Degrees of freedom	5% significance	Value	Degrees of freedom	5% significance
All results	All values	30	143.364	29	S	104.004	1	S
Type of comparison	Headrest versus no headrest	28	136.919	27	S	91.887	1	S
	Adjustable versus fixed headrest	2	6.309	1	S	12.253	1	S
Injury severity for	Fatal injuries	7	12.203	6	S	2.125	1	NS
headrest versus none	Non-fatal injuries not specified	4	44.118	3	S	94.601	1	S
	Non-fatal neck injuries	17	46.426	16	S	28.333	1	S
Fatal injuries for	Before-and-after study—type 1	6	10.373	5	NS	4.950	1	S
headrest versus none	Before-and-after study—type 2	1	Not defined for a single study			0.004	1	NS
Non-fatal injuries for	Case control studies	2	23.020	1	S	90.676	1	S
headrest versus none	Before-and-after studies	2	0.000	1	NS	25.022	1	S
Neck injuries for	Case control studies	7	33.673	6	S	20.417	1	S
headrest versus none	Before-and-after studies	10	12.215	9	NS	8.454	1	S
Neck injuries for	Driver injuries	2	0.073	1	NS	17.357	1	S
case-control studies	Driver and passenger injuries	5	29.215	1	S	7.445	1	S
Driver and passenger neck	Adjustable headrests	2	0.001	1	NS	0.395	1	NS
injuries for case-control	Fixed headrests	1	Not defined for a single study				1	S
studies	Type of headrest not specified	2	17.938	1	S	0.001	1	S

By successively partitioning a set of results into smaller subsets as shown in Table 1, it is to some extent possible to avoid mixing results that differ significantly from each other. But Table 1 also shows the limitations of the χ^2 test of the homogeneity of results. In this comparatively small data set of 30 results, seven successive partitionings of the results were needed to account for all sources of systematic variation in results that were included in the study. The moderator variables were considered one at a time, not in conjunction, as one would in a multivariate analysis.

A set of results will fail the homogeneity test if there is systematic variation between the results in the set. If one were to go strictly by the results of this test, presenting mean results from a number of studies would make sense only if the variation in results was purely random. This criterion of homogeneity is likely to be too stringent in accident research. The effects of a certain safety measure on accident occurrence or injury severity is no doubt influenced by a large number of factors. However, as argued by Hauer (1991), the belief that one has to account for all these factors before anything can be concluded with respect to the effects of a measure, quickly leads to paralysis. In the first place, one will never know all the moderating factors. In the second place, apparently systematic variation in results can be generated not just by truly moderating factors, but also by statistical artifacts like incomplete or inaccurate accident reporting, whose effects cannot be separated from those of the true moderator variables.

To the extent that there is a large amount of systematic variation in a set of results, it may be more appropriate to adopt a random effects model for estimating the weighted mean and its standard error. The basics of the random effets model are shown in Appendix A. Appendix B shows how applying the random effects model to the headrest data affects the statistical weights assigned to each result and the weighted mean. The random effects model leads to a considerable flattening of the statistical weights. Moreover, it shifts the value of the weighted mean towards an unweighted mean. The standard error of the mean becomes greater, since the value of the statistical weights is considerably reduced. In the headrest data, the sum of the statistical weights is 6029 in the fixed effects model and 765 in the random effects model.

TESTING SKEWNESS IN A SAMPLE OF RESULTS

If a set of results is highly skewed, a mean result can be misleading in the sense that a large majority of the results may lie to one side of the mean value. Testing for skewness in a set of results from evaluation studies is easy in a funnel graph diagram. Figure 3 shows an example of such a test for the 30 results plotted in Fig. 1, using fixed effects statistical weights.

In a distribution skewed to the left (with a longer tail to the right than to the left), the mode will be smaller than the median, which in turn will be smaller than the mean. In a distribution skewed to the right, the converse applies. In Fig. 2, the weighted mean is 0.877 (i.e. a 12.3% reduction in the probability of

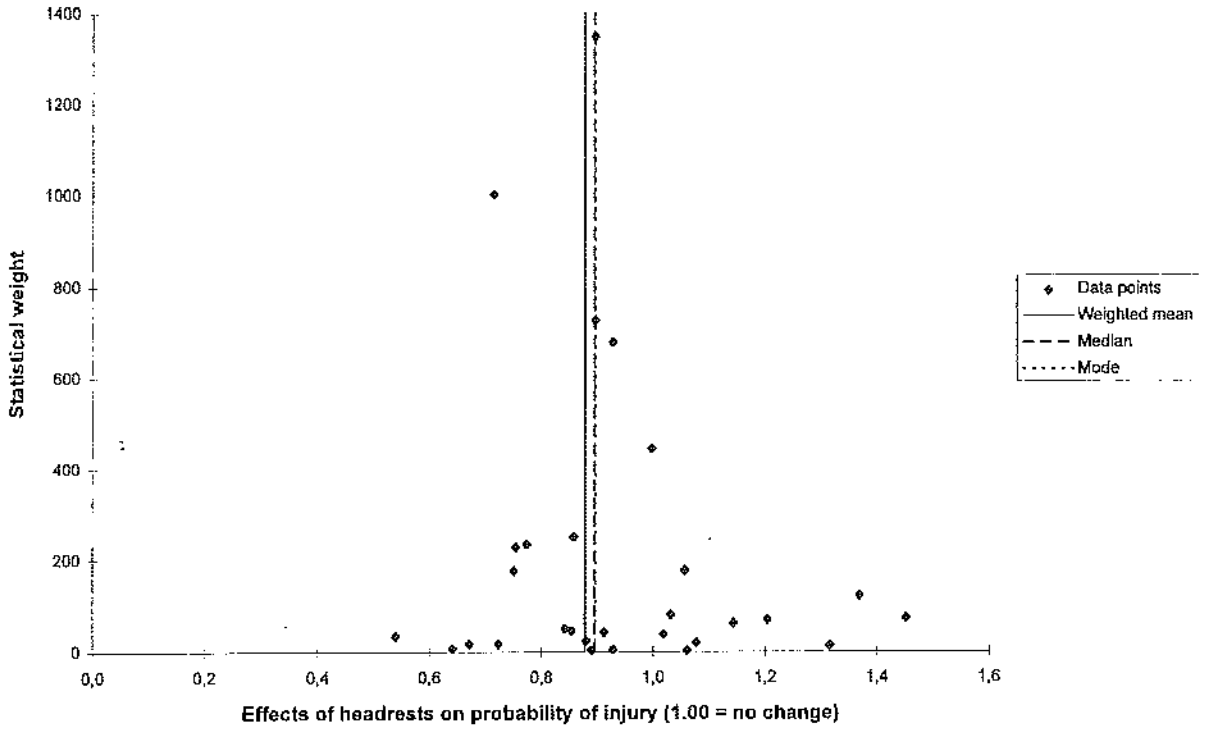


Fig. 3. Weighted mean, median and mode of results of studies that have evaluated the effects of headrests in cars on the probability of injury.

injury; shown be the solid vertical line), the weighted median (the result that divides the sum of statistical weights for all results into two equal parts; shown by the spaced line) is 0.894 and the mode (the result with the largest statistical weight; shown by the dotted line) is 0.896. This indicates a very slight skewness to the right, but the mode, median and mean do not differ significantly from each other. In this case, the mean is a good summary of the central tendency in the sample of results.

TESTING THE MODALITY OF A DISTRIBUTION OF RESULTS

The modality of a distribution denotes the number of humps in it. A distribution with just one hump is unimodal, a distribution with two humps is bimodal, etc. In multimodal distributions, it is usually more informative to estimate a mean for each mode, not just an overall mean that pastes over all modes. Based on a funnel graph diagram, it is easy to determine the modality of the distribution of results.

Fignre 4 shows a frequency distribution of results, based on their statistical weights (fixed effects model), compiled from the funnel graph plot of Fig. 1. The concentration of results around the weighted mean value (0.877) is clearly visible. The distribution is unimodal and slightly skewed to the left (with a tail to the right).

A frequency distribution based on the number of results, rather than their statistical weights, is similar to the one presented in Fig. 4.

TESTING THE SENSITIVITY OF THE MEAN TO OUTLYING DATA POINTS

Any sample of results contains extreme data points. The extreme data points will, however, not necessarily bias the mean. In the first place, if the extreme data points are symmetrically distributed between the two tails of the distribution, they will tend to cancel out each other. In the second place, extreme data points tend to be based on smaller statistical weights than less extreme data points and will therefore contribute less to the mean.

How, then, is the outlier bias argument given in the introduction to be interpreted? Bias can be introduced if a single data point significantly affects the mean. An outlying data point will therefore be defined as any data point that significantly affects the mean value of a set of results. In order to determine if a set of results contains outlying data points in this sense, the influence of each data point on the mean can be assessed by omitting it and estimating the mean based on the remaining $g-1$ data points. If the omission of a data point leads to a significant change in the estimate of the mean, the omitted data

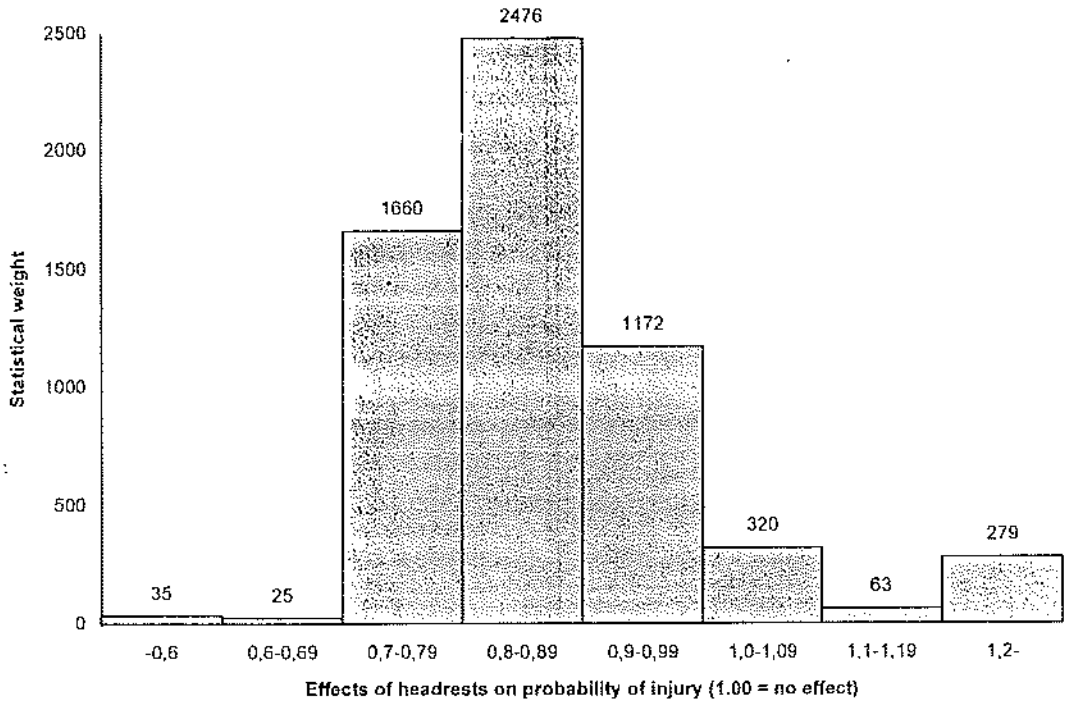


Fig. 4. Distribution of results of results of studies that have evaluated the effects of headrests in cars on the probability of injury.

point will be defined as outlying. Figure 5 shows the upper and lower 95% confidence limits for the meau of the 30 data points in Fig. 1, according to a fixed effects model, estimated by successively omitting one data point.

The boldface horizontal line shows the overall mean, based on all 30 data points. It is seen that the overall mean lies within the 95% confidence limits of all the estimates based on 29 data points, except one. The case in point is, however, a borderline case. The 95% confidence limits of the overall mean, which are not shown in Fig. 5, partly overlaps the 95% confidence limits of the estimates of the mean of $g-1$ data points. The general impression from Fig. 5 is that the mean is very stable and hardly affected by the omission of any single data point. This shows that there is no outlier bias in the mean of this distribution.

THE POSSIBLE PRESENCE OF PUBLICATION BIAS

Light and Pillemer (1984), who introduced funnel graph diagrams as an element of meta-analysis, argue that one can find indications of publication bias by carefully studying such diagrams. They distinguish between two forms of publication bias. One form of publication bias occurs when results that are not statistically significant are less likely to be published than results that are statistically significant. The other form of publication bias occurs when results that are regarded as 'unfavourable' or 'negative' (for example, results showing increases in the number of accidents or injuries) are less likely to be published than results that are regarded as desirable. If any of these forms of publication bias are present, they can affect the shape of the distribution of data points in a funnel graph. Bias against statistically insignificant results will tend to give the funnel graph a hollow core, with few data points. Bias against results in a certain direction will tend to cut off one of the tails of the funnel graph.

In the absence of direct evidence, it is of course impossible to know the extent to which the results shown in a funnel graph are affected by publication bias. One cannot know what a funnel graph would look like if it contained the results of every study that has even been made, published as well as unpublished. It is, however, possible to get a benchmark for an informal judgement by preparing a funnel graph in which a distinction is made between four categories of data points:

(1) data points showing a significant reduction;
(2) data points showing an insignificant reduction;
(3) data points showing an insignificant increase; and
(4) data points showing a significant increase.

Fig. 6 shows such a funnel graph for the 30 data points in Fig. 1.

By counting the number of data points in these four categories, it is possible to make an informal

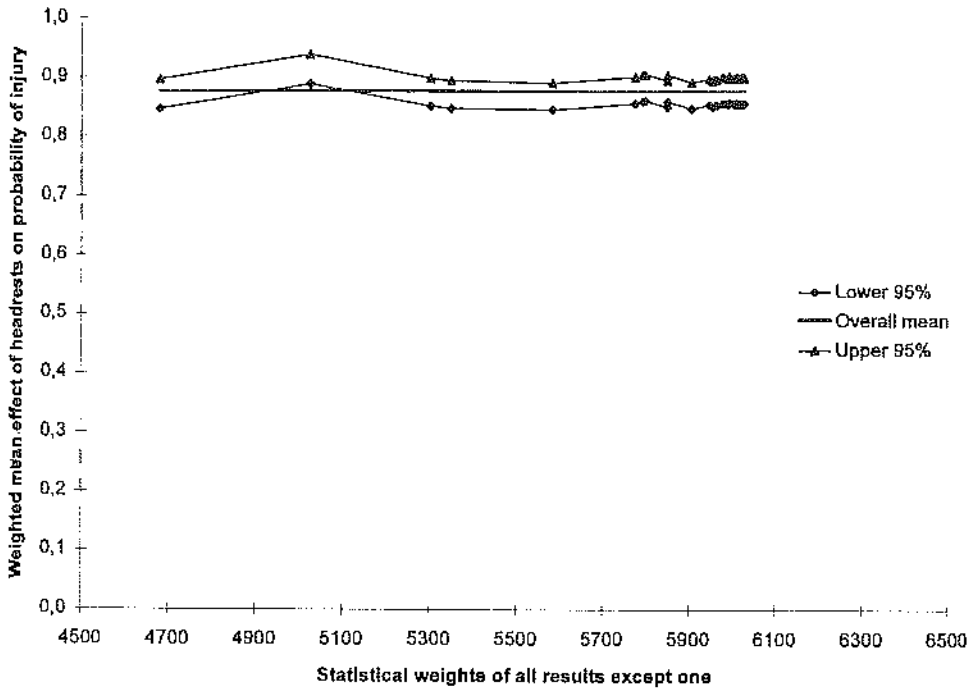


Fig. 5. The sensitivity of the weighted mean effect of headrests in cars on the probability of injury to outlying data points.

assessment. Figure 6 contains 19 data points that are not statistically significant. The fact that 19 out of 30 data points were not statistically significant indicates that there is no strong bias against insignificant results. By the same token, the fact that 10 out of 30 data points indicated an adverse safety effect of headrests hardly suggests the presence of a strong bias against unwanted results. Although indications like these fall short of hard evidence, they are perhaps sufficiently clear to point towards the conclusion that these data do not seem to be strongly affected by publication bias.

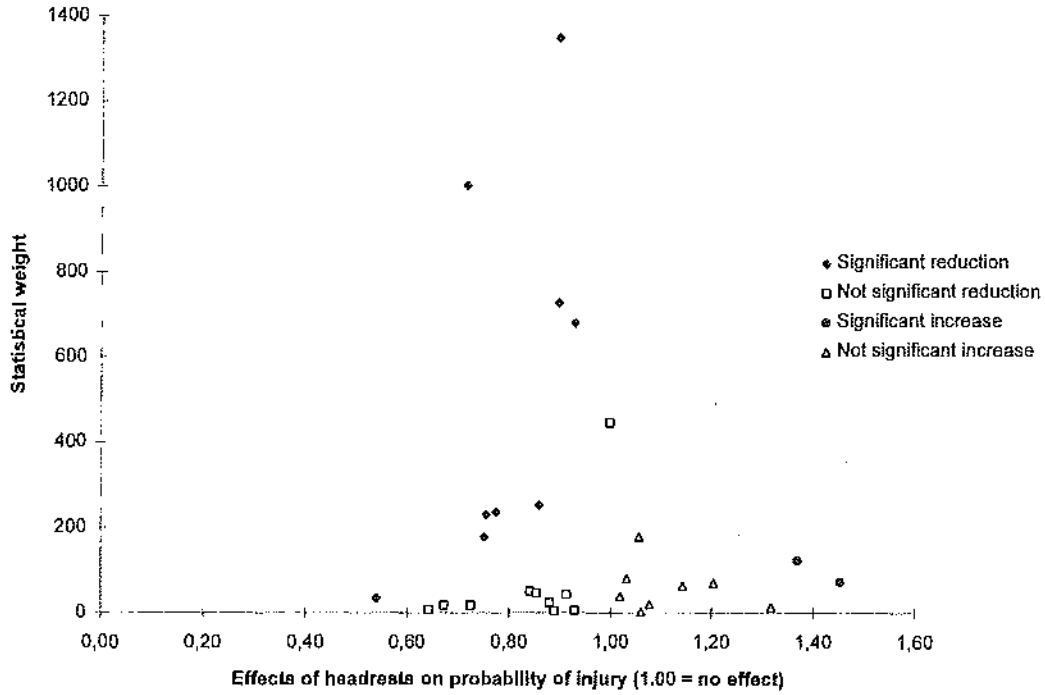


Fig. 6. Funnel graph diagram for headrests sorting data points according to statistical significance and direction of effect.

ASSESSING THE UNCERTAINTY OF A WEIGHTED MEAN RESULT BY MEANS OF THE JACK-KNIFE TECHNIQUE

Multiple results from the same study tend to be correlated. This means that g results originating from n studies, when $g > n$, are likely to vary less than g results originating from g studies will. Including multiple, correlated results from the same study in a meta-analysis may lead to an underestimate of the uncertainty of the mean result. Mosteller and Tukey (1968) have proposed using the so-called jack-knife technique to deal with this problem. Application of this technique to meta-analysis, is equivalent to basing the analysis on weighted mean results within each study, rather than on the original, multiple results of each study. Appendix B shows the application of the jack-knife technique to the headrest data set. Table 2 compares the weighted mean and its standard error for four models of analysis:

(1) fixed effects, nsing all results;
(2) fixed effects, applying jack-knifed results;
(3) random effects, using all results; and
(4) random effects, using jack-knifed results.

The choice of model of analysis is seen to affect both the weighted mean and its standard error. Using jack-knifed data does not affect the value of the weighted mean, but leads to a larger standard error. The standard error of the mean is more than six times greater for the random effects model with jack-knifed data than for for fixed effects model using all results.

Whereas the choice between a fixed effects and a random effects model of analysis can be based on the results of the homogeneity test, no similar test is available for determining whether multiple results from the same study are too highly correlated to be treated as independent in a meta-analysis. In fact, the meaning of the term 'correlated' is not perfectly clear when it comes to multiple results from the same study. Correlation usually refers to the relationship between variables X aud Y, not to the relationship between results 1 aud 2 from the same study. It is important to distinguish between cases in which multiple results from a study simply happen to coincide to a high degree, and cases in which there are features of the study design that create dependency between the results of the study. An example of the latter would be a study in which variables A, B, C and D are all derived from the same theoretical concept and intended to capture its empirical referent. Another example would be a study using multiple dependent variables, in which, for example, variable C is defined in terms of variables A and B. In this case, results applying to variable C are likely to be correlated with those applying to variables A and B. A third case would be a time-series analysis, in which multiple measurements of the same variable are auto-correlated. While the presence of correlation between multiple results of the same study can be tested in these three cases, it may seem overly conservative to opt for a jack-knife technique of analysis whenever a significant correlation is found.

DISCUSSION

Generalization is one of the basic characteristics of research. This characteristic is particularly promiuent in meta-analyses, which attempt to draw general conclusions based on formal analyses, often involving a large number of studies made over an extended period of time, very often in different countries and by means of different research methods. Can the results of studies that differ in these and a number of other ways be generalized at all? Critics of meta-analyses have pointed out a number of problems in generalizing, in the form of weighted or unweighted estimates of the mean result, the results of studies that differ in important respects.

A mean result is generally regarded as a meaningful summary of a set of results if: (a) the spread of results around the mean is 'well behaved': and if (b) the reported results are not known to be a biased sample of the type of studies that the results refer to. The extent to which condition (a) is met, can be determined by testing:

Table 2. Weighted mean effect of headrests on injury probability according to the fixed effects and random effects models of analysis

Model of analysis	Treatment of multiple results from same study	Effect of headrests on probability of injury (1.00 = no effect; < 1.00 = reduction)			Size of 95% confidence interval
		Lower 95%	Best estimate	Upper 95%	
Fixed effects	All results included	0.855	0.877	0.899	0.044
	Results jack-knifed	0.841	0.877	0.918	0.077
Random effects	All results included	0.853	0.915	0.983	0.130
	Results jack-knifed	0.794	0.915	1.074	0.280

(1) the skewness of the distribution to which the mean applies;
(2) the modality of the distribution to which the mean applies; and
(3) the sensitivity of the mean to outlying values, that is to results that are highly atypical of the distribution.

Simple methods for testing these characteristics of the set of resilts to which weighted mean results apply have been presented in this paper.

A more stringent interpretation of the condition of well behavedness requires that the distribution of results around the mean should contain random, sampling variation only. The extent to which this condition is met can be tested in meta-analysis by using the χ^2 test of homogeneity proposed by Fleiss (1981). It seems too stringent to say that a mean makes sense only if it applies to a sample containing random variation exclusively. Surely, it makes sense to say that the mean daily temperature in Oslo in January is lower than in June, although there will normally be systematic (i.e. larger than random) variations in daily temperatures in both months. Glass et al. (1981) are entirely correct in pointing out that the 'apples and oranges' argument made by critics of meta-analysis is inconsistent. How can critics know that there are, to stick to their metaphor, both apples and oranges in a sample of results? It is only as a result of a detailed and systematic review of the studies, that is only by means of a meta-analysis, that such a conclusion can be justified. The presence of systematic variation in a set of results does not necessarily make a weighted mean meaningless, but suggests that using a random effects model of analysis is more appropriate than using a fixed effects model of analysis.

The condition that the sample of reported results should not be known to be a biased sample of the studies that have been made (condition b) may appear too weak. It would of course be better to know that a sample of results was representative of all studies made, or better yet, that it contained all studies ever made. It is, however, rarely possible to know this. For many subjects, literally hundreds of studies have been made. No approach to searching the literature can guarantee the retrieval of every study, published as well as unpublished, that has ever been made about a subject. Meta-analysis will never be able to exclude the possibility of publication bias and must try to account for it as best it can. The inspection of funnel graph diagrams is one of several methods for assessing publication bias. It is not a formal test, but it does give an indication.

Finally, as far as the issue of correlation between multiple results of the same study is concerned, the concept would appear to make sense only when:
(1) a study uses multiple dependent variables, that are conceptually or empirically related; or
(2) multiple measurements have been made of the same dependent variable.

The concept of correlation between multiple results of the same study makes less sense when applied to a data set of the sort used for illustration in this paper. When testing the presence of correlation between multiple results of the same study, it is essential to take account for the fact that a certain amount of correlation can be expected to arise from chance alone.

CONCLUSIONS

This paper has presented a set of simple statistical and graphical techniques of analysis that can be used in meta-analysis to assess the strength of a number of criticisms that have been made against summarizing the results of a number of studies in terms of a weighted or unweighted mean result. More specifically, the following problems and associated techniques for dealing with them have been presented.

(1) The presence of significant heterogeneity in a sample of results can be tested by means of the χ^2 test for homogeneity, and a random effects model of analysis adopted if results are highly heterogeneous.
(2) The presence of skewness in a sample of results can be tested by comparing the mode, the median and the mean of the sample.
(3) The modality of a distribution, that is the number of humps in the distribution, can be assessed by compiling a frequency distribution of the results and examining its shape.
(4) The sensitivity of the mean to atypical results (outliers) can be tested by removing one result at a time, estimating the mean of the remaining $g-1$ results and comparing it to the mean of all g results.
(5) The possible presence of publication bias can be assessed by visual inspection of funnel graph diagrams in which data points have been sorted according to statistical significance and direction of effect.
(6) The uncertainty of a mean result based on g results from n studies, when $g>n$, can be assessed by means of the jack-knife technique.

Brief illustrations taken from road safety evaluation studies are given for all these techniques.

REFERENCES

Bangert-Drowns, R. L. (1986) Review of developments in meta-analytic method. *Psychological Bulletin* **99**, 388–399.

Cameron, M. H. and Wessels, J. P. (1979) The effectiveness of Australian design rule 22 for head restraints (Report CR 5). Melbourne, Road Safety and Traffic Authority, Victoria.

Cook, T. D. and Campbell, D. T. (1979) Quasi-experimentation. *Design and Analysis Issues for Field Settings*. RandMcNally, Chicago.

Elliott, B. (1993) Road safety mass media campaigns: a meta analysis (Report CR 118). Department of Transport and Communications, Federal Office of Road Safety, Canberra, Australia.

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

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

Fleiss, J. L. (1981) *Statistical methods for rates and proportions*, 2nd edn. Wiley, New York.

Fleiss, J. L. and Gross, A. J. (1991) 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* **2**, 127–139.

Glass, G. V., McGaw, B. and Smith, M. L. (1981) *Meta-analysis in Social Research*. Sage, Beverly Hills, CA.

Hauer, E. (1991) Should stop yield? Matters of method in safety research. *ITE-Journal* **61**, 25–31.

Kahane, C. J. (1982) An evaluation of head restraints. Federal motor vehicle safety standard 202 (Report DOT HS 806 108). US Department of Transportation, National Highway Traffic Safety Administration, Washington, DC.

Light, R. J. and Pillemer, D. B. (1984) *Summing Up. The Science of Reviewing Research*. Harvard University Press, Cambridge, MA.

McLean, A. J. (1974) *Collection and Analysis of Collision Data for Determining the Effectiveness of Some Vehicle Systems*. Motor Vehicle Manufacturers Association, Detroit, MI.

Mosteller, F. and Tukey, J. W. (1968) Data analysis, including statistics. In *The Handbook of Social Psychology*, eds G. Lindzey and E. Aronson, 2nd edn, Vol. 2, Research Methods, pp. 80–203. Addison-Wesley, Reading, MA.

Nygren, Å. (1984) Injuries to car occupants—some aspects of the interior safety of cars. *Acta Oto-Laryngologica Scandinavica* Suppl., 395.

Nygren, Å., Gustafsson, H. and Tingvall, C. (1985) effects of different types of headrest in rear-end collisions. *Proceedings of Tenth Experimental Safety Vehicle Conference*, pp. 85–90. National Highway Traffic Safety Administration, Washington, USA.

O'Neill, B., Haddon, W., Kelley, A. B. and Sorenson, W. W. (1972) Automobile head restraints-frequency of neck injury claims in relation to the presence of head restraints. *American Journal of Public Health* **62**, 399–406.

Shadish, W. R. and Haddock, C. K. (1994) Combining estimates of effect size. In *The Handbook of Research Synthesis*, eds H. Cooper and L. V. Hedges, pp. 261–281. Russell Sage Foundation, New York.

States, J. D., Balcerak, J. C., Williams, J. S., Morris, A. T., Babcock, W., Palvino, R., Riger, R. and Dawley, R. E. (1972) Injury frequency and head restraint effectiveness in rear-end impact accidents. *Proceedings of Sixteenth Stapp Car Crash Conference*, pp. 228–245. Society of Automotive Engineers, New York, USA.

Wagenaar, A. C., Zobeck, T. S., Williams, G. D. and Hingson, R. (1995) 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.

APPENDIX A

Mathematical appendix

Notation

y_i	result i expressed in terms of the natural logarithm of the odds ratio
w_i	the statistical weight of result i, estimated by $1/v_i$
v_i	the variance of result i
g	the number of results
n	the number of studies

Weighted mean effect in the fixed effects model of meta-analysis The weighted mean effect based on g results is estimated according to:

$$\bar{y}=\exp\left(\frac{\sum_{i=1}^{g} w_i y_i}{\sum_{i=1}^{g} w_i}\right) \tag{1}$$

where exp denotes the exponential function. The variance, v_i of each result is estimated by:

$$v_i=\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D} \tag{2}$$

where A, B, C and D are the four numbers that enter the calculation of the odds ratio, and the statistical weight, w_i, of each result is $1/v_i$.

Statistical test of homogeneity of results The extent to which the various y_i results differ from zero (i.e. no association, or no effect) can be assessed by estimating:

$$\chi^2_{\text{total}}=\sum_{i=1}^{g} w_i y_i^2 \tag{3}$$

This test statistic has a χ^2 distribution with g degrees of freedom. To determine the presence of systematic variation in the y_i effects, the χ^2 can be partitioned into two components:

$$\chi^2_{\text{total}}=\chi^2_{\text{hom og}}+\chi^2_{\text{assoc}} \tag{4}$$

The quantity $\chi^2_{\text{hom og}}$ essentially measures the presence of systematic variation between the g measures of effect, and the quantity χ^2_{assoc} assesses the significance of the weighted mean effect. The two components of the χ^2 are estimated as follows:

$$\chi^2_{\text{hom og}}=\sum_{i=1}^{g} w_i\left(y_i-\bar{y}\right)^2 \tag{5}$$

which is χ^2 distributed with $g-1$ degrees of freedom (df), and:

$$\chi^2_{\text{assoc}}=\frac{\left(\sum_{i=1}^{g} w_i y_i\right)^2}{\sum_{i=1}^{g} w_i} \tag{6}$$

which is χ^2 distributed with 1 df.

Statistical weights in the random effects model of meta-analysis If the test of homogeneity shows that there is a large systematic variation in study results, one can account for this when

combining results by adding a systematic variance component to the variance of each result used in determining the statistical weight of that result:

$$v_i^* = \sigma_\theta^2 + v_i \tag{7}$$

in which σ_θ^2 is the systematic variance component. The estimator of the systematic variance component [Shadish and Haddock (1994), Equations (18)–(23)] is:

$$\sigma_\theta^2 = \left[\chi^2_{\text{hom } vg} - (g-1) \right] / c \tag{8}$$

in which c is estimated according to:

$$c = \sum_{i=1}^{g} w_i - \left[\sum_{i=1}^{g} w_i^2 / \sum_{i=1}^{g} w_i \right] \tag{9}$$

The statistical weight of each result in the random effects model is $1/v_i^*$. Application of a random effects model leads to a considerable flattening of the statistical weights, see Appendix B. The weighted mean in a random effects model may be closer to an unweighted mean of the results than to the weighted mean of a fixed effects model.

The jack-knife technique of combining multiple results from the same study Assume that n studies have yielded g results, and that $g > n$, that is, at least one study has multiple results. To the extent that multiple results from the same study are stochastically dependent on each other (correlated), the variance of the g results may be smaller than the variance of n results. An artifactual shrinkage of the variance, due to multiple, correlated results from the same study, can be eliminated by means of a jack-knifed estimate of the mean and a corresponding jack-knifed estimate of the standard error of the mean. A jack-knifed estimate of the mean is obtained by taking the mean of pseudovalues. The pseudovalues are defined by:

$$y_{*j} = n\bar{y}_{\text{all}} - (n-1)y_j \tag{10}$$

where n is the number of studies, $\bar{y}_{\text{all}}$ is the overall mean and y_{*j} is the pseudovalue for each study. A pseudovalue is essentially the difference between the overall mean and the mean of all studies minus one. When applying the jack-knife technique to meta-analyses relying on the logodds method, the pseudovalues are identical to the weighted mean result for each study.

The lower and upper 95% confidence interval for the standard error of the jack-knifed estimate of the mean, can be estimated by Mosteller and Tukey (1968) (p. 135):

$$95\%_{\text{lower}} = \sum_{i=1}^{n} \left(w_i / \sum_{i=1}^{n} w_i \right) \left\{ \exp\left[\left(w_i y_i \right) / w_i \right] - 1.96/\sqrt{w_i} \right\} \tag{11}$$

and the corresponding upper 95% confidence interval is estimated by adding 1.96. Otherwise the formula is identical.

APPENDIX B — *see next page*

APPENDIX B

Calculation of pseudovalues for assessing the uncertainty of the weighted mean of 30 estimates of the effects of headrests in cars by means of the jack-knife technique

Study No.	Result No.	Result	Fixed effects weight	Random effects weight	Fixed effects mean of $n-1$ studies	Random effects mean of $n-1$ studies	Fixed effects pseudovalues	Random effects pseudovalues	Fixed effects weight of pseudovalues	Random effects weight of pseudovalues
1	1	0.773	234.826	38.822	0.881	0.924	0.773	0.773	234.826	38.822
2	2	0.723	17.782	12.864						
	3	0.928	5.365	4.810	0.877	0.919	0.766	0.774	23.148	17.674
3	4	0.912	43.007	22.345	0.877	0.915	0.912	0.912	43.007	22.345
4	5	0.879	24.576	16.080						
	6	1.077	20.281	14.123						
	7	0.889	4.000	3.683						
	8	0.671	17.845	12.897						
	9	1.061	2.590	2.453						
	10	0.642	7.386	6.374	0.877	0.921	0.851	0.846	76.678	55.610
5	11	1.056	177.740	36.865						
	12	1.368	122.062	33.678						
	13	0.714	1001.150	44.447						
	14	1.031	80.775	29.516						
	15	1.451	73.444	28.477						
	16	1.204	70.181	27.973						
	17	1.143	62.820	26.725						
	18	0.841	50.802	24.281						
	19	1.018	38.318	21.010						
	20	0.997	444.693	42.107						
	21	0.896	726.547	43.713						
	22	0.896	1348.348	44.961						
	23	0.750	177.206	36.842						
	24	0.927	678.572	43.528	0.803	0.814	0.892	0.980	5052.660	484.123
6	25	0.857	251.664	39.256						
	26	0.754	229.205	38.665	0.883	0.929	0.806	0.804	480.869	77.921
7	27	0.854	46.110	23.155						
	28	0.539	34.681	19.867						
	29	0.879	24.358	15.986						
	30	1.316	13.209	10.288	0.879	0.927	0.788	0.803	118.358	69.296
Sum/mean		0.935	6029.545	765.791	0.877	0.915	0.877	0.915	6029.545	765.791

Paper 7

Pergamon

Accid. Anal. and Prev., Vol. 30, No. 1, pp. 101–118, 1998

0001-4575/98 $19.00 + 0.00

PII: S0001-4575(97)00068-7

ARE ROAD SAFETY EVALUATION STUDIES PUBLISHED IN PEER REVIEWED JOURNALS MORE VALID THAN SIMILAR STUDIES NOT PUBLISHED IN PEER REVIEWED JOURNALS?

RUNE ELVIK*

Institute of Transport Economics, PO Box 6110 Etterstad, N-0602 Oslo, Norway

(Received 26 March 1997)

Abstract—The peer review system of scientific journals is commonly assumed to prevent seriously flawed research from getting published. This paper compares the quality of 44 road safety evaluation studies published in peer reviewed journals to the quality of 79 evalutation studies dealing with the same safety measures, but not published in peer reviewed journals, in terms of seven criteria of study validity. Studies were scored for validity in terms of (1) sampling technique, (2) total sample size, (3) mean sample size for each result, (4) specification of accident or injury severity, (5) study design, (6) number of confounding factors controlled and (7) number of moderator variables specified. Confounding factors are all factors that distrurb the attribution of a causal relationship between the safety measure being evalutated and the observed changes in safety, moderator variables are all variables that influence the size of the effect of the safety measure. Very few statistically reliable differences in study validity were found between studies published in peer reviewed journals and studies not published in such journals. There was, at best, a weak tendency for studies published in peer reviewed journals to score higher for validity. An interaction was found between author affiliation and type of publication with respect to study validity. Studies published in peer reviewed journals by authors who were at a university scored highest for validity. For a number of reasons, this study must be regarded as exploratory and its results as indicative only. The study does, however, point to a line of research that might be worth pursuing in larger and more rigorous studies.

Keywords—Road safety, Evaluation study, Study validity, Peer review, Scientific journal

INTRODUCTION

Can the findings of studies that have evaluated the effects of, for example, a traffic safety measure, be trusted? Some of the most prominent researchers in road safety have argued that many studies are fatally flawed and ought to be rejected as nonscientific. In reviewing studies that have evaluated the safety effects of road design, it was noted by Ezra Hauer (1988, p. 3): "As I moved from one inquiry to another and realized how shallow are the foundations for what passes for knowledge, it gradually dawned on me that ignorance about the safety repercussions of common elements of highway design or traffic management is not the exception." A few pages later (Hauer, 1988, p. 12), he added that "the situation is reminiscent of the Middle Ages. What was accepted as true was the knowledge passed on from the ancients and made believable by uncritical repetition."

Remarks to the same effect have been made by Leonard Evans (1991, p. 379): "Increasing the importance of peer-reviewed literature is the most effective way to discard the plethora of nonscientific results which overwhelm this field. The value of many papers is very negative; not only do they spread misinformation, but they may oblige competent researchers to squander their time refuting nonsense." Adding that "the peer-review system is subject to all the frailties to which humans succumb", Evans nevertheless concludes (Evans, 1991, p. 380) that "the average quality, importance, objectivity, and the technical correctness of the peer-reviewed literature is substantially higher than the nonrefereed literature."

Cynthia Crossen (1994, p. 178) is more skeptical. She notes that "the peer review system is stretched thin. The sheer volume of biomedical journals—some 15,000 journals publish about 250,000 articles a month—puts insupportable demands on the system.

*Author for correspondence: Tel: 00 47 2257 3800; fax: 00 47 2257 0290; e-mail: rune.elvik@toi.no

Peer reviewers are unpaid volunteers, and they cannot take the time to scrutinize the raw data, let alone replicate the research." She adds that a paper that is rejected in one journal, can often be published in another journal, concluding that the main function of peer review may be to decide not whether a paper is published, but where it is published.

Who is right? Leonard Evans or Cynthia Crossen? This paper explores this question by comparing the quality of studies that have evaluated the safety effects of five different traffic safety measures, depending on whether the studies were published in peer reviewed journals or not. The following research problems are discussed:

(1) What is meant by the term study quality? Is it possible to measure study quality objectively?

(2) What is the role of peer review in scientific publishing? Which other factors can affect study quality?

(3) Are road safety evaluation studies published in peer reviewed journals of higher quality (more valid) than similar studies not published in peer reviewed journals?

The third question is the main question to be dealt with. A brief discussion of the other two questions is, however, needed in order to answer the main question.

MEASURING STUDY QUALITY

The notion that it is possible to assess the quality of scientific studies is basic to the peer review system. If it was impossible to tell a good study from a bad one, the whole idea of subjecting studies to peer review would stop making sense. Nevertheless, universally accepted and easily applied standards of study quality do not exist. Rosenthal (1991, p. 130) points out that "bad studies are too often those whose results we do not like." This point of view is too pessimistic and overly cynical. Fairly elaborate frameworks for measuring study quality have been developed, notably by Campbell and Stanley (1966), and Cook and Campbell (1979) for evaluation research, Mitchell and Carson (1989) for contingent valuation studies and Chalmers et al. (1981) for clinical trials.

In this paper, the validity framework of Cook and Campbell (1979) has been used to assess study quality. In this framework, a distinction is made between four types of validity and 33 specific threats to validity. The four types of validity are (1) statistical conclusion validity, which refers to the numerical accuracy and representativeness of the results of a study, (2) construct (or theoretical) validity, which refers to the success in making theoretical concepts operational, (3) internal validity, which refers to the possibility of inferring causality in the relationship between variables, and (4) external validity, which refers to the possibility of generalizing the results of a study to other contexts than the one in which the study was made. Studies are rated in terms of how well they account for the various threats to validity. All the threats to validity identified by Cook and Campbell will not always be relevant. Hence, as pointed out by Wortman (1994), one must be selective in using the validity framework. In this paper, studies are rated in terms of the following seven validity characteristics:

(1) Sampling technique: How were study units selected?
(2) Total sample size: What was the total number of accidents or injuries in the study?
(3) Mean sample size: What was the mean number of accidents or injuries which specific results in each study were based on?
(4) Specification of accident or injury severity: Was the severity of accidents or injuries specified?
(5) Study design: What type of design was used?
(6) Confounding variables controlled: Which, from a list of three specific confounding variables, did the study control for?
(7) Moderator variables specified: Which, from a list of two specific moderator variables, did the study specify?

These criteria were chosen because they were regarded as important for the quality of road safety evaluation studies and because they were comparatively easy to apply. The seven criteria of study quality are obviously not exhaustive and reflect the author's opinion with respect to the importance of the various threats to validity identified by Cook and Campbell. Criteria 1, 2 and 3 refer to statistical conclusion validity, criterion 4 partly to statistical conclusion validity, partly to external validity and criteria 5, 6 and 7 mainly to internal validity. Table 1 elaborates the criteria of study validity. For reasons of space, a detailed discussion of the criteria is impossible in this paper. The criteria will only be briefly discussed (see Table 1).

Sampling technique

The best sampling technique is random sampling or studying the whole population to which one wishes to generalize. Many evaluation studies, however, rely on various forms of systematic sampling or convenience samples. Self-selected samples (people volunteering for a safety programme) are also quite common. Random sampling and population surveys

Table 1. Coding of studies by criteria of validity

Variable	Categories of variable	Numerical score
Sampling technique	Population survey (study of entire population)	3
	Random sample (from known sampling frame)	3
	Systematic sample (chosen according to specific criteria)	2
	Convenience sample (chosen arbitrarily or by reference to data availability)	1
	Self selected sample (volunteers for a study or treatment)	1
	Sampling technique not stated	1
Total sample size	Sum of statistical weights of all results reported in study, estimated according to logodds method of meta-analysis	Measured directly
Mean sample size	Mean statistical weight of each reported result in study, estimated according to logodds method of meta-analysis	Measured directly
Accident of injury severity specified	Accidents or injuries specified according to severity (eg fatal, injury, property damage only)	1
	Accident or injury severity not specified	0
Study design	Experimental design (controlled, randomized trial)	4
	Well controlled quasi-experimental or observational designs, for example Before-and-after study explicitly controlling for all major confounding factors Case-control study applying multivariate analysis to control for confounders Time-series analysis employing multivariate techniques or comparison series Multivariate analysis of cross-section data based on explicit statistical model	3
	Weakly controlled quasi-experimental or observational designs, for example Before-and-after study controlling for some, but not all major confounders Case-control study in which subjects are stratified according to confounders Time-series analysis with no comparison series and no multivariate analysis Multivariate analysis of cross-section data not based on an explicit model	2
	Inadequate study designs, for example Simple before-and-after studies (no comparison group) Simple case-control studies (no confounders accounted for) Simple case studies Theoretical estimates of effect based on assumptions that cannot be tested	1
Specific confounders controlled	All confounders from a list of three major confounders controlled (see Table 2)	3
	Two of three confounders from a list of three major confounders controlled	2
	One of three confounders from a list of three major confounders controlled	1
	No confounders controlled from a list of three major confounders	0
Specification of moderator variables	Both moderator variables from a list of two specified	2
	One moderator variable from a list of two specified	1
	No moderator variable from a list of two specified	0

have been assigned a score of three, systematic sampling a score of two and the other sampling techniques, including unknown, a score of one.

Total sample size

The total sample size for all results reported in a study has been defined in terms of the statistical

weight assigned to the sample in the logodds method of combining results from fourfold tables (Fleiss, 1981). A fairly common design is a before-and-after study with a comparison group. The results of such a study are based on the number of accidents or injuries in the treated group before (denoted B) and after (denoted A) treatment, and on the corresponding numbers in the comparison group (denoted D and C, respectively). The statistical weight of the logodds of the estimate of treatment effect $[\ln\{(A/B)/(C/D)\}]$ is:

$$\text{Statistical weight} = 1/(1/A + 1/B + 1/C + 1/D)$$

If the study lacks a comparison group, the terms $1/C$ and $1/D$ drop out. In an experimentally designed study using after data only, the terms $1/B$ and $1/D$ drop out.

Mean sample size

Many studies present several results referring to subsets of the study and one overall result. The mean sample size for each result presented in a study was defined in the same way as the total sample size.

Accident or injury severity specified

Studies that at least made a distinction between fatal accidents, injury accidents and property-damage-only accidents were rated as better than studies mixing these levels of accident severity. By the same token, studies that at least made a distinction between fatal injuries, severe injuries, slight injuries and no injury, were rated as better than studies not making it clear which level of injury severity results applied to.

Study design

Studies were classified in four groups with respect to study design. Experimentally designed studies were rated as best and given the score of four. By an experiment is meant a controlled, randomized trial, in which study subjects are randomly assigned either to one or more treatment conditions or to a control condition not receiving any of the tested treatments. Well-controlled quasi-experimental or observational studies were rated as second best, with a score of three. To qualify for this group, a before-and-after study, for example, must use a comparison group and explicitly control for all major confounding factors, at least for all those listed in Table 2 (to be discussed below). Case-control studies must use multivariate analysis to control for all major confounders. The next step on the ladder is weakly controlled quasi-experimental studies or observational studies, given a score of two. There is admittedly, a grey zone here, in which one may be in doubt with respect to the classification of a study as well controlled or weakly controlled. The classification of study design was based on considerations of: (1) how the comparison group was chosen and its size, (2) how sophisticated the techniques of analysis used in the study were and (3) whether the study controlled for all confounding variables listed in Table 2 or not. The poorest types of study design were labelled inadequate and include, for example, simple before-and-after studies not controlling for any confounders or simple case-control studies. These studies were given the score of one.

Confounding variables controlled and moderating variables specified

Classifying studies in four groups by study design is a rather crude way of measuring their quality. A need was felt for a more explicit assessment of whether studies have controlled for known confounders or not. To this end, the classification of relevant variables shown in Fig. 1 and made operational in Table 2 was developed.

Figure 1 is most relevant for non-experimental studies, which make up the vast majority of road safety evaluation studies. There are five basic categories of variables that are relevant in such studies. The independent variable or variables is the safety measure whose effects researchers want to measure. The dependent variable or variables is the change in the expected number of accidents or injuries that is causally attributable to the safety measure. The causal chain from the safety measure to the number of accidents or injuries always includes one or more mediator variables that mediate the effect of the measure. In the case of speed limits, for examples, changes in driving speed would be the most important mediator variable. Confounding variables are all variables that affect both the use of the safety measure and the number of accidents or injuries. An example of a common confounding variable in before-and-after studies is the number of accidents that was recorded before the safety measure was introduced.

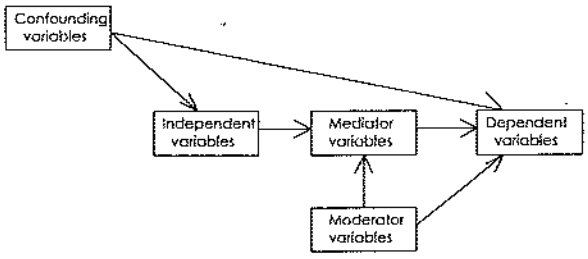


Fig. 1. Generic types of variables in non-experimental studies illustrated by a simplified causal diagram.

Table 2. Lists of confounders and moderator variables by study design for non-experimental studies

Measures lists apply to	Confounder or moderator	Before-and-after designs and time series analyses	Case-control designs and analyses of cross-section data
Roundabouts	Confounders	Regression-to-mean General trends Changes in traffic volume	Number of legs in junction Type of traffic control Speed limit
	Moderators	Number of legs in junction Type of traffic control	Number of legs in junction Type of traffic control
Blackspot treatment	Confounders	Regression-to-mean General trends Accident migration	These types of design have not been applied in studies of road accident blackspot treatment
	Moderators	Type of blackspot Type of treatment	
Daytime running lights (DRL) on cars	Confounders	Regression-to-mean General trends Other safety measnres	Self selection bias Driver characteristics Other safety measures
	Moderators	Type of accident affected Level of DRL use	Type of accident affected Level of DRL use
Seat belts	Confounders	These types of design have not been applied in studies of seat belts	Occupant age Seating position Impact speed
	Moderators		Type of belt Type of accident
Periodic motor vehicle inspection	Confonnders	Regression-to-mean General trends Driver characteristics	Self selection bias Type of traffic environment Driver characteristics
	Moderators	Vehicle technical condition Annual driving distance	Vehicle technical condition Annual driving distance

If this number was unusually high, a subsequent reduction (regression to the mean) must be expected, even if the measure is ineffective in reducing the number of accidents. Moderator variables are all variables that affect the size of the effect of the safety measure on accidents or injuries. Roundabouts, for example, reduce the theoretical maximum number of conflict points between the various turning movements in a junction from 32 to 9 in four leg junctions and from 9 to 6 in three leg junctions. If everything else is equal, the safety effect of roundabouts will be greater in four leg junctions than in three leg junctions. The number of legs in a junction is therefore a moderator variable for the effect of roundabouts.

In some study designs, especially case-control studies, the same set of variables can be both confounding variables and moderating variables. In before-and-after studies, on the contrary, the confounding and moderating variables are usually not identical. Table 2 lists the specific confounding and moderating variables that were defined for studies that have evaluated the safety effects of the five road safety measures selected for this study. The safety measures are roundabouts, blackspot treatment, daytime running lights on cars, seat belts (effects for each user) and periodic motor vehicle inspection.

The specific confounding or moderator variables listed are unique for each of the five measures, but some of the variables are common for more than one measure. In particular, regression to the mean, is listed for three of the five measures.

This paper does not permit a detailed description of how studies were classified in terms of the seven criteria described above. The appendix lists all studies that were included and shows the classification of the studies.

THE ROLE OF THE PEER REVIEW SYSTEM

The ideal function of the peer review system is to weed out bad research and thus prevent science from sinking into a morass of untested speculation and personal idiosyncrasies. In real life, the peer review system will, in the words of Leonard Evans, be "subject to all the frailties to which humans succumb". But exactly how bad or how good will the peer review system be in sorting out good studies from bad? Few studies have reported on this question, and a comprehensive survey of them will not be presented in this paper.

In an experiment made by The American Economic Review (Blank, 1991), double blind reviewing was compared to single blind reviewing. The experiment was made to test if there was any substance in allegations of sex discrimination made against the journal. The results indicated a slightly lower acceptance rate for papers subjected to double blind reviewing than for papers subjected to single blind reviewing. In about half the cases, however, reviewers were able to guess the identity of the author in double blind reviewing. The experiment vindicated the review system, in that it found no evidence of sex discrimination. Moreover, papers submitted by authors affiliated with high ranking universities were more often accepted than papers submitted by authors affiliated with lower ranking universities.

One of the most troublesome findings for the peer review system, is the demonstration of publication bias in scientific journals (Light and Pillemer, 1984; Begg and Berlin, 1988; Dickersin and Min, 1993). Publication bias denotes the tendency to reject papers for publication if their main findings are not statistically significant or are in an unexpected direction. Apparently, both editors and reviewers are less tolerant of papers that fail to confirm previous findings, or whose findings are not statistically significant, than of otherwise identical papers confirming previous findings at a conventional level of statistical significance.

Peer review is just one of several factors that may affect the quality of research. In a peer review of road safety research funded by the Swedish Transport Research Board, Elvik et al. (1993) suggested that research done at universities tends to be of higher quality than research done at institutes or consultancies doing contract research on a commercial basis. Arguments given for this hypothesis included: (1) the presence of a formal career structure rewarding high quality research at universities, colloquially referred to as the "publish or perish" system; (2) a more developed theoretical basis for research in well-defined academic subject areas than in multidisciplinary evaluation research; and (3) the absence of economic incentives to do "quick and dirty" research to win a competitive bid for a research contract. No doubt a host of other factors influencing research quality can be imagined.

STUDY RETRIEVAL AND ANALYSIS

Study retrieval

To compare the quality of road safety evaluation studies published in peer reviewed journals to the quality of studies not published in peer reviewed journals, studies that have evaluated five road safety measures were retrieved. Studies were retrieved by means of a systematic literature search that included examining selected journals, going through publications catalogues from selected research institutes, examining conference proceedings and examining the list of references in previously published meta-analyses. The journals that were scanned, included all journals listed in Table 3, as well as Australian Road Research, Ergonomics, Human Factors, Policy Sciences, Public Roads, Recherche—Transports—Sécurité (INRETS Research Review) and Risk Analysis. In general, all volumes after about 1970 were examined. The publications catalogues of the following institutions were studied and relevant studies identified: the Institute of Transport Economics (Norway), the Swedish Road and Transport Research Institute (VTI, Sweden), the Technical Research Centre of Finland (VTT, Finland), the Danish Road Safety Research Council (RfT, Denmark), the Nordic Council of Ministers and associated institutions, The Nordic Association for Traffic Engineering (Nordisk vegteknisk forbund), The SWOV Institute for Road Safety Research (Netherlands), the German Federal Road and Traffic Research Institute, (BASt, Germany), the Transport Research Laboratory (TRL, Great Britain), the French National Road Transport Research Institute (INRETS, France), the Australian Road Research Board (ARRB, Australia), the Transportation Research Board (TRB, United States) and the Organization of Economic Cooperation and Development (OECD, France).

Conference proceedings from the recent 10 years were examined for regular conferences, including the VTI Forskardager (VTI Annual Research Conference, in January each year in Linköping, Sweden), the annual European or Transatlantic conference hosted by VTI and others (named Road Safety in Europe or Road Safety on Two Continents every other year) and the PTRC Summer Annual Meeting (recently renamed the European Transport

Table 3. List of peer reviewed journals and classification of author affiliation

Variable	Categories of variable
Type of publication	Peer reviewed journal, alphabetically (only journals found in the data set listed): Accident Analysis and Prevention American Journal of Optometry American Journal of Public Health ITE-Journal (formerly Traffic Engineering) Journal of the American Medical Association Journal of Risk and Insurance Journal of Safety Research Journal of Traffic Medicine Journal of Transport Economics and Policy Traffic Engineering and Control Zeitschrift für Verkehrssicherheit
	Other types of publication (reports issued by research institutes etc)
Author affiliation	University, including hospital with teaching functions (2)
	Research institute or consulting group (1)
	Business firm (not a consulting firm) (0)
	Government agency or other affiliations (0)
Academic rank of most senior author	University professor (not including assistant or associate professors) (2)
	Other formal academic position at university, hospital or research institute (1)
	No formal academic position (0)

Forum). Studies were also retrieved from previous meta-analyses, including those of Elvik (1996, 1997). The approach to literature search adopted in this paper does not guarantee that every unpublished study is retrieved. In fact, almost all studies that were found have been published, in the sense that a report having the performing institution's name on the cover, and sometimes an ISBN number, has been printed and is available to the public. However, only a minority of the reports and papers that were found had been published in peer reviewed jonrnals.

Study inclusion criteria

Studies were included if: (1) The number of accidents or injuries was reported. Studies not reporting this information could not be included, because it was impossible to determine their sample size. (2) The final publication status could be determined. Studies that were published both in peer reviewed journals and in other forms, were counted only as peer reviewed publications.

Peer reviewed journals versus other publications

Table 3 lists the peer reviewed journals that were found in this data set. It is recognized that some of the studies that were not published in these journals may nevertheless have been subject to some form of peer review. Conference papers are sometimes peer reviewed, although the reviewing of conference papers is often regarded as more lax than that of scientific journals.

Author affiliation and rank

To the extent that information was available, author affiliation and rank was coded, as indicated in Table 3. A distinction was made between universities (a value of two), research institutes or consultancies (one) and other institutions, not having research as their primary function (zero). Professors were given the score of two, other formal academic positions the score of one and authors not having, or not stating, a formal academic position the score of zero. These codes were used merely as labels, and were not used as numerical variables in the analyses reported below.

Statistical analysis

Study quality was summarized by means of the arithmetic mean of scores for each of the seven

criteria of study validity that were used. Strictly speaking, the use of arithmetic means is inappropriate for most of these criteria, because they are measured on an ordinal scale only. An arithmetic mean requires variables that are measured on an interval scale of measurement. In the present study, however, the absolute values of the mean scores are not of primary interest. The main purpose of this study is to compare studies published in peer reviewed journals to studies not published in such journals. For this purpose, the mean was chosen as the most simple and understandable statistic.

The statistical significance of differences in mean scores was tested by means of the T-test. The normal approximation of T to Z (the number of standard deviations of the standard normal distribution) was used, although T is not identical to Z in small samples. The approximation is, however, close enough for the present study, considering the coarseness of the measurements that are compared.

RESULTS

Figure 2 presents the number of studies included by decade and type of publication. There were altogether 44 studies published in peer reviewed journals and 79 studies not published in peer reviewed journals.

The number of studies was 28 for roundabouts (five in journals, 23 other), 36 for blackspot treatment (14 in journals, 22 other), 17 for daytime running lights (five in journals, 12 other), 29 for seat belts (11 in journals, 18 other) and 13 for periodic motor vehicle inspection (nine in journals, four other).

Table 4 presents the results of the validity comparison by type of safety measure. The table contains altogether 35 (5×7) comparisons. Inspection of the p values in the right column of Table 4 shows that most of the differences found in study quality are far from statistical significance at conventional levels. The mean quality of studies published in peer reviewed journals was higher than for other studies in 20 cases. In 15 cases the mean quality of studies not published in peer reviewed journals was higher than for studies published in peer reviewed journals.

These results indicate, at best, a weak tendency for studies published in peer reviewed journals to be of higher quality (more valid) than other studies. The absence of statistically significant differences in study quality is no doubt due in large measure to the small number of studies compared. In order to compare the quality of all the 44 studies published in journals to all the 79 studies not published in journals, a ratio of mean validity scores was computed, that is the mean validity score (from Table 4) of studies published in peer reviewed journals divided by the mean validity score of studies not published in peer reviewed journals. A weighted mean of these ratios was then estimated using weights inversely proportional to the number of studies. For example, the

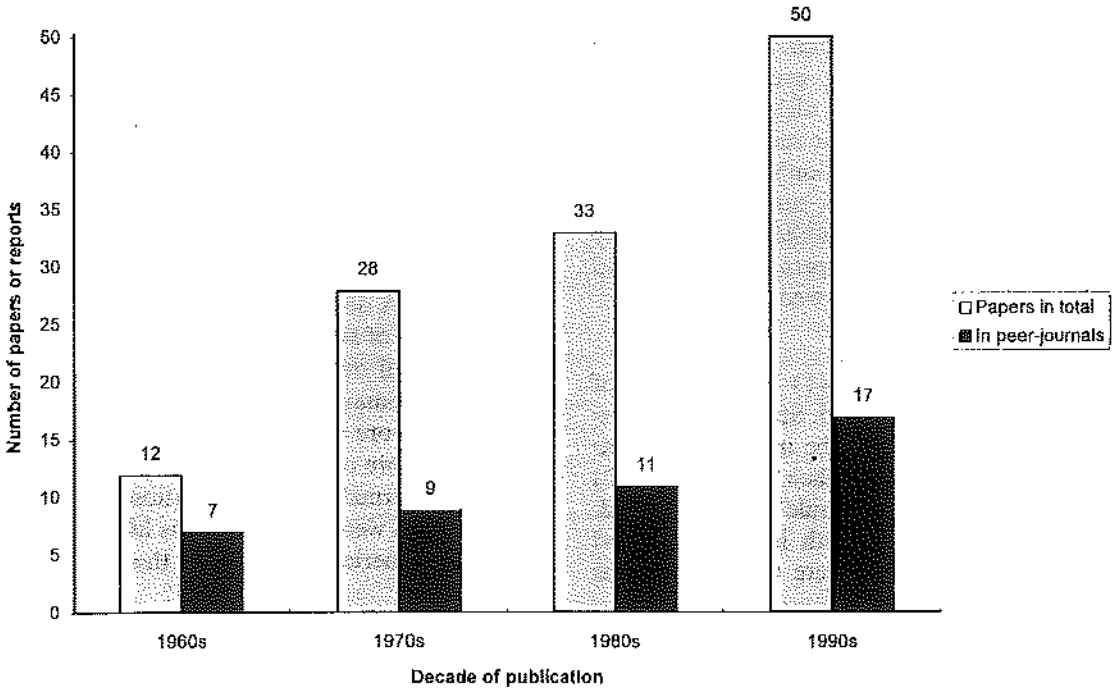


Fig. 2. Number of studies in total and in peer reviewed journals by decade of publication.

Table 4. Comparison of the validity of papers published peer reviewed journals and other types of publications

Measure	Criteria of study validity	Mean scores (standard error) by type of publication		*P*-value
		Peer reviewed journals	Other types of publication	
Roundabouts	Sampling technique	2.00 (0.00)	1.68 (0.10)	0.01
	Total sample size	473.34 (266.64)	231.58 (75.72)	0.47
	Mean sample size	228.94 (135.65)	61.52 (19.54)	0.28
	Accident severity specified	1.00 (0.00)	0.95 (0.04)	0.23
	Study design	1.60 (0.45)	2.26 (0.15)	0.27
	Confounders controlled	0.60 (0.45)	1.42 (0.20)	0.21
	Moderators specified	0.80 (0.42)	1.47 (0.14)	0.22
Blackspot treatment	Sampling technique	1.71 (0.13)	1.36 (0.11)	0.15
	Total sample size	353.33 (147.84)	142.82 (42.19)	0.27
	Mean sample size	111.21 (27.43)	42.37 (14.38)	0.10
	Accident severity specified	0.64 (0.14)	0.82 (0.09)	0.44
	Study design	2.29 (0.23)	1.91 (0.16)	0.34
	Confounders controlled	1.50 (0.30)	0.96 (0.18)	0.27
	Moderators specified	1.43 (0.26)	1.64 (0.16)	0.62
Daytime running lights on cars	Sampling technique	1.40 (0.45)	2.58 (0.20)	0.07
	Total sample size	3232.11 (3527.38)	3473.26 (916.47)	0.97
	Mean sample size	443.02 (431.15)	1322.99 (643.20)	0.43
	Accident severity specified	0.40 (0.27)	0.83 (0.12)	0.27
	Study design	2.20 (0.82)	1.83 (0.34)	0.75
	Confounders controlled	1.40 (0.76)	1.25 (0.43)	0.90
	Moderators specified	1.00 (0.35)	1.58 (0.16)	0.25
Seat belts	Sampling technique	2.09 (0.09)	2.33 (0.11)	0.24
	Total sample size	1767.37 (858.82)	300.89 (110.24)	0.14
	Mean sample size	130.20 (63.40)	31.58 (7.63)	0.17
	Injury severity specified	1.00 (0.00)	0.94 (0.06)	0.32
	Study design	2.64 (0.20)	2.28 (0.20)	0.37
	Confonnders coutrolled	2.00 (0.30)	1.44 (0.25)	0.31
	Moderators specified	1.36 (0.15)	1.28 (0.16)	0.78
Periodic motor vehicle inspection	Sampling technique	2.00 (0.18)	2.75 (0.29)	0.11
	Total sample size	14,260.97 (7230.15)	60,802.65 (31,290.51)	0.23
	Mean sample size	3187.89 (1367.70)	29,729.62 (28,338.53)	0.38
	Accident severity specified	0.89 (0.12)	1.00 (0.00)	0.35
	Study design	2.67 (0.25)	2.25 (0.55)	0.60
	Confounders coutrolled	1.67 (0.25)	1.25 (0.55)	0.60
	Moderators specified	0.33 (0.25)	0.00 (0.00)	0.18

weight of studies evaluating roundabouts was:

$$\text{Weight} = 1/(1/5 + 1/23) = 4.107$$

The weighted mean validity ratio was estimated by means of the logodds method (Fleiss, 1981). Upper and lower 95% confidence limits for the weighted mean validity ratio were estimated by relying on the fixed effects model of Fleiss (1981). Figure 3 presents the results.

Studies published in peer reviewed journals score higher for validity than studies published elsewhere in terms of total sample size, mean sample size, study design and number of confounders controlled, but slightly lower in terms of sampling technique, specification of accident or injury severity and specification of moderator variables. The differences are statistically significant for total sample size and mean sample size, but not for any of the other criteria of study validity. Indirectly, these results lend support to the publication bias hypothesis, since studies based on large samples are more likely to obtain statistically significant findings than studies based on small samples.

Table 5 compares study validity for studies published in peer reviewed journals and other studies, depending on author affiliation.

The most consistent differences in study quality are found for studies written by authors with a university affiliation. Studies by university affiliated authors published in peer reviewed journals score higher for validity than studies by university affiliated authors not published in such journals for five of seven criteria of study validity. The differences are statistically significant at the 5% level for two of the criteria. As far as studies written by authors with other affiliations are concerned, no consistent differences in study quality were found between studies

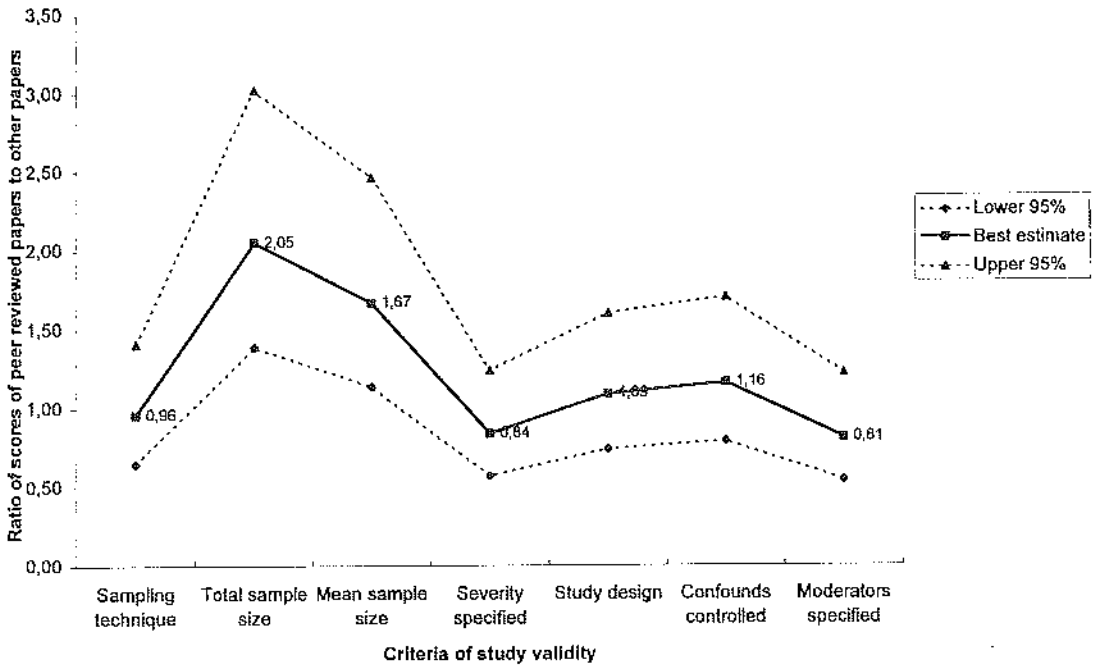


Fig. 3. Ratio of validity scores for studies published in peer reviewed journals to studies published in other types of publication.

published in peer reviewed journals and studies published elsewhere.

DISCUSSION

Are road safety evaluation studies published in peer reviewed journals of better quality (more valid) than studies published elsewhere? If taken at face value, the findings of this study indicate that the answer is something like: No, but to the extent that any differences can be found at all, it is more likely than not that they are of higher quality. This finding hardly lends much support to the calls made by Hauer, Evans and others to strengthen the place of the peer reviewed literature in road safety research in order to purge the field of non-scientific findings.

This study has a number of rather severe limitations whose combined effect is to make it exploratory only, rather than definite. In the first place, the validity scoring system does not include all study characteristics that are relevant in judging study quality. For example, the quality of accident or injury data, the choice of appropriate statistical techniques of analysis, the quality of operational definitions of theoretical concepts and the awareness of the authors of study limitations are aspects of study validity that were not scored in the present study. Including these, or other study characteristics, in the validity scoring system could in principle have changed the results. In practice, however, it is unlikely that the main findings would have been very different. Various aspects of study quality tend to be correlated. Hence, adding more items to the validity scoring system would not necessarily have added much more information to the system.

A more serious objection to the present study, is that the coding of studies according to validity was performed by one person only and that the publication status of each study was known to this person. Ideally speaking, two or more independent coders ought to have been used and the coders ought to be blinded from knowing the publication status of each study. However, the time and other resources available for the study made it impossible to use a rigorous coding procedure. For replicability, the coding of all studies is reproduced in the appendix to the paper.

A third limitation of this study is the small sample size. The sample is unlikely to be complete, especially with respect to studies not published in journals. It cannot be ruled out that omission of some unpublished studies may have biased the results. The plausibility of this argument depends, however, on the nature of the unpublished studies. Unpublished doctoral or masters dissertations can sometimes be quite rigorous and employ rather good research designs. Unpublished in-house reports by, for example, highway agencies, on the contrary, tend to be of quite poor quality, at least as evidenced in the sample of studies compiled for this paper.

A fourth limitation of the study is that the dichotomy used between peer reviewed journals and other publications is probably too simple to ade-

Table 5. Comparison of scores for study validity in peer reviewed journals and other types of publication by author affiliation

Author affiliation	Criteria of study validity	Mean scores (standard error) by type of publication		*P*-value
		Peer reviewed journals	Other types of publication	
University	Sampling technique	1.90 (0.11)	1.71 (0.17)	0.50
	Total sample size	6267.63 (3604.07)	6667.79 (6722.80)	0.97
	Mean sample size	1126.20 (624.11)	216.49 (170.40)	0.25
	Accident severity specified	0.84 (0.09)	0.82 (0.10)	0.92
	Study design	2.63 (0.14)	1.88 (0.20)	0.03
	Confounders controlled	1.79 (0.19)	0.94 (0.23)	0.04
	Moderators specified	0.58 (0.18)	1.24 (0.19)	0.07
Research institute or consultancy	Sampling technique	2.10 (0.19)	2.31 (0.12)	0.48
	Total sample size	4446.57 (2110.70)	5219.89 (3316.83)	0.89
	Mean sample size	1306.77 (779.38)	3758.70 (3272.61)	0.55
	Accident severity specified	0.90 (0.11)	0.97 (0.03)	0.62
	Study design	2.20 (0.38)	2.22 (0.15)	0.99
	Confounders controlled	1.50 (0.42)	1.50 (0.19)	1.00
	Moderators specified	1.20 (0.26)	1.47 (0.14)	0.51
Business firm or government agency	Sampling technique	1.67 (0.13)	1.67 (0.11)	1.00
	Total sample size	514.20 (410.44)	594.53 (299.89)	0.91
	Mean sample size	38.30 (10.98)	455.89 (300.02)	0.18
	Accident severity specified	0.67 (0.13)	0.87 (0.06)	0.30
	Study design	2.13 (0.28)	2.07 (0.16)	0.88
	Confounders controlled	1.27 (0.33)	1.13 (0.18)	0.80
	Moderators specified	1.60 (0.17)	1.47 (0.13)	0.65

quately capture the several forms of peer review that are likely to be found. In fact, peer reviewing will range from formal double blind reviewing, on through single blind reviewing, open reviewing, informal reviewing (comments by colleagues) to no reviewing at all. The strictness of reviewing may depend on factors unrelated to study quality, like the supply of manuscripts to a journal (when the supply dries up, reviewing becomes more lax to fill journal pages), the availability of other journals (it is easier to reject a paper, if it can be referred to a different journal) and how famous the author is (Nobel laureates are less likely to be rejected). Ideally speaking, all these aspects of the review process ought to be known in order to assess its performance.

Moreover, the existence of a contagion effect from journal papers to other studies cannot be ruled out. Researchers may take a good journal paper as a model for their own research and try to imitate it, even if their own report never gets published in a scientific journal. To the extent that such a contagion effect from good journal papers exists, the peer review system serves a useful function indeed by raising the quality of unpublished studies. In a comparison like the one made in this paper, however, the existence of a contagion effect will tend to reduce the differences in study quality between journal papers and other studies.

Finally, it cannot be ruled out that the standards of reviewing are too lax in some journals. This study has found examples of quite poor studies that have been published in peer reviewed journals. Imperfections are likely to be found even in the most rigorous experimentally designed studies. But the dominance of weakly controlled or inadequate non-experimental study designs in road safety evaluation studies is overwhelming. Only one of the 123 studies included in this paper was an experiment. Road safety evaluation research is largely: (1) non-experimental research, (2) based on incomplete or poor data, (3) hampered by a paucity of strong theory to support the interpretation of findings, (4) performed by government agencies or consultancies operating under severe time and cost constraints, (5) on behalf of sponsors who take a vested interest in the measures that are being evaluated and are likely to be less than perfectly objective in their assessment of the results. It is perhaps not surprising that the quality of large parts of this research leaves much to be desired.

The inherent limitations of the peer review system should not be forgotten. Peer review comes too late in the research process to influence study design, sampling plan, and the analysis of data. Sometimes peer review may lead to a reanalysis of data, but in many cases the modifications peer reviewed papers undergo as a result of reviewing, will be limited to rephrasing the conclusions and giving a more extensive discussion of results. Study design, the quality of data and the results of a study are usually not affected by peer review.

CONCLUSIONS

This paper compared the quality of road safety evaluation studies published in peer reviewed journals

to the quality of similar studies not published in peer reviewed journals. Studies that evaluated the safety effects of five road safety measures were included. These measures were roundabouts, blackspot treatment, daytime running lights for cars, seat belts and periodic motor vehicle inspection. A total of 123 evaluation studies were included. 44 studies were published in peer reviewed journals, 79 were not. Studies were compared in terms of seven criteria of study validity: (1) sampling technique, (2) total sample size, (3) mean sample size for each result, (4) specification of accident or injury severity, (5) study design, (6) specific confounding variables controlled (from a list of three variables), (7) specific moderator variables examined (from a list of two moderator variables). No clear differences in study quality were found between studies published in journals and other studies in terms of these criteria of validity. There was a slight tendency for studies published in peer reviewed journals to score higher for validity than other studies. The study reported in this paper has several limitations and should therefore be regarded as exploratory only. The results are indicative, rather than definite. The line of research explored in this paper might perhaps lead to clearer results if pursued in larger and more rigorously designed studies.

APPENDIX

List of studies with coded values for each study

Authors	Year	Country	Type of measure	Sampling technique	Sample size	No of results	Mean spl size	Acc severity specified	Study design	Confounders controlled	Moderators specified	Type of publication	Author affiliation	Author ac rank
Lalani	1975	GB	1	2	91,307	1	91,307	1	1	0	0	1	0	0
Green	1977	GB	1	2	238,199	3	79,400	1	2	1	2	0	1	1
Lahrmann	1981	DK	1	2	98,536	4	24,634	1	3	2	2	0	0	0
Cederstrnd	1983	S	1	2	1,204,054	12	100,338	1	3	3	2	0	1	1
Senneset	1983	N	1	1	5,053	1	5,053	1	2	1	2	0	1	1
Brüde	1985	S	1	2	18,570	1	18,570	1	3	2	2	0	1	1
Johannessen	1985	N	1	2	13,022	2	6511	1	3	2	2	0	1	1
Hall	1988	GB	1	2	523,024	2	261,512	1	3	2	2	0	1	1
Nygaard	1988	N	1	1	7442	1	7442	1	2	1	1	0	0	0
Glæver	1990	N	1	2	102,904	3	34,301	1	3	2	1	0	1	1
Tudge	1990	AUS	1	2	766,863	2	383,432	1	3	3	1	0	0	0
VanMinnen	1990	NL	1	2	134,758	2	67,379	1	1	0	2	0	1	1
Jørgensen	1991	DK	1	2	36,429	3	12,143	1	2	1	1	0	2	2
Brüde	1992	S	1	2	1,139,899	12	94,992	1	3	3	2	0	1	1
Dagersten	1992	CH	1	1	12,253	2	6127	1	1	0	1	0	0	0
Holzwarth	1992	D	1	2	6972	2	3486	1	1	0	2	1	0	0
Hydén	1992	S	1	2	15,588	2	7794	1	1	0	2	0	2	2
Jørgensen	1992	DK	1	2	52,510	4	13,128	1	2	1	2	0	2	2
Kristiansen	1992	N	1	2	204,944	2	102,472	1	2	1	1	0	0	0
Schnüll	1992	D	1	2	14,364	4	3591	1	2	1	2	0	1	1
Brilon	1993	D	1	1	105,086	2	52,543	1	2	1	1	0	2	2
Jørgensen	1994	DK	1	2	32,257	4	8064	1	3	2	2	0	2	2
Schoon	1994	NL	1	2	729,930	2	364,965	1	1	0	1	1	1	1
Saim	1994	N	1	1	4000	1	4000	1	2	1	1	0	2	1
Voss	1994	D	1	2	252,959	6	42,160	1	3	2	1	1	1	1
Huber	1995	CH	1	2	1,285,533	2	642,767	1	2	1	0	1	1	1
Oslo Vei	1995	N	1	1	37,800	1	37,800	1	2	1	0	0	0	0
Flannery	1996	USA	1	1	9326	1	9326	0	1	0	0	0	2	1
Exnicios	1967	USA	2	1	77,432	4	19,358	1	1	0	2	0	0	0
Maio	1967	USA	2	1	306,014	14	21,858	0	1	0	2	0	0	0
Wilson	1967	USA	2	2	80,202	4	20,051	1	2	1	2	0	0	0
Tamburri	1968	USA	2	2	160,534	4	40,134	1	2	1	2	1	0	0
Hammer	1969	USA	2	2	127,582	6	21,264	1	2	1	2	1	0	0
Dearinger	1970	GB/USA	2	1	113,980	2	56,990	0	1	0	2	0	2	1
Duff	1971	GB	2	1	193,907	12	16,159	1	1	0	2	1	0	0
Hatherly	1971	GB	2	1	158,782	2	79,391	0	1	0	2	1	0	0
Karr	1972	USA	2	2	502,261	2	251,131	1	3	2	2	0	0	0
Hvoslef	1974	N	2	1	89,792	5	17,958	1	2	1	2	0	0	0
OECD	1976	F	2	1	24,792	1	24,792	1	2	1	2	0	0	0
Hatherly	1977	GB	2	2	97,121	1	97,121	1	2	1	2	1	0	0
Vodahl	1977	N	2	1	50,992	5	10,198	1	3	2	2	0	2	1
Jørgensen	1979	DK	2	2	138,930	4	34,733	1	2	1	2	0	0	0
St vegvesen	1983	N	2	1	4800	1	4800	1	2	1	2	0	0	0
Boyle	1984	GB	2	2	2,118,741	8	264,843	1	3	2	2	1	2	1
Elvik	1985	N	2	1	17,933	1	17,933	1	2	1	0	0	1	1
Lovell	1986	USA	2	2	650,535	4	162,634	0	3	2	2	1	2	2
Persaud	1987	USA	2	2	327,695	1	327,695	0	3	3	2	1	2	1
Christensen	1988	N	2	2	166,547	4	41,637	1	2	1	0	0	1	1
Mountain	1989	GB	2	2	344,758	3	114,919	1	3	3	0	1	2	1
Corben	1990	AUS	2	2	831,194	9	92,355	1	2	1	2	0	2	1
Flagstad	1990	N	2	1	32,917	6	5486	1	2	1	1	0	2	0
Wong	1990	USA	2	1	17,850	1		0	1	0	2	1	0	0

Lalani	1991	USA	2	2	199,140	6	33,190	0	2	1	2	1	0	0
Retting	1991	USA	2	2	26,255	2	13,128	1	1	0	1	0	0	0
Sorensen	1991	DK	2	1	2083	2	1042	1	1	0	2	0	0	0
Kolster Ped	1992	DK	2	1	99,238	18	5513	1	1	0	2	0	0	0
Mountain	1992	GB	2	2	336,511	2	168,256	1	3	3	0	1	2	1
Mountain	1992	GB	2	2	27,640	1	27,640	1	3	2	0	1	2	1
Væra	1992	DK	2	2	131,735	40	3293	1	3	2	2	0	0	0
Holmskov	1993	DK	2	1	186,086	6	31,014	0	3	2	2	0	2	1
Gregory	1994	GB	2	1	185,840	1	185,840	1	3	2	0	1	2	1
Mountain	1994	GB	2	2	219,702	1	219,702	1	3	3	0	0	2	1
Legassick	1995	GB	2	1	20,188	1	20,188	1	2	1	2	0	0	0
Proctor	1995	GB	2	1	18,962	1	18,962	0	1	0	2	0	0	0
Allen	1964	USA	3	1	133,251	2	66,626	1	1	0	0	1	2	2
Cantilli	1965	USA	3	1	2767	2	1384	0	4	3	1	1	0	0
Cantilli	1970	USA	3	1	14,855	2	7428	0	4	3	1	1	0	0
Andersson	1976	SF	3	3	4,990,826	6	831,804	1	1	0	2	0	1	1
Andersson	1981	S	3	3	7,697,200	4	1,924,300	1	1	0	2	0	1	1
Attwood	1981	CDN	3	2	8533	2	4267	1	4	3	1	0	0	0
Stein	1985	USA	3	2	51,214	2	25,607	0	4	3	1	0	1	1
Vaaje	1986	N	3	3	4,219,762	2	2,109,881	1	1	0	2	0	1	1
Sparks	1989	CDN	3	1	355,033	4	88,758	1	1	0	1	0	1	1
Hocherman	1991	ISR	3	3	244,261	1	244,261	1	1	0	1	0	2	1
Elvik	1993	N	3	3	15,851,458	8	1,981,432	1	1	1	2	1	1	1
Hansen	1993	DK	3	3	3,858,862	5	771,772	1	2	3	2	0	1	1
Kuratorium	1993	A	3	2	4,712,547	10	471,255	1	2	1	1	0	1	1
Sparks	1993	CDN	3	1	158,233	1	158,233	0	1	0	1	1	1	1
Arora	1994	CDN	3	3	7,703,743	1	7,703,743	0	1	0	2	0	0	0
Hansen	1995	DK	3	3	6,841,281	5	1,368,256	1	2	3	2	0	1	1
Hollo	1995	H	3	3	995,834	3	331,945	1	2	2	2	0	1	1
Bohlin	1967	S	4	3	287,682	34	8461	1	3	2	1	0	0	0
Bäckström	1974	S	4	2	15,738	2	7869	1	2	1	1	1	0	0
Kahane	1974	USA	4	2	410,851	8	51,356	1	3	2	2	0	0	0
Reinfurt	1976	USA	4	2	309,844	6	51,641	1	1	0	1	0	2	1
Dalgaard	1977	DK	4	2	90,268	6	15,045	1	2	1	1	0	2	2
Dk statistik	1977	DK	4	2	199,814	2	99,907	1	3	2	1	0	0	0
Hartemann	1977	F	4	2	27,771	6	4629	1	3	2	1	0	1	1
Huelke	1977	USA	4	2	989,433	32	30,920	1	2	1	1	1	2	2
Sabey	1977	GB	4	3	144,982	3	48,327	1	1	0	0	0	1	1
Toomath	1977	NZ	4	2	15,154	6	2526	1	2	1	1	0	0	0
Hobbs	1978	GB	4	3	74,978	6	12,496	1	1	0	2	0	1	1
Perchonok	1978	USA	4	2	180,390	2	90,195	1	2	1	1	0	1	0
Partyka	1979	USA	4	2	21,294	2	10,647	1	1	0	1	0	0	0
Norin	1980	S	4	2	23,340	2	11,670	0	3	2	0	0	0	0
Thomas	1980	F	4	3	515,760	6	85,960	1	3	3	2	0	1	1
Cameron	1981	AUS	4	2	788,944	6	131,491	1	1	0	1	1	1	0
Hobbs	1981	GB	4	3	68,953	9	7661	1	2	1	2	0	1	1
Hobbs	1984	GB	4	3	42,873	6	7146	1	3	2	2	0	1	1
Evans	1986	USA	4	2	387,826	36	10,773	1	3	3	2	1	0	0
Evans	1988	USA	4	2	185,793	12	15,483	1	3	3	2	1	0	0
Partyka	1988	USA	4	2	1,787,595	47	38,034	1	3	3	2	0	0	0
Tunbridge	1988	GB	4	2	21,662	4	5416	1	2	1	1	0	1	1
Maghsoodloo	1989	USA	4	2	8,502,586	36	236,183	1	3	2	1	1	2	2
Krafft	1990	S	4	2	107,902	6	17,984	1	3	2	1	1	2	2
Conn	1993	USA	4	2	13,455	1	13,455	1	3	3	2	1	1	1
Dean	1995	USA	4	3	2,170,881	3	723,627	1	3	3	2	1	1	1
Elvik	1995	N	4	2	1,192,884	69	17,288	1	3	3	2	0	1	1
Huelke	1995	USA	4	2	225,684	2	112,842	1	3	2	1	1	2	2
Evans	1996	USA	4	2	6,052,865	46	131,584	1	3	2	1	1	0	0
Mayer	1963	USA	5	3	111,018,056	39	2,846,617	1	1	0	0	0	2	1
Buxbaum	1966	USA	5	2	740,187	4	185,047	1	2	1	0	1	2	1
Fuchs	1967	USA	5	2	7,718,176	1	7,718,176	1	3	2	0	1	1	1
Colton	1968	USA	5	2	4,200,715	2	2,100,358	1	2	1	0	1	2	1
Foldvary	1971	USA	5	2	4,727,273	1	4,727,273	1	2	1	0	0	0	0
Little	1971	USA	5	2	64,821,560	36	1,800,599	1	2	1	0	1	2	1
Schroer	1979	USA	5	1	639,280	1	639,280	0	2	1	1	1	2	1
Crain	1980	USA	5	3	24,181,020	3	8,060,340	1	3	2	0	0	1	1
VanMatre	1981	USA	5	3	23,147,428	2	11,573,714	1	3	2	0	1	2	2
Berg	1984	S	5	3	103,284,233	1	103,284,233	1	3	2	0	0	1	1
Loeb	1984	USA	5	2	648,209	1	648,209	1	3	2	0	1	2	1
Fosser	1992	N	5	2	15,496,119	12	1,291,343	1	4	3	2	1	1	1
Moses	1992	USA	5	2	10,937,078	4	2,734,270	1	3	2	0	1	2	1

For reasons of space, only the first author is listed for each study.

The variables have been coded as follows:

Variable	Code
Type of measure	1 = roundabout 2 = blackspot treatment 3 = daytime running lights 4 = seat belts 5 = periodic motor vehicle inspection
Sampling technique	3 = population study or random sample 2 = systematic sample 1 = self selected sample, convenience sample or unknown sampling technique
Total sample size	Sum of statistical weights for logodds method of meta-analysis
Nnmber of results	Measured directly
Mean sample size	Mean statistical weight for each result
Accident or injury severity specified	1 = accident or injury severity specified (at least for categories of fatal, injnry and non-injury) 0 = accident or injnry severity not specified
Stndy design	4 = experimental design 3 = well controlled quasi-experimental or non-experimental design 2 = weakly controlled quasi-experimental or non-experimental design 1 = inadequate study design (no control of confounding factors)
Confounders controlled	3 = all three from a list of confounders controlled (see Table 2) 2 = two of three listed confonnders controlled 1 = one of three listed confounders controlled 0 = no confounding factor from a list of three controlled
Moderators specified	2 = two moderators from a list of two specified (see Table 2) 1 = one of two moderators specified 0 = no moderator from a list of two specified
Type of publication	1 = peer reviewed journal 0 = other type of publication
Author affiliation	2 = university or hospital with research and teaching functions 1 = research institute or consnltancy 0 = bnsiness firm, government agency or nnknown affiliation
Anthor academic rank	2 = full professor 1 = other formal academic rank 0 = no formal academic rank or unknown

Roundabouts

Brilon, W., Stnwe, B. and Drews, O. (1993) Sicherheit und Leistnngsfähigkeit von Kreisverkehrsplätzen. FE Nr 77359/91. Lehrstuhl für Verkehrswesen, Ruhr-Universität Bochnm, Germany.

Brüde, U. and Larsson, J. (1985) Korsningsåtgärder vidtagna inom vägförvaltningarnas trafiksäkerhetsarbete. Regressions- och åtgärdseffekter. VTI-rapport 292. Statens väg- och trafikinstitnt, Linköping, Sweden.

Brüde, U. and Larsson, J. (1992) Trafiksäkerhet i tätortskorsningar. VTI-meddelande 685. Statens väg- och trafikinstitut, Linköping, Sweden.

Cedersund, H-Å. (1983) Cirkulationsplatser. VTI-meddelande 361. Statens väg- och trafikinstitut, Linköping, Sweden.

Dagersten, A. (1992) Roundabonts in Switzerland and Sweden. Thesis 72. University of Lund, Lund Institnte of Technology, Department of Traffic Planning and Engineering, Lund.

Flannery, A. and Datta, T. K. (1996) Modern roundabouts and traffic crash experience in the United States. Paper 960658. Transportation Research Board Annnal Meeting, Washington DC, Jannary 7–11.

Giæver, T. (1990) Ulykkesfrekvenser i rundkjøringer og signalregulerte kryss. Rapport STF63 A90002. SINTEF Samferdselsteknikk, Trondheim, Norway.

Green, H. (1977) Accidents at off-side priority roundabouts with mini or small islands. TRRL Laboratory Report 774. Transport and Road Research Laboratory, Crowthorne, Berkshire.

Hall, R. D. and McDonald, M. (1988) Jnnction design for safety. Paper presented at Roads and Traffic 2000, in *Proceedings of the International Road and Traffic Conference*, Vol. 4–2, pp. 147–151. Berlin, 6–9 September.

Holzwarth, J. (1992) Ausserorts-Kreisverkehrsplätze zur Unfallstellenbeseitigung. Ergebnisse zweier Modellvorhaben in Baden-Württemberg, *Strassenverkehrstechnik*, 36, 142–146.

Huber, C.A. (1995) Sicherheit von Kreiselanlagen. Erfahrungen und vorläufige Empfeblungen, *Zeitschrift für Verkehrssicherheit*, 41, 83–85.

Hydén, C., Odelid, K. and Várhelyi, A. (1992) Effekten av generell hastighetsdämpning i tätort. Resultat av et storskaligt försök i Växjö. I. Huvudrapport. Lunds Tekniske Høgskola, Institutionen för trafikteknik, Lund, Sweden.

Johannessen, S. (1985) Rundkjøringer. Forslag til retningslinjer basert på data om 35 rundkjøringer. Rapport STF63 A85008. SINTEF Samferdselsteknikk, 1985, Trondheim, Norway.

Jørgensen, N. O. (1991) Rundkørslers kapacitet og sikkerhed. Dokumentasjonsrapport. Danmarks Tekniske Højskole, Institut for veje, trafik og byplan, København, Denmark.

Jørgensen, E. and Jørgensen, N. O. (1992) Er der mere nyt om rundkørsler? *Dansk Vejtidsskrift*, **12**, 29–31.

Jørgensen, E. and Jørgensen, N. O. (1994) Sikkerhed i nyere danske rundkørsler. *Proceedings Trafikdage ved Aalborg Universitets Center (AUC)*, 28–30 August, pp. 191–198.

Kristiansen, P. (1992) Erfaringer med rundkjøringer i Akershus. Statens vegvesen Akershus, Oslo, Norway.

Lahrmann, H. (1981) Rundkørsler: Trafiksikkerhed, geometrisk udformning, kapacitet. Vejdirektoratet, Sekretariatet for Sikkerhedsfremmende Vejforanstaltninger, Næstved, Denmark.

Lalani, N., (1975) The impact on accidents of the introdnction of mini, small and large roundabonts at major/minor priority junctions. *Traffic Engineering and Control*, **16**, 560–561.

Nygaard, H. C. (1988) Erfaringer med rundkjøringer i Akershus. Statens vegvesen Akershus, Oslo, Norway.

Oslo Veivesen (1995) Ulykkesanalyse. Rundkjøringer i Oslo. Oslo Veivesen Trafikksikkerhetskontoret, Oslo, Norway.

Schnüll, R., Haller, W. and Von Lübke, H. Sicherheitsanliegen bei der Umgestaltung von Knotenpunkten in Städten. Forschungsberichte der Bundesanstalt für Strassenwesen (BASt) 253. Bundesanstalt für Strassenwesen, Bergisch-Gladbach, Germany.

Schoon, C. C. and Van Minnen, J. (1993) Ongevallen op rotondes II. Tweede onderzoek naar de onveiligheid van rotondes vooral voor fietsers en bromfietsers. R-93-16. Stichting Wetenschappelijk Onderzoek Verkeersveiligheid, SWOV, Leidschendam, Netherlands.

Seim, R. (1994) Analyse av kryssulykker i Akershus fylke 1990–93. Hovedoppgave i samferdselsteknikk. Institntt for samferdselsteknikk, Norges Tekniske HøgskoleTrondheim, Norway.

Senneset, G. (1983) Rnndkjøringer. Del II Hovedrapport. Erfaringer fra utvalgte rundkjøringer i Norge. Rapport STF63 A83001 II. SINTEF Samferdselsteknikk, Trondheim, Norway.

Tudge, R. T. (1990) Accidents at ronudabonts in New South Wales. Proceedings of the 15th ARRB Conference, Part 5, 331–349. Australian Road Research Board, Vermont South, Australia.

VanMinnen, J. (1990) Ongevallen op rotondes. Vergelijkende studie van de onveiligheid op een aantal locaties waar een kruispunt werd vervangen door een "nieuwe" rotonde. R-90-47. Stichting Wetenschappelijk Onderzoek Verkeersveiligheid, SWOV, Leidschendam, Netherlands.

Voss, H. (1994) Zur Verkehrssicherheit innenörtlicher Knotenpunkte. *Zeitschrift für Verkehrssicherheit*, **40**, 68–72.

Blackspot treatment

Boyle, A.J. and Wright, C. C. (1984) Accident "migration" after remedial treatment at accident blackspots. *Traffic Engineering and Control*, **25**, 260–267.

Christensen, P. (1988) Utbedringer av ulykkespunkter på riksveger og kommunale veger i perioden 1976–1983. Erfaringsrapport. TØI-rapport 0009. Transportøkonomisk Institutt, Oslo.

Corben, B. F., Ambrose, C. and Wai, F. C. (1990) Evaluation of accident black spot treatments. Report 11. Monash University, Melbourne, Australia, Accident Research Centre, 1990.

Dearinger, J. A. and Hutchinson, J. W. (1970) Cross Section and Pavement Surface. Chapter 7 of Traffic Control and Roadway Elements—Their Relationship to Highway Safety. Revised Editiou. Highway Users Federation for Safety and Mobility, Washington DC.

Duff, J.T. (1971) The effect of small road improvements on accidents. *Traffic Engineering and Control*, **12**, 244–245.

Elvik, R. (1985) Regresjonseffekt i ulykkespunkter. En empirisk undersøkelse på riksveger i Vest-Agder. Arbeidsdokument av 9.9.1985 (prosjekt O-1146). Transportøkonomisk Institutt, Oslo.

Exnicios, J. F. (1967) Accident reduction through channelization of complex intersections. In *Improved Street Utilization Through Traffic Engineering*, pp. 160–165. Highway Research Board, Special Report 93. Highway Research Board, Washington DC.

Flagstad, K. (1990) For-etter analyse av trafikksikkerhetstiltak i Bergen. Hovedoppgave i samferdselsteknikk. Institutt for samferdselsteknikk, Trondheim, Norges Tekniske Høgskole.

Gregory, M. and Jarrett, D. F. (1994) The long-term analysis øf accident remedial measures at high-risk sites in Essex. *Traffic Engineering and Control*, **35**, 8–11

Hammer, C.G. (1969) Evaluation of minor improvements. *Highway Research Record*, **286**, 33–45.

Hatherly, L.W. and Lamb, D. R. (1971) Accident prevention in London by road surface improvements. *Traffic Engineering and Control*, **12**, 524–529.

Hatherly, L.W. and Young, A. E. (1977) The location and treatment of urban skidding hazard sites. *Transportation Research Record*, **623**, 21–28.

Holmskov, O. and Lahrmann, H. (1993) Er sortpletbekæmpelse vejen frem? *Dansk Vejtidsskrift*, **2**, 3–9.

Hvoslef, H. (1974) Trafikksikkerhet i Oslo. Problemstilling, analyse og løsninger. Oslo veivesen, Oslo, Norway.

Jørgensen, E. (1979) Sikkerhedsmæssig effekt af mindre anlægsarbejder. Effektstudie. Sekretariatet for Sikkerhedsfremmende Vejforanstaltninger, Næstved, Denmark, Vejdirektoratet.

Karr, J. I. (1972) Evaluation of minor improvements—part 8, grooved pavements. Final Report. Report CA-HY-TR-2151-4-71-00. California Division of Highways, Sacramento, CA.

Kolster Pedersen, S., Knlmala, R., Elvestad, B., Ivarsson, D. and Thnresson, L. (1992) Trafiksäkerhetsåtgärder i Väg- och Gatumiljö. Exempel bämtade från de nordiska länderna nnder 1980-talet. Nnrdiske Seminar- og Arbejdsrapporter 1992, p. 607. Ministerråd, København, Nordisk.

Lalani, N. (1991) Comprehensive Safety Program Produces Dramatic Results. *ITE-Journal*, October, 31–34.

Legassick, R. (1995) Tbe case for ronte studies in road traffic accident analysis investigations. Paper presented at the conference Strategic Highway Research Program and Traffic Safety, Prague, The Czech Repnblic, September 21–22. Preprint for Sessions 21/9.

Lovell, J. and Hauer, E. (1986) The safety effect of conversion to all-way stop control. *Transportation Research Record*, 1068, 103–107

Malo, A. F. (1967) Signal Modernization. In: Improved Street Utilization Through Traffic Engineering, pp. 96–113. Highway Research Board, Special Report 93. Highway Research Board, Washington DC.

Mountain, L. and Fawaz, B. (1989) The area-wide effects of engineeriug measnres on road accident occnrrence. *Traffic Engineering and Control*, 30, 355–360.

Mountain, L. and Fawaz, B. The effects of engineering measures on safety at adjacent sites. *Traffic Engineering and Control*, 33, 15–22.

Monntain, L., Fawaz, B. and Sineng, L. (1992) The assessment of changes in accident freqnencies on link segments: a comparison of four methods. *Traffic Engineering and Control*, 33, 429–431.

Mountain, L., Fawaz, B., Wright, C., Jarrett, D. and Lupton, K. (1994) Highway improvements and maintenance: their effects on road accidents. *Paper presented at the 22nd PTRC Summar Annual Meeting*, 12–16 September. Proceedings of Seminar J., pp. 151–161.

OECD Road Research Gronp (1976) Hazardous Road Locations. Identification and Countermeasures. OECD, Paris.

Persaud, B.N. (1987) "Migration" of accident risk after remedial blackspot treatment. *Traffic Engineering and Control*, **28**, 23–26.

Proctor, S. (1995) An independent review of 3M "Road Safety" products. Paper presented at the conference Strategic Highway Research Program and Traffic Safety, Prague, The Czech Republic, September 21–22. Preprint for Sessions 22/9.

Retting, R. A. (1991) Improving Urban Traffic Safety: A Multidisciplinary Approach. Experiences From New York City 1983–1989. Prepared in conjunction with the Volvo Traffic Safety Award 1991. Thompson Printing, Belleville, NJ.

Statens vegvesen (1983) Veiledning. Håndbok 115. Analyse av ulykkessteder. Statens vegvesen, Oslo.

Sørensen, M. (1991) Forsøg med særlig afmærkning af uheldskryds. *Dansk Vejtidsskrift*, **5**, 17–19.

Tamburri, T.N., Hammer, C. J., Glennon, J. C. and Lew, A (1968) Evaluation of minor improvements. *Highway Research Record*, 257, 34–79.

Vodahl, S. B. and Johannessen, S. (1977) Ulykkesfrekvenser i kryss. Arbeidsnotat nr 7. Resultater av før/etterundersøkelsen. Oppdragsrapport 178. Norges Tekniske Høgskole, Forskningsgruppen, Institutt for samferdselsteknikk, Trondheim.

Værø, H. (1992A) Effekt af sortpletbekæmpelse i Hillerød. København, Vejdirektoratet, Trafiksikkerhedsafdelingen.

Værø, H.(1992B) Effekt af sortpletbekæmpelse i Nyborg. København, Vejdirektoratet, Trafiksikkerhedsafdelingen.

Værø, H. (1992C) Effekt af sortpletbekæmpelse i Silkeborg. København, Vejdirektoratet, Trafiksikkerhedsafdelingen.

Værø, H. (1992D) Effekt af sortpletbekæmpelse i Skælskør. København, Vejdirektoratet, Trafiksikkerhedsafdelingen.

Wilson, J. E. (1967) Simple Types of Intersection Improvements. In: Improved Street Utilization Through Traffic Engineering, 144–159. Highway Research Board, Special Report 93. Highway Research Board, Washington DC.

Wong, S-Y. (1990) Effectiveness of Pavement Grooving in Accident Reduction. *ITE Journal*, July, 34–37.

Daytime running lights

Allen, M. J. and Clark, J. R. (1964) Automobile running lights—a research report. *American Journal of Optometry and Archives of American Academy of Optometry*, **41**, 293–315.

Andersson, K. and Nilsson, G. (1981) The effects on accidents of compulsory use of running lights during daylight in Sweden. VTI-report 208A. National Road and Traffic Research Institute, Linköping, Sweden.

Andersson, K., Nilsson, G. and Salusjärvi, M. (1976) Effekt på trafikolyckor av rekommenderad och påkallad användning av varsellius i Finland. VTI-rapport 102. Statens väg- och trafikinstitut, Linköping, Sweden.

Arora, H., Collard, D., Robbins, G., Welbourue, E. R. and White, J. G. (1994) Effectiveness of Daytime Running Lights in Canada. Report TP 12298 (E). Transport Canada, Ottawa, Canada.

Attwood, D. A. (1981) The Potential of Daytime Running Lights as a Vehicle Collision Countermeasure. SAE Technical Paper 810190. Society of Automotive Engineers, Warrendale, PA.

Cantilli, E. J. (1965) Daylight "lights-on" plan by port of New York authority. *Traffic Engineeering*, *17*, December.

Cantilli, E.J. (1970) Accident experience with parking lights as running lights. *Highway Research Record*, 332, 1–13.

Elvik, R. (1993) The effects on accidents of compulsory use of daytime running lights for cars in Norway. *Accident Analysis and Prevention*, **25**, 383–398.

Hansen, L. K. (1993) Kørelys i Danmark. Effektvurdering af påbudt kørelys i dagtimerne. Notat 2/1993. Rådet for Trafiksikkerhedsforskning, København, Denmark.

Hansen, L. K. (1995) Kørelys. Effektvurdering baseret på uheldstal efter knap 3 års erfaring med kørelys. Arbejdsrapport 1/, 1995. Rådet for Trafiksikkerhedsforskning, København, Denmark.

Hocherman, I. and Hakkert, A. S. (1991) The use of daytime running lights during the winter months in Israel—evaluation of a campaign. Proceedings of the third workshop of ICTCT in Cracow, Poland, November 1990, pp. 123–131. Bulletin 94, University of Lund, Sweden, Lund Institute of Technology, Department of Traffic Planniug and Engineering.

Hollo, P. (1995) Changes of the DRL-regulations and their effect on traffic safety in Hungary. *Paper presented at the conference Strategic Highway Safety Program and Traffic Safety*, Prague, The Czech Republic, September 20–22, 1995. Preprint for sessions on September 21.

Kuratorium für Verkehrssicherheit (1993) Institut für Verkehrstechnik und Unfallstatistik. Fahren mit Licht—auch am Tag. Analyse der verkehrsunfälle beim Kraftwagendienst der Österreichischen Bundesbahnen und bei der Österreichischen Post- und Telegraphenverwaltung nach Einführung der Verwendung des Abblendlichtes auch am Tag. Wien, Austria, August.

Sparks, G. A., Neudorf, R. D. and Smith, A. E. (1989) An analysis of the use of daytime running lights in the CVA fleet in Saskatchewan. Traffic Safety Services Department, SaskAuto, Saskatoon, Saskatchewan.

Sparks, G.A., Neudorf, R. D., Smith, A. E., Wapman, K. R. and Zador, P. L. (1993) The effect of daytime running lights on crashes between two vehicles in Saskatchewan: a study of a government fleet. *Accident Analysis and Prevention*, **25**, 619–625.

Stein, H. (1985) Fleet Experience with Daytime Running Lights in the United States. SAE Technical Paper 851239. Society of Automotive Engineers, Warrendale, PA.

Vaaje, T. (1986) Kjørelys om dagen reduserer ulykkestallene. Arbeidsdokument av 15.8.1986, Q-38 CRASH. Transportøkonomisk institutt, Oslo, Norway.

Seat belts

Bohlin, N. I. (1967) A Statistical Analysis of 28,000 Accident Cases with Emphasis on Occupant Restraint Value. SAE Technical Paper 670925. Society of Automotive Engineers, New York, NY (reprinted 1968).

Bäckström, C-G., Andersson, C-E., Forsman, E. and Nilsson, I.-E. (1974) Road accidents with SAAB 99. *Journal of Traffic Medicine*, **2** (1), 1–5.

Cameron, M. H. (1981) The effect of seat belts on minor and severe injuries measured on the abbreviated injury scale. *Accident Analysis and Prevention*, **13**, 17–27.

Conn, J. M., Chorba, T. L., Peterson, T. D., Rhodes, P. and Annest, J. L. Effectiveness of safety-belt use: A study using hospital-based data for nonfatal motor-vehicle crashes. *Journal of Safety Research*, **24**, 223–232.

Dalgaard, J. B. (1977) Dræbt i bil. Ulykkesårsager og selevirkning. En trafikmedicinsk undersøgelse. Århus, Denmark, Retsmedicinsk institut.

Danmarks Statistik (1977) Færdselsuheld 1976. Kap 4, Analyse af sikkerhedsselens skadeforebyggende virkning. Danmarks Statistik, København, Denmark.

Dean, J. M., Reading, J. C. and Nechodom, P. J. (1995) Overreporting and measured effectiveness of seat belts in motor vehicle crashes in Utah. *Transportation Research Record*, 1485, 186–191.

Elvik, R. (1995) Virkninger av bilbelter i Norge. Arbeidsdokument TST/0667/95. Transportøkonomisk institntt, Oslo.

Evans, L. (1986) The effectiveness of safety belts in preventing fatalities. *Accident Analysis and Prevention*, **18**, 229–241.

Evans, L. (1988) Rear seat restraint system effectiveness in preventing fatalities. *Accident Analysis and Prevention*, **20**, 129–136.

Evans, L. (1996) Safety-belt effectiveness: the infinence of crash severity and selective recruitment. *Accident Analysis and Prevention*, **28**, 423–433.

Hartemann, F., Thomas, C., Henry, C., Foret-Bruno, J-Y., Faverjon, G., Tarriere, C., Got, C. and Patel, A. (1977) Belted or not-belted: The only difference between two matched samples of 200 car occupants. Paper 770917. *Proceedings of Twenty-First Stapp Car Crash Conference*, pp. 97–150.

Hobbs, C. A. (1978) The effectiveness of seat belts in reducing injuries to car occupants. TRRL Laboratory Report 811. Transport and Road Research Laboratory, Crowthorne, Berkshire.

Hobbs, C. A. (1981) Car occupant injury patterns and mechanisms. TRRL Supplementary Report 648. Transport and Road research Laboratory, Crowthorne, Berkshire.

Hobbs, C. A and Mills, P. J. (1984) Injury probability for car occupants in frontal and side impacts. TRRL Laboratory Report 1124. Transport and Road Research Laboratory, Crowthorne, Berkshire.

Huelke, D. F. and Compton, C. P. The effects of seat belts on injury severity of front and rear seat occupants in the same frontal crash. *Accident Analysis and Prevention*, **27**, 835–838.

Huelke, D. F., Lawson, T. E., Scott, R. and Marsh, J. C. (1977) The effectiveness of belt systems in frontal and rollover crashes. *Journal of Traffic Medicine*, **5** (1), 8–21.

Kahane, C. J. (1974) Usage and Effectiveness of Seat and Shoulder Belts in Rural Pennsylvania Accidents. NHTSA Technical Note DOT HS-801 398. US Department of Transportation, National Highway Traffic Safety Administration, Washington DC.

Krafft, M., Nygren, C. and Tingvall, C. (1990) Rear seat occupant protection. A study of children and adults in the rear seat of cars in relation to restraint use and characteristics. *Journal of Traffic Medicine*, **18** (2), 51–60.

Maghsoodloo, S., Brown, D. B. and Shieh, Y-I. (1989) A quantification of the impact of restraining systems on passenger safety. *Journal of Safety Research*, 20, 115–128.

Norin, H., Nilsson-Ehle, A., Saretok, E. and Tingvall, C. (1980) Injury—reducing effect of seat belts on rear seat passengers. Volvo Car Corporation and The Swedish Road Safety Office, Göteborg and Borlänge, Sweden.

Partyka, S. C. (1979) Fatal accidents in the first fifteen months of the National Crash Severity Study. *Proceedings of Twenty-Third Conference of the American Association for Automotive Medicine*, pp. 77–89, Louisville, KY, October 3–6.

Partyka, S. C. (1988) Papers on Adult Seat Belts—Effectiveness and Use. Report DOT HS 807 285. US Department of Trausportation, National Highway Traffic Safety Administration, Washington DC.

Perchonok, K., Ranney, T. A., Baum S., Morris, D. F. and Eppich, J. D. (1978) Hazardous Effects of Highway Features and Roadside Objects. Volume 2: Findings. Report FHWA-RD-78-202. US Department of Transportation, Federal Highway Administration, Washington.

Reinfurt, D. W., Silva, C. Z. and Seila, A. F. (1976) A Statistical Analysis of Seat Belt Effectiveness in 1973–75 Model Cars Involved in Towaway Crashes. Report DOT-HS-5-01255, US Department of Transportation, National Highway Traffic Safety Administration, Washington DC.

Sabey, B. E., Grant, B. E. and Hobbs, C. A. (1977) Alleviation of injuries by use of seat belts. TRRL Supplementary Report 289. Transport and Road Research Laboratory, Crowthorne, Berkshire.

Thomas, C., Faverjon, G., Henry, C., Tarriere, C., Got, C. and Patel, A. (1980) Comparative study of 1624 belted and 3242 non-belted occupants: results on the effectiveness of seat belts. *Proceedings of the Twenty-Fourth Conference of the American Association for Automotive Medicine*, pp. 422–436, October 7–9.

Toomath, J. B. (1977) Compulsory seat belt legislation in New Zealand. *Proceedings of the Sixth International Conference of the International Association for Accident and Traffic Medicine*, pp. 21–39, Melbourne, Australia, January 31–February 4, 1977.

Tunbridge, R. J., Everest, J. T., Wild, B. R. and Johnstone, R. A. (1988) An in-depth study of road accident casualties and their injury patterns. Research Report 136. Transport and Road Research Laboratory, Crowthorne, Berkshire.

Periodic motor vehicle inspection

Berg, G., Danielsson, S. and Jnnghard, O. (1984) Trafiksäkerhet och periodisk fordonskontroll. VTI-rapport 281. Väg- och Trafikinstitntet, Linköping, Sweden.

Buxbaum, R.C. and Colton, T. (1966) Relationship of motor vehicle inspection to accident mortality. *Journal of the American Medical Association*, **197**, 31–36.

Colton, T. and Buxbaum, R. C. (1968) Motor vehicle inspection and accideut mortality. *American Journal of Public Health*, **58**, 109–1099.

Crain, W. M. (1980) Vehicle Safety Inspection Systems. How Effective? AEI studies 258. American Entreprise Institute for Public Policy Research, Washington DC.

Foldvary, L. A. (1971) A Review of Vehicle Inspection in relation to road safety. Report NR/9. Australian Department of Transport, Canberra.

Fosser, S. (1992) An experimental evaluation of the effects of periodic motor vehicle inspection on accident rates, *Accident Analysis and Prevention*, **24**, 599–612.

Fuchs, V.R. and Leveson, I. (1967) Motor accident mortality and compulsory inspection of vehicles. *Journal of the American Medical Association*, **201**, 657–661

Little, J.W. (1971) Uncertainties in evaluating periodic motor vehicle inspection by death rates. *Accident Analysis and Prevention*, **3**, 301–313

Loeb, P.D. and Gilad, B. (1984) The efficacy and cost-effectiveness of vehicle inspection. *Journal of Transport Economics and Policy*, **18**, 145–164

Mayer, A. J. and Hoult, T. F. (1963) Motor Vehicle Inspection. A Report on Current Information, Measurement, and Research. Wayne State University, Institute for Regional and Urban Studies.

Moses, L. N. and Savage, I. (1992) The effectiveness of motor carrier safety audits. *Accident Analysis and Prevention*, **24**, 479–496

Schroer, B.J. and Peyton, W. F. (1979) The effects of automobile inspections on accident rates. *Accident Analysis and Prevention*, 11, 61–68.

VanMatre, J.G. and Overstreet, G. A. (1981) Motor vehicle inspection and accident mortality: A re-examination. *Journal of Risk and Insurance*, **48**, 423–435.

REFERENCES

Begg, C. B. and Berlin, J. A. (1988) Publication bias: a problem in interpreting medical data. *Journal of the Royal Statistical Society, Series A* **151**, 3, 419–463.

Blank, R. M. (1991) The effects of double-blind versus single-blind reviewing: Experimental evidence from the American Economic Review. *American Economic Review* **81**, 1041–1067.

Campbell, D. T. and Stanley, J. A. (1966) *Experimental and Quasi-Experimental Designs for Research*. RandMcNally, Chicago.

Chalmers, T. C., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D. and Ambroz, A. (1981) A method for assessing the quality of a randomized clinical trial. *Controlled Clinical Trials* **2**, 31–49.

Cook, T. D. and Campbell, D. T. (1979) *Quasi-Experimentation. Design and Analysis Issues for Field Settings*. RandMcNally, Chicago.

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

Dickersin, K. and Min, Y.-I. (1993) Publication bias: The problem that won't go away. *Annals of the New York Academy of Sciences*, **703**, 135–148.

Elvik, R. (1996) A meta-analysis of studies concerning the safety effects of daytime rnnning lights on cars. *Accident Analysis and Prevention* **28**, 685–694.

Elvik, R. (1997) Evaluations of road accident blackspot treatment: A case of the Iron Law of evaluation studies?. *Accident Analysis and Prevention* **29**, 191–199.

Elvik, R., Salusjärvi, M. and Syvänen, M. (1993) *Peer-Review of TFB-Funded Research on Road Safety*. TFB-Information 11-1993. The Swedish Transport Research Board, Stockholm.

Evans, L. (1991) *Traffic Safety and the Driver*. VanNostrand Reinhold, New York.

Fleiss, J. L. (1981) *Statistical methods for rates and proportions*. Second edition. Wiley, New York.

Hauer, E. (1998) A case for science-based road safety design and management. *Paper presented at Highway Safety At the Crossroads, San Antonio, TX, March 1988* (Quoted from manuscript as submitted to the conference), ed. R.F. Stammer. Proceedings published by American Society of Civil Engineers.

Light, R. J. and Pillemer, D. B. (1985) *Summing Up. The*

Science of Reviewing Research. Harvard University Press, Cambridge.

Mitchell, R. C. and Carson R. T. (1989) *Using Surveys to Value Public Goods: The Contingent Valuation Method*. Washington DC, Resources for the Future. The Johns Hopkins University Press, Baltimore.

Rosenthal, R. M. (1991) Meta-analytic Procedures for Social Research. Revised Edition. *Applied Social Research Methods*, Vol. 6. Sage Publications, Newbury Park.

Wortman, P. M. (1994) Judging research quality. In *The Handbook of Research Synthesis*, eds. H. Cooper and L.V. Hedges, pp. 97–109. Russell Sage Foundation, New York.