Rune Elvik

TØI report 984/2008

Institute of Transport Economics
Norwegian Centre for Transport Research

Making sense of road safety evaluation studies

Developing a quality scoring system

TØI report
984/2008

Making sense of road safety evaluation studies

Developing a quality scoring system

Rune Elvik

ISSN 0808-1190
ISBN 978-82-480-0919-1 Paper version
ISBN 978-82-480-0920-7 Electronic version

Oslo, October 2008

Title: Making sense of road safety evaluation

Author(s): *Rune Elvik*

TØI report 984/2008
Oslo, 2008-10
140 pages
ISBN 978-82-480-0919-1 Paper version
ISBN 978-82-480-0920-7 Electronic version
ISSN 0808-1190

Financed by:
The Research Council of Norway

Project: 2581 SIP - meta-analysis

Project manager: Rune Elvik

Quality manager: Marika Kolbenstvedt

Key words:
Meta-analysis; Evaluation studies; Road safety measures; Study quality; Numerical quality scale

Summary:

The report discusses the development of a numerical quality scoring system for road safety evaluation studies. Previous research related to quality assessment of scientific studies is reviewed, but it is concluded that this research offers few guidelines with respect to quality scoring of road safety evaluation studies. A quality scale is nevertheless proposed and tested on a few road safety evaluation studies.

Tittel: Vurdering av kvaliteten på undersøkelser om virkninger av trafikksikkerhetstiltak

Forfatter(e): *Rune Elvik*

TØI rapport 984/2008
Oslo: 2008-10
140 sider
ISBN 978-82-480-0919-1 Papirversjon
ISBN 978-82-480-0920-7 Elektronisk versjon
ISSN 0808-1190

Finansieringskilde:
Norges forskningsråd

Prosjekt: 2581 SIP - meta-analyser

Prosjektleder: Rune Elvik

Kvalitetsansvarlig: Marika Kolbenstvedt

Emneord:
Meta-analyse; Evalueringsstudier; Trafikksikkerhetstiltak; Kvalitet på undersøkelser; Kvalitetsskala

Sammendrag:

Rapporten drøfter muligheten for å utvikle et tallmessig mål på kvaliteten på undersøkelser om virkninger av trafikksikkerhetstiltak. Tidligere forskning på området gjennomgås, og det konkluderes med at denne gir relativt liten veiledning om hvordan man best kan måle kvaliteten på undersøkelser. En skala for å bedømme kvaliteten på undersøkelser om virkninger av trafikksikkerhetstiltak blir likevel foreslått.

Language of report: English

The report can be ordered from:
Institute of Transport Economics, The library
Gaustadalleen 21, NO 0349 Oslo, Norway
Telephone +47 22 57 38 00 - www.toi.no

Rapporten kan bestilles fra:
Transportøkonomisk institutt, Biblioteket
Gaustadalleen 21, 0349 Oslo
Telefon 22 57 38 00 - www.toi.no

Preface

Thousands of road safety evaluation studies have been published. A summary of the results of many of these studies is given in the “Handbook of Road Safety Measures” (Elvik and Vaa, Elsevier Science, Oxford, 2004), which is currently being updated. The Handbook of Road Safety Measures presents detailed information regarding the effects on road safety of close to 130 road safety measures. The detailed and sometimes very precise information presented gives readers an impression that very extensive knowledge exists concerning the effects of road safety measures.

It is true that there is extensive knowledge. It is, however, also a fact that not all road safety evaluation studies are as methodologically rigorous as one would want them to be. There are many methodologically weak studies, and it is easy to give examples showing how various shortcomings of these studies influence their findings. The objective of this report is to develop a system for assessing the quality of road safety evaluation studies and assign a numerical score to study quality.

The report is the final documentation of the Strategic Research Programme (SIP): “Using meta-analysis to summarise knowledge in transport research”, which was formally finished in 2004. Due to other commitments, finalising this report has taken considerably longer than expected. A previous report, entitled Topics in meta-analysis (report 692, was authored by Peter Christensen and published in 2003. Rune Elvik, having acted as manager of the research programme, is the author of the present, final report. Head of Department Marika Kolbenstvedt has been in charge of quality assurance.

The topic discussed in this report turned out to be more complicated than originally envisaged. It was hoped that a well-justified numerical quality scoring system for road safety evaluation studies could be developed by surveying previous research related to the topic. Unfortunately, previous research did not provide much guidance. A numerical quality scoring system is nevertheless proposed in the report, but it must be admitted that this system contains some elements of arbitrariness. Assessing study quality numerically is so complex, that some degree of arbitrariness appears to be inevitable in any system developed for this purpose.

Oslo, October 2008
Institute of Transport Economics

Lasse Fridstrøm
Managing director

Marika Kolbenstvedt
Head of department

Contents

Summary

Sammendrag

1 The importance of study quality 1
1.1 What is study quality and does it matter? 1
1.2 Problems discussed in this report...2

2 A systematic approach to the assessment of study quality 4
2.1 Elements of a systematic approach 4
2.2 The effect of study quality on study results: a sample of horror stories from road safety evaluation studies 5
2.2.1 Case 1: Black spot treatment 5
2.2.2 Case 2: Evaluation of road safety measures in Norway 6
2.2.3 Case 3: The effects of technical inspections of heavy vehicles.......................7
2.2.4 Case 4: Claiming to control for confounding factors that were actually not controlled for 9
2.3 An ideal quality scoring system 11

3 A review of scales for study quality..14
3.1 What can be learnt from existing quality scales?..14
3.2 Main findings of review 15
3.2.1 Scales that have been reviewed 15
3.2.2 The concept of study quality underlying the scales – items covered 17
3.3 Lessons to be learnt...25

4 What do the experts think?...28
4.1 A survey of ten prominent road safety experts 28
4.2 Results of the survey 28
4.3 Study quality assessment for the Highway Safety Manual.............................33
4.4 Lessons from listening to the experts and studying a quality rating system proposed by one of them 36

5 Testing a pilot scoring system for study quality 38
5.1 The scoring system...38
5.2 Testing reliability and validity 42
5.2.1 Reliability...42
5.2.2 Validity...44
5.3 Lessons from the pilot testing 45

6 A typology of study designs and threats to validity 48
6.1 The validity of study quality assessments...48
6.2 Study designs used in road safety evaluation studies 49
6.3 Important threats to validity in road safety evaluation studies 50
6.4 How well do different designs control for threats to internal validity? 51
6.5 A hierarchy of study designs...53

6.6 Summary of lessons learnt 53

7 Elements of a rationally justified quality scoring system for road safety evaluation studies 55

7.1 The problem of arbitrariness in quality scoring systems 55

7.2 Elements of a rational (non-arbitrary) quality scoring system 56

7.3 A review of methodological research regarding road safety evaluation studies 62

7.3.1 Experimental study designs 62

7.3.2 Before-and-after studies 64

7.3.3 Cross-section studies 66

7.3.4 Case-control studies 70

7.3.5 Multivariate accident models 73

7.3.6 Time-series analysis 77

7.4 Lessons from methodological research 78

8 A formal quality scoring system for road safety evaluation studies 80

8.1 Framework for the system 80

8.2 The scoring system 81

8.2.1 Is there a need for ranking study designs? 81

8.2.2 Items common to all study designs 81

8.2.3 Items specific to each study design 86

8.2.4 Assigning weights to items included in quality scoring 90

8.3 Application of the system – test cases 92

8.3.1 Converting scores to a bounded scale 93

8.3.2 Experiment versus multivariate study 93

8.3.3 Before-and-after studies 96

8.3.4 Cross-section studies 96

8.3.5 Case-control studies 97

8.3.6 Multivariate analyses 98

8.3.7 Time series analyses 98

8.3.8 Comparison of findings 98

8.4 Summary of main lessons 100

9 The use of quality scores in meta-analysis 102

9.1 Different approaches to the treatment of study quality in meta-analysis.. 102

9.2 The statistical treatment of quality scores 103

9.3 Example 1: Daytime running lights 104

9.4 Example 2: Roundabouts 107

9.5 Example 3: Speed and road accidents 108

9.6 The treatment of study quality in meta-analysis 109

10 Discussion and conclusions 111

10.1 Discussion 111

10.2 Conclusions 113

11 References 115

Appendix 1: Test of pilot quality scoring system 128

Appendix 2: Scoring of 18 studies by means of quality scale 131

TØI report 984/2008
Author(s): Rune Elvik
Oslo 2008, 140 pages

Summary:

Making sense of road safety evaluation studies

This report presents a systematic approach to assessing the quality of road safety evaluation studies. These are studies that evaluate the effects of road safety measures. The report is the final documentation of a strategic research programme on the use of meta-analyses to summarise knowledge in transport research, funded by the Research Council of Norway.

Background and research problem

Literally thousands of road safety evaluation studies have been reported. A large share of these studies are referred to in the Handbook of Road Safety Measures, which is continually being expanded and updated. This book presents quite detailed information about the effects of nearly 130 road safety measures, possibly giving readers the impression that this is a topic where extensive knowledge exists.

It is correct that many studies have been made, but the quality of these studies varies considerably. It is easy to give examples of bad studies, and it is easy to show how the methodological shortcomings of these studies have influenced their findings. The report gives some examples of this. This forms the background for asking the main research question addressed by this report:

Is it possible to assess the quality of road safety evaluation studies in a systematic way, preferably by means of a numerical scale for study quality?

Many attempts have been made to develop numerical scales intended to measure study quality, in particular in medicine. A serious objection to nearly all these scales is that they are to a large extent arbitrary, in the sense that no reasons are given for the selection of items included, nor for the weighting of these items. Study quality is, in other words, a concept that cannot easily be operationalised (made measurable).

Is a non-arbitrary scoring of studies for quality possible?

Major emphasis is put in this report on developing an approach to assessing study quality that minimises the element of arbitrariness. To this end, the report consists of the following studies:

- A review of previously developed scales for study quality,

The report can be ordered from:
Institute of Transport Economics, Gaustadalléen 21, NO 0349 Oslo
Telephone: +47 22 57 38 00   Telefax: +47 22 60 92 00

- A survey of how leading road safety researchers understand the concept of study quality and what they think about trying to measure study quality numerically,
- Developing and testing a pilot version of a scale intended to measure study quality,
- Developing a typology of study designs and threats to validity in road safety evaluation studies,
- A review of methodological research that has investigated how various aspects of study design and analysis influence the findings of road safety evaluation studies.

Existing scales for study quality

A total of 35 scales for measuring study quality have been reviewed. The review is not intended to be exhaustive. Most of the scales reviewed were developed in medicine. Only a few scales developed for assessing the quality of road safety evaluation studies were identified.

Very few of the scales are based on a formal definition of the concept of study quality. The items covered by the scales vary considerably and reflect widely divergent views about what constitutes study quality. A total of 158 variables were coded to capture the contents of the scales; these variables were subsequently reduced to 12 main categories. It is, however, not clear that all of the 12 main categories address aspects of study quality; it can be argued that some of them do not. Reliability is not known for all of the scales; it appears to be satisfactory when ever known. Validity has hardly been tested; some of the few tests reported make little sense.

On the whole, it must be concluded that the review of existing scales for study quality confirms the criticism that has been made against such scales, namely that the scales are arbitrary, subjective, not well justified and almost never tested in a scientifically defensible way. The scales are, in other words, the result of sloppy work and studying them produced nothing that could be used in developing a scale for measuring the quality of road safety evaluation studies.

Expert views about study quality

Four open questions dealing with the quality of road safety evaluation studies were asked to a convenience sample of 10 leading road safety researchers around the world. Eight replies were received. The answers showed that there is no consensus about the meaning of the concept of study quality. It was not possible to develop a concise definition of the concept based on the replies given in the survey. Opinions also differed with respect to which are the most common weaknesses of road safety evaluation studies. Several experts did, however, state that poor control for confounding factors was a major weakness of many road safety evaluation studies. As far as the possibility of developing a numerical score for study quality was concerned, most experts did not reject this idea, but many voiced concern about the large element of arbitrariness (or subjectivity) involved in scoring studies for quality.

One of the researchers who answered the survey, Ezra Hauer, has recently developed a numerical scale for assessing study quality, intended for use in the forthcoming Highway Safety Manual in the United States. This scale is presented and some elements of it have been used in the scale proposed in this report.

A pilot version of a quality scale

In 2000, a pilot version of a numerical scale for measuring the quality of road safety evaluation studies was developed by the author of this report. The scale consisted of 10 items, each of which was scored on an ordinal scale. Five researchers scored five studies each independently of each other in a pilot test of the scale. The scale was found to have an acceptable level of reliability. Testing the validity of the scale turned out not to be possible. The idea was originally to use the conception of study quality extracted from the survey of the experts as a "gold standard" and compare the scale to this standard. However, expert opinion on study quality turned out to be too divergent to serve as a gold standard.

Another lesson learnt in testing the scale, was that its discriminative power appeared to be small. All five studies selected were assigned almost the same score for quality, although the initial impression of these studies was that their quality differed. The scale was rejected and has not subsequently been used.

A typology of study designs and threats to validity

The lessons learnt from studying existing quality scales, expert opinion and a pilot version of a quality scale suggested that a broad perspective on study quality and a wide-ranging survey of factors influencing study quality need to be adopted in order to develop a numerical scale for study quality. For this purpose, a typology of study designs and threats to internal validity in road safety evaluation studies was developed.

The most commonly applied study designs in road safety evaluation studies (there are many versions of each design) are:

1. Experiments (randomised, controlled trials; rarely used)
2. Before-and-after studies (many versions exist; a very common design)
3. Cross-section studies (without statistical modelling; used to be common)
4. Case-control studies (applied mostly to evaluate injury-reducing measures)
5. Multivariate accident models (statistical models; is becoming more common)
6. Time-series analysis (applied in alcohol-control studies; otherwise rare)

For each of these study designs, major threats to internal validity were identified. Internal validity refers to the possibility of inferring a causal relationship between a road safety measure and changes in road safety.

Methodological research

In order to select items to be included in a scale intended to measure study quality, it is necessary to know which aspects of study methodology influence study findings and how large the influence is. A study is of good quality if there is a small probability that methodological weaknesses influenced study findings.

Accordingly, methodological research is research designed to assess how various aspects of study design and methods influence, or may influence, study findings. This type of research can serve as a basis for developing a scale intended to measure study quality, by identifying items to be included (which aspects of study methods are relevant) and by providing a basis for assigning weights to the items included (if aspect A of the method is found to exert a stronger influence on study findings than aspect B).

Methodological research related to road safety evaluation studies was reviewed. The amount of methodological research varies considerably between different study designs; hence more is known about potential sources of error for some designs than for others. Results turned out to be difficult to interpret. It was found that even such well-known sources of error as not controlling for regression-to-the-mean in before-and-after studies did not always influence study findings greatly. When lack of control for regression-to-the-mean did in fact influence study findings, neither the direction nor the size of the impact were consistent. It has almost become a canon of faith that not controlling for regression-to-the-mean will invariably result in a gross exaggeration of the effects of the road safety measure. This was not found to be the case. Results are, unfortunately, a lot more untidy. Still, they underscore the importance of controlling for potentially confounding factors.

The review of methodological research did therefore not provide a useful basis for assigning weights to different items in a scale designed to assess study quality.

A scale for assessing the quality of road safety evaluation studies

The attempts that have been made to develop a scientific foundation for developing a numerical scale for assessing study quality must be rated as largely unsuccessful. Despite this, a scale has been developed and is presented in this report. As any other scale found in the literature, the scale presented in this report contains a large element of arbitrariness. At the current stage of knowledge, this appears to be inevitable. The choice facing researchers is either: (A) To conclude that there is no way of measuring the concept of study quality in a scientifically defensible way, or: (B) To try to measure study quality, fully recognising the fact that not all elements of the scale used can be fully justified by referring to well-established knowledge.

The scale consists of two parts. Part one, standard items, are common to all study designs employed in road safety evaluation studies. Part two consists of items that have been customised to each study design. The scale has a bounded range. A perfect study will score 1; a worthless study will score 0. The various study designs have not been ranked; thus, a good study employing any design may attain a score close to 1 for quality. The standard items count for 50 %; the design-specific items counts for the other 50 %. Weights have been assigned to the items making up each part of the scale.

The scale is based on criteria of internal validity; i.e. operational criteria of causality designed to help assess the basis for inferring a causal relationship between a road safety measure and observed changes in road safety. These criteria have been developed and applied in several previous studies. The number of items that must be scored varies somewhat according to study design, but is between 10 and 20.

The scale was tested by applying it to 18 studies. These studies scored between 0.863 for the best study and 0.131 for the worst study. The reliability and validity of the scale is not known.

The treatment of study quality in meta-analysis

Several approaches can be taken to the treatment of study quality in meta-analysis. The following three approaches are all defensible:

1. Identify items of study quality, score each item and use a variable representing each item as an explanatory variable in a meta-regression analysis,
2. Develop an overall quality score and use it as an explanatory variable in meta-regression analysis,
3. Assign a quality weight to each study and adjust the statistical weight of study by means of the quality weight. Studies scoring close to 0 for quality will then have their weight greatly reduced.

Examples are given of all these approaches.

TØI-rapport 984/2008
Forfatter(e): Rune Elvik
Oslo 2008, 140 sider

Sammendrag:

Vurdering av kvaliteten på undersøkelser om virkninger av trafikksikkerhetstiltak

Denne rapporten presenterer en undersøkelse av muligheten for å utvikle et systematisk opplegg for å vurdere kvaliteten på undersøkelser om virkninger av trafikksikkerhetstiltak. Rapporten utgjør den siste rapporten fra det strategiske instituttprogrammet "Bruk av meta-analyser til kunnskapsoppsummering i transportforskning", som formelt pågikk fra 2000 til 2004.

Bakgrunn og problemstilling

Omfanget av forskning øker på nesten alle fagområder og det er en stor utfordring å sammenfatte foreliggende kunnskap på en konsis og riktig måte. Ett av problemene man møter på mange fagområder, er at kvaliteten på foreliggende undersøkelser varierer. Man ønsker da å legge mest vekt på de beste undersøkelsene. Dette krever at man kan bedømme kvaliteten på undersøkelser på en systematisk måte.

Det foreligger i dag flere tusen studier om virkninger av trafikksikkerhetstiltak. Mange av disse studiene er oppsummert i Trafikksikkerhetshåndboken, som er under kontinuerlig oppdatering og utvikling. I Trafikksikkerhetshåndboken presenteres til dels svært detaljerte opplysninger om virkninger av mange trafikksikkerhetstiltak, noe som kan gi inntrykk av at det foreligger omfattende kunnskap om virkninger av slike tiltak.

Det er riktig at det er utført omfattende forskning om virkninger av trafikksikkerhetstiltak, men dessverre er ikke all denne forskningen av like god kvalitet. Det er lett å finne eksempler på dårlige undersøkelser, og det er lett å vise eksempler på hvordan svakheter ved de dårlige undersøkelsene har påvirket resultatene av dem. I rapporten gis en rekke slike eksempler.

På denne bakgrunn, er hovedproblemstillingen denne rapporten tar sikte på å besvare:

Er det mulig å bedømme kvaliteten på undersøkelser om virkninger av trafikksikkerhetstiltak på en systematisk måte, fortrinnsvis i form av en tallmessig skala for undersøkelsers kvalitet?

Det er tidligere, spesielt i medisin, gjort mange forsøk på å utvikle tallmessige mål på undersøkelsers kvalitet. En tungtveiende innvending mot de aller fleste av disse målene, er at de i stor grad er vilkårlige, det vil si at det i liten grad gis noen begrunnelse av hva som inngår i dem og hvordan ulike poster er vektet i forhold til hverandre. Kvalitet på undersøkelser er følgelig et begrep det er vanskelig å operasjonalisere på en velbegrunnet måte.

Kan en ikke-vilkårlig skala for kvalitet utvikles?

I denne rapporten er det lagt vekt på å etablere et grunnlag for å utvikle en skala for kvalitet der de ulike elementene i størst mulig grad begrunnes, slik at vilkårligheten reduseres. For å oppnå dette, er flere tilnærmingsmåter valgt:

- Gjennomgang av tidligere utviklede skalaer for tallfesting av undersøkelsers kvalitet,
- Spørreundersøkelse blant ledende trafikksikkerhetsforskere om hva de legger i begrepet kvalitet og om de mener det er mulig å lage et tallmessig mål på undersøkelsers kvalitet,
- Utvikling og testing av en pilotversjon av en skala for kvalitet på undersøkelser om trafikksikkerhetstiltak,
- Utvikling av en typologi av undersøkelsesopplegg i studier av trafikksikkerhetstiltak og mulige feilkilder i slike undersøkelser,
- Gjennomgang av metodologisk forskning om hvilken betydning ulike feilkilder kan ha for resultatene av undersøkelser om virkninger av trafikksikkerhetstiltak.

Tidligere kvalitetsskalaer

35 ulike skalaer som er utviklet for å måle kvaliteten på undersøkelser er gjennomgått. De aller fleste av disse skalaene er utviklet i medisin. Kun et fåtall skalaer for trafikksikkerhetsstudier ble funnet.

De færreste skalaer for kvalitet på undersøkelser bygger på en klar definisjon av begrepet kvalitet. Det varierer svært mye hva som inngår i skalaene, og hele 158 variabler ble kodet for å definere innholdet i de 35 skalaene. Disse 158 variablene kan reduseres til 12 hovedkategorier. Det er imidlertid høyst tvilsomt om alle disse kategoriene har særlig mye med undersøkelsers kvalitet å gjøre. Reliabiliteten er testet for noen av skalaene og har vist seg å være god. Validiteten av skalaene er i liten grad testet, og de få tester som foreligger er til dels meningsløse.

I det hele tatt må det konkluderes med at gjennomgangen av tidligere utviklede skalaer for undersøkelsers kvalitet langt på veg bekrefter den kritikk som har vært reist mot slike skalaer, nemlig at de er vilkårlige, ubegrunnede, subjektive og ikke testet på en vitenskapelig holdbar måte. Praktisk talt ingen ting av nytte for å utvikle en skala for kvalitet på undersøkelser om virkninger av trafikksikkerhetstiltak kom ut av gjennomgangen av foreliggende skalaer.

Ekspertoppfatninger om kvalitet

Fire åpne spørsmål om kvalitet på undersøkelser om virkninger av trafikksikkerhetstiltak ble sendt til 10 av verdens ledende trafikksikkerhetsforskere. Det kom 8 svar. Svarene viste at oppfatningene om hva som ligger i begrepet kvalitet varierer mye. Det var ikke mulig å formulere en kort og konsis definisjon av begrepet på grunnlag av de svar som ble gitt. Det var også ulike oppfatninger om hva som er de vanligste svakhetene ved studier av trafikksikkerhetstiltak. Mange nevnte imidlertid dårlig kontroll for bakenforliggende eller andre forstyrrende variabler (”confounding factors”) som en viktig feilkilde. Når det gjaldt mulighetene for å måle kvaliteten på undersøkelser tallmessig, var de fleste ikke avvisende til tanken om dette, men det fantes en viss skepsis til om et slikt mål ville inneholde et for stort element av vilkårlighet.

En av de forskere som ble spurt, Ezra Hauer, har nylig utviklet en skala for å bedømme kvaliteten på undersøkelser som ledd i utviklingen av Highway Safety Manual i USA. Denne skalaen presenteres og visse elementer i den er benyttet i den skala som er utviklet i denne rapporten.

En pilotversjon av en skala

Det ble i 2000 utviklet en pilotversjon av en skala for kvalitet på undersøkelser om virkninger av trafikksikkerhetstiltak. Skalaen bestod av 10 poster som ble scoret med en ordinal skala. Fem forskere scoret uavhengig av hverandre fem studier om virkninger av trafikksikkerhetstiltak for å teste skalaen. Skalaen hadde en akseptabel reliabilitet. Det viste seg å være umulig å teste dens validitet. Tanken var opprinnelig å gjøre dette ved å sammenholde skalaen med en ”gullstandard”, representert ved de ledende trafikksikkerhetsforskernes oppfatning om kvalitet. Det viste seg imidlertid at disse oppfatningene var så sprikende at de ikke kunne brukes som validitetskriterium.

En annen erfaring med skalaen var at den i liten grad diskriminerte mellom de fem utvalgte undersøkelsene. Alle fikk tildelt omtrent samme poengsum, selv om det på forhånd var antatt at disse undersøkelsene representerte arbeider med ulik kvalitet. Skalaen ble forkastet og har ikke vært benyttet etter pilotstudien.

Typologi av undersøkelsesopplegg og feilkilder

Erfaringene med å gjennomgå tidligere kvalitetsskalaer, spørre ledende forskere, samt teste en pilotversjon av en skala viste at man for å utvikle en hensiktsmessig skala for kvalitet må bygge på en bred forståelse av begrepet kvalitet og en omfattende gjennomgang av faktorer som påvirker undersøkelsers kvalitet. Til dette formål er det utviklet en typologi av undersøkelsesopplegg og mulige feilkilder i hvert undersøkelsesopplegg.

De vanligste undersøkelsesopplegg i studier om virkninger av trafikksikkerhets-tiltak er (det finnes flere varianter av hvert opplegg):

1. Eksperimenter (randomiserte, kontrollerte forsøk; brukes sjelden)
2. Før-og-etter undersøkelser (mange varianter; brukes ofte)
3. Tverrsnittsstudier (uten statistisk modellering; tidligere mye brukt)
4. Case-control studier (brukes mest om skadereduserende tiltak)
5. Multivariate, statistiske ulykkesmodeller (brukes mer og mer)

6. Tidsrekkeanalyser (brukes mye om promillekjøring; lite ellers)

For hvert av disse oppleggene, ble de viktigste feilkildene knyttet til intern validitet identifisert. Med intern validitet menes grunnlaget for å trekke slutninger om årsakssammenheng mellom det undersøkte tiltaket og endringer i trafikksikkerheten.

Metodologisk forskning

For å kunne bestemme hva som skal inngå i en skala for kvalitet, må man ha kunnskap om hva som påvirker kvaliteten på en undersøkelse og hvor store virkninger ulike feilkilder kan ha på resultatene av en undersøkelse. En undersøkelse av god kvalitet kan defineres som en undersøkelse der det er lite sannsynlig at svakheter ved metoden har påvirket resultatene av undersøkelsen.

Med metodologisk forskning menes forskning der formålet er å studere hvordan ulike metodeproblemer og svakheter påvirker, eller kan påvirke, resultatene av en undersøkelse. Slik forskning kan gi et bidrag til grunnlaget for kvalitetsskalaer ved å identifisere hva som skal inngå i skalaene (hvilke sider ved metoden er relevante) og hvor stor betydning de ulike faktorene skal tillegges (betyr metodefeil A som regel mer for resultatene enn metodefeil B?)

Det ble gjort en gjennomgang av metodologisk forskning om studier av virkninger av trafikksikkerhetstiltak. Omfanget av denne forskningen varierer en god del mellom de ulike typer undersøkelsesopplegg som benyttes; mer er derfor kjent om mulige feilkilder ved noen opplegg enn ved andre. Resultatene var vanskelige å tolke. Det viste seg at selv velkjente og relativt godt utforskede feilkilder som manglende kontroll for regresjon mot gjennomsnittet slett ikke alltid påvirker resultatene av en undersøkelse nevneverdig. Når manglende kontroll for regresjon mot gjennomsnittet har betydning, viser det seg, noe overraskende, at feilen kan gå i begge retninger. Det har vært nærmest opplest og vedtatt at manglende kontroll for regresjon mot gjennomsnittet alltid og uten unntak fører til at tiltakets virkning overvurderes betydelig. Slik er det ikke. Bildet er dessverre langt mer uklart.

Gjennomgangen av metodologisk forskning ga følgelig ikke noe brukbart grunnlag for å tilordne vekter til ulike poster i en skala for kvalitet.

En skala for kvalitet på undersøkelser om virkninger av trafikksikkerhetstiltak

Til tross for at forsøkene på å etablere en forskningsmessig begrunnelse for en skala for kvalitet på undersøkelser om virkninger av trafikksikkerhetstiltak i det store og hele må betegnes som mislykkede, er en slik skala likevel foreslått i rapporten. I likhet med enhver annen skala som har vært utviklet, har denne skalaen et betydelig element av vilkårlighet. Dette synes, som rapporten viser, foreløpig å være umulig å unngå. Det valg man står overfor, er derfor enten å konkludere med at kvalitet er noe som ikke kan måles på en god nok måte, eller å måle kvalitet med en skala der de enkelte elementer ikke fullt ut kan begrunnes med henvisning til veletablert kunnskap.

Skalaen består av to deler. Den ene delen er felles for alle typer undersøkelsesopplegg. Den andre delen er skreddersydd til hver type undersøkelsesopplegg. Skalaen er normert slik at en fullkommen undersøkelse scorer 1, en helt verdiløs undersøkelse scorer 0. De ulike typene undersøkelsesopplegg er ikke innbyrdes rangordnet; en god undersøkelse kan følgelig score 1 uansett hvilket opplegg den har benyttet. Den felles delen av skalaen teller 50 %; den del som er spesifikk for hver type undersøkelsesopplegg teller 50 %. Ulike vekter er tilordnet de ulike poster som inngår i skalaen.

Skalaen bygger på kriterier for intern validitet, det vil si operasjonelle kriterier som angir hvor godt en har oppfylt betingelsene for å trekke slutninger om en årsakssammenheng mellom et tiltak og endringer i trafikksikkerheten. Disse kriteriene er utviklet og anvendt gjennom en rekke tidligere studier. Antallet poster som må sjekkes for å bedømme en undersøkelses kvalitet varierer noe mellom de ulike undersøkelsesoppleggene, og ligger mellom 10 og 20.

Skalaen er testet på 18 undersøkelser. Disse scorer verdier som ligger mellom 0.863 for den beste og 0.131 for den dårligste undersøkelsen. Skalaens reliabilitet og validitet er foreløpig ukjent og bør testes på flere undersøkelser.

Behandling av undersøkelsers kvalitet i meta-analyser

Flere tilnærmingsmåter kan tenkes til behandling av undersøkelsers kvalitet i meta-analyser. Tre tilnærmingsmåter betraktes som forsvarlige:

1. Man kan identifisere ulike aspekter ved undersøkelsers kvalitet og bruke en variabel som representerer hvert aspekt som uavhengig variabel i en meta-regresjonsanalyse.
2. Man kan tilordne hver undersøkelse en generell score for kvalitet og bruke denne som uavhengig variabel i meta-regresjonsanalyse.
3. Man kan tilordne hver undersøkelse en generell score for kvalitet mellom 0 og 1 og justere de statistiske vektene som tilordnes hver undersøkelse for undersøkelsens kvalitet. Undersøkelser som scorer nær null vil da få redusert sin vekt tilsvarende.

Alle disse tre tilnærmingsmåtene gir mening og kan forsvares. Eksempler på bruk av dem blir gitt. Tilnærmingsmåte 3 kan begrunnes med at studier av lav kvalitet kan gi mer sprikende resultater enn studier av høy kvalitet, og derfor bør telle mindre.

1 The importance of study quality

1.1 What is study quality and does it matter?

There is an abundance of road safety evaluation studies. The Handbook of Road Safety Measures (Elvik and Vaa 2004) refers to more than 1,500 studies. Hundreds of new studies are reported every year. Trying to keep abreast of new studies is a full-time job. But is it worthwhile to read all new road safety evaluation studies? The answer depends on whether the results of these studies can be trusted or not. Not all road safety evaluation studies present findings that can be trusted. But is it possible to reliably identify good studies and weed out bad studies? One would hope so. Reading many studies is very time-consuming, and most people, who are too short on time to even read the few key studies that they ought to read, will simply not have the time, nor the inclination to go through a detailed and time-consuming exercise for the purpose of rating a study for its quality.

But why bother? Why not simply take the results of all studies at face value? Many would say that any attempt to rate studies by quality is bound to be arbitrary. “Bad studies tend to be those whose results we do not like” (Rosenthal 1991A, page 130). We must ask: can study quality be assessed without knowing study results?

The main reason for trying to assess study quality is that problems related to the varying quality of studies will not go away. If one accepts the idea that study quality is a meaningful concept, and if one accepts the idea that the quality of studies is likely to vary, then one should also accept the need for assessing study quality in a systematic way. A large part of transport research, in particular road safety research, is applied. Policy advice based on bad studies can lead to a waste of resources.

There is no standard definition of the concept of study quality. In fact, the lack of a universally accepted definition is one of the reasons why some researchers think that trying to systematically assess study quality is too difficult and cannot be done in a non-arbitrary way. The implications of such a point of view are too dire to contemplate. Scientific journals try to identify papers that are worthy of publication by means of peer review. If this process involves little more than the personal prejudices of those reviewing a paper, it is worthless as a means of identifying credible research. Students get grades for their work; again if this process is completely arbitrary, how can the most talented students be recruited to research?

While recognising that it is difficult to assess the quality of research, this report will argue that assessing study quality is both necessary and possible. The fact that such assessments have often appeared to be arbitrary does not mean that the task of assessing study quality is impossible; it just means that past efforts have not taken the task seriously enough and have not approached it in a way that adequately reflects its complexity. True, there is no standard definition of study quality. The elaboration of the concept of study quality is itself an important element of a systematic assessment of study quality. As a starting point, the following definition of study quality (Christensen 2003) was adopted:

Study quality denotes the extent to which a study is free of methodological weaknesses that may affect the results.

This reports takes this definition of study quality as the starting point of a research process whose objective is to develop an objective and systematic method for assessing the methodological quality of road safety evaluation studies and summarising the results of study quality assessment in terms of a numerical quality score.

The definition of study quality given above may seem narrow. It disregards aspects of a study that would normally be regarded as part of quality, such as originality and timeliness, conciseness of presentation or depth of argument. This is true, but the definition given above is tailored to the need for assessment of study quality within the framework of meta-analysis, in which summary estimates of effect should ideally be as free of methodological sources of error as possible.

1.2 Problems discussed in this report

The main questions to be discussed in this report are:

1. Is it possible to develop widely accepted operational definitions of the concept of study quality?
2. Is it possible to assign numerical values to aspects of study quality in a non-arbitrary way?
3. Can various aspects of study quality be summarised into an overall numerical quality score?
4. How can the reliability and validity of a scoring system for study quality be assessed?
5. What is the best way of accounting for varying study quality in meta-analysis?

As noted, opinions on these issues are divided. Sander Greenland, for example, offers the following comments on quality scoring of studies (1994, 295, 296):

"Perhaps the most insidious form of subjectivity masquerading as objectivity in meta-analysis is ”quality scoring”. This practice subjectively merges objective information with arbitrary judgments in a manner that can obscure important sources of heterogeneity among study results. … I wholeheartedly condemn quality scores because they conflate objective study properties (such as study design) with subjective and often arbitrary quality weighting schemes."

These comments are a sobering reminder of the complexity of the task of developing a system designed to score studies for quality. It is important to note that Greenland does not condemn the assessment of study quality. On the contrary, he argues that such an assessment is needed, but that it should not be summarised in terms of an overall quality score. "With proper analysis of quality-score *components*, quality scores (and any fix-ups) are superfluous and, without component analyses, quality scores can be misleading" (1994, 300; emphasis in original).

2 A systematic approach to the assessment of study quality

2.1 Elements of a systematic approach

A research process typically involves the following stages:

1. Formulating the research problem
2. Surveying previous studies and the current state of knowledge
3. Developing a theoretical framework for the study
4. Developing a study design
5. Obtaining a sample and collecting data
6. Analysing data
7. Interpreting the results of analysis
8. Writing and presenting a research report

Scientific research is characterised by adherence to rules and procedures; it relies on methodological principles that demarcate it from activities that are not scientific. The assessment of study quality should be approached in the same rigorous manner. More specifically, it involves:

1. A survey of previous research designed to develop systems for assessing study quality
2. A survey of how leading researchers define quality in research and criteria for assessing it
3. A description of the characteristics of an ideal method for assessing study quality
4. The development of a framework for assessing study quality, in terms of a typology of study designs and/or a typology of factors affecting study quality
5. The development of a pilot instrument for assessing study quality and the testing of the instrument
6. The development of a more permanent instrument for assessing study quality and guidance on the use of the instrument
7. Periodic revisions of the instrument for assessing study quality based on experience gained by using it.

Subsequent chapters of the report will go into each of these stages. Before embarking on this research, a few examples will be given of how the quality of road safety evaluation studies can influence their results, in case some readers are doubtful about this. These examples are intended to convince readers of the need for critically assessing the quality of road safety evaluation studies.

2.2 The effect of study quality on study results: a sample of horror stories from road safety evaluation studies

2.2.1 Case 1: Black spot treatment

In a paper published in Accident Analysis and Prevention in 1997 (Elvik 1997, reprinted in Elvik 1999), studies that have evaluated the effects on accidents of black spot treatment were compared with respect to the confounding variables they had controlled for. The studies were classified according to whether or not they controlled for the following potentially confounding factors in before-and-after studies of black spot treatment:

1. Regression-to-the-mean
2. Changes in traffic volume
3. Long-term trends in the number of accidents
4. Accident migration, that is the tendency for accidents to “migrate” from treated black spots to other locations.

The assessment of studies according to control for these factors was generous: Studies that claimed to have controlled for any of the confounding factors were treated as having done so, although some studies did not explain in sufficient detail how they had controlled for the confounding factors.

Figure 1 gives a sample of the results of the study. It shows the percentage change in the number of injury accidents attributed to black spot treatment, depending on which confounding factors studies controlled for.

In simple before-and-after studies that did not control for any of the four confounding factors, an impressive accident reduction of 55 % was attributed to black spot treatment. In studies that controlled for regression-to-the-mean, long-term trends and accident migration, the effect attributed to black spot treatment was zero. The more confounding factors a study controlled for, the smaller were the effects attributed to black spot treatment. No study was found that had controlled for all the four confounding factors listed above.

Now, some readers might wonder how we can know that a potentially confounding factor actually did confound a study. The answer is simple. If the effect attributed to the road safety measure differs depending on whether or not the potentially confounding factor was controlled for, then it does in fact confound study results. Potentially confounding factors do not, of course, always actually confound the results of a study. If, for example, there are no long-term trends in accidents, then this factor cannot confound. The point is that it is not possible to know whether a potentially confounding factor actually confounds a study unless we control for it. The fact that a certain factor is potentially confounding is, in other words, a sufficient condition for trying to control for it.

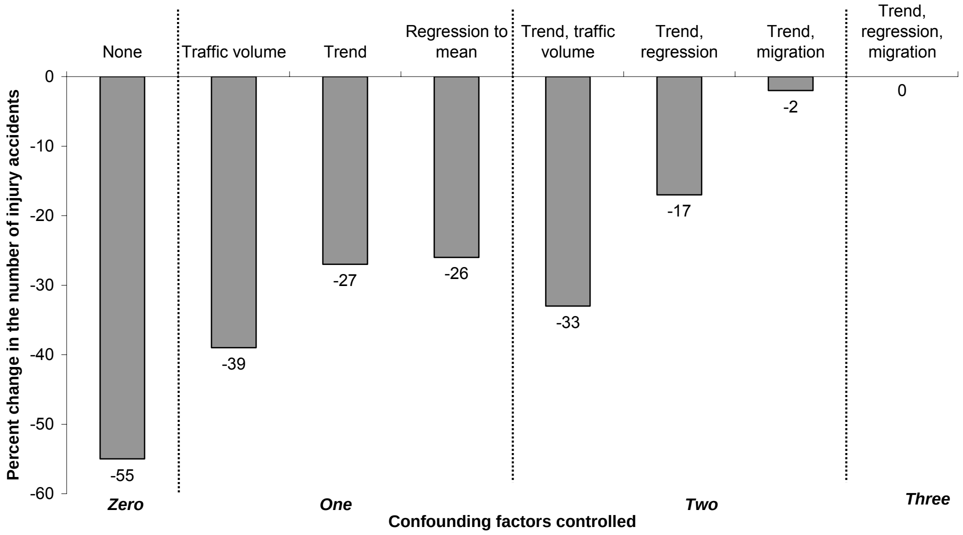

Figure 1: The importance of confounding factors in before-and-after studies of black spot treatment. Source: Elvik 1997

Consider the pattern found in Figure 1. It has been claimed 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 "black spots" demonstrating high returns from relatively low cost measures have been reported worldwide." (quoted from Elvik 1997). Is this claim justified? Take a look at Figure 1 and judge for yourself.

2.2.2 Case 2: Evaluation of road safety measures in Norway

The second example is based on a number of evaluations of road safety measures in Norway (Elvik 2002A). Figure 2 presents the key findings of this study.

Nine different road safety measures were evaluated, all by means of before-and-after studies. For some of the measures, more than one evaluation study has been reported. Each study controlled for regression-to-the-mean and general trends in a larger area in which study sites were located. Controls for these confounding factors were introduced in such a way, that it was easy to remove them, thus producing the results a naïve before-and-after study would have yielded.

As can be seen from Figure 2, the effects attributed to the nine road safety measures were almost always greater, in some cases substantially greater, when no confounding factors were controlled for, than when the effects of the confounding factors were removed. On the average, the uncontrolled estimate of effect was an accident reduction of 31 %. The mean of the controlled estimates of effect was an accident reduction of 19 %.

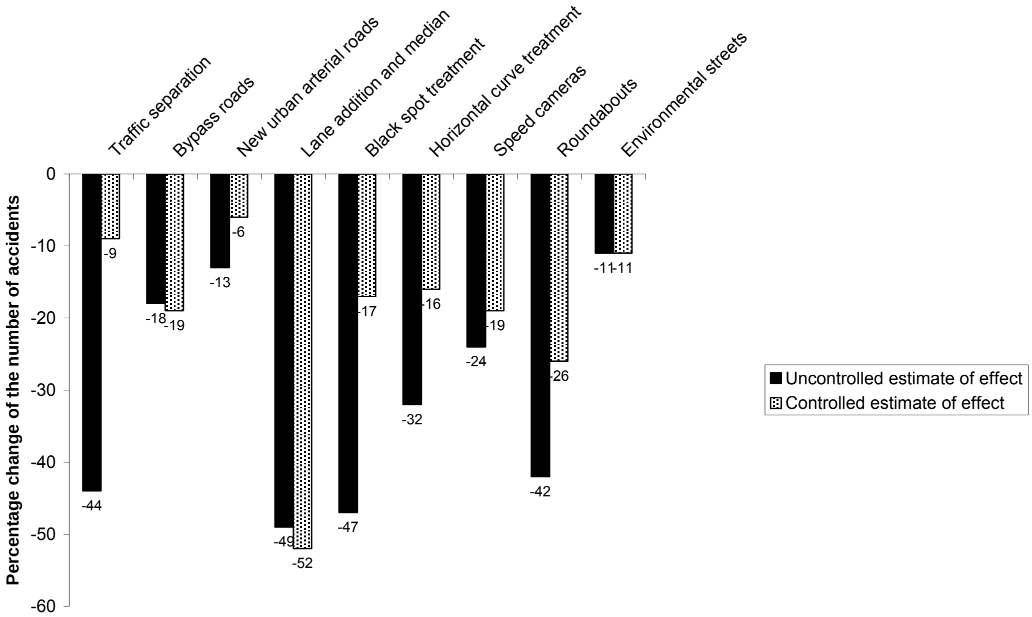

Figure 2: Comparison of controlled and uncontrolled estimates of the effects of nine road safety measures evaluated in Norway. Based on Elvik 2002A.

Figure 2 shows that the effect of confounding factors is sometimes very strong, much greater than the effect of the road safety measure being evaluated. Moreover, it shows that the effects of confounding factors do not always go in the same direction. While one would often expect lack of control for confounding factors to be associated with an overestimation of the effect of a road safety measure, this is not always the case.

2.2.3 Case 3: The effects of technical inspections of heavy vehicles

The third case illustration refers to an evaluation of the effects on road safety of technical inspections of heavy vehicles in Norway (Elvik 2002B).

The study employed multiple regression to model year-to-year changes in accident rate for trucks and buses in Norway. The principal variable of interest was the number of technical inspections per vehicle per year. The study did, however, control for long-term trend, the number of new drivers recruited each year, and the business cycle (measured as percentage annual growth of the gross domestic product of Norway).

Figure 3 shows the effect attributed to technical inspections, depending on which confounding variables were controlled for. The effect attributed to technical inspections was reduced from a 12 % reduction of accident rate to a 7 % reduction of accident rate, as the number of confounding variables controlled for increased from zero to three. This trend is worrying. Would any effect of technical inspections remain if, say, ten confounding variables had been controlled for? Unfortunately, the study did not allow for controlling for more than three confounding variables. Relevant data on other potentially confounding variables were not available. Moreover, the study relied on just 12 data points, making it difficult to control for more than three or four variables.

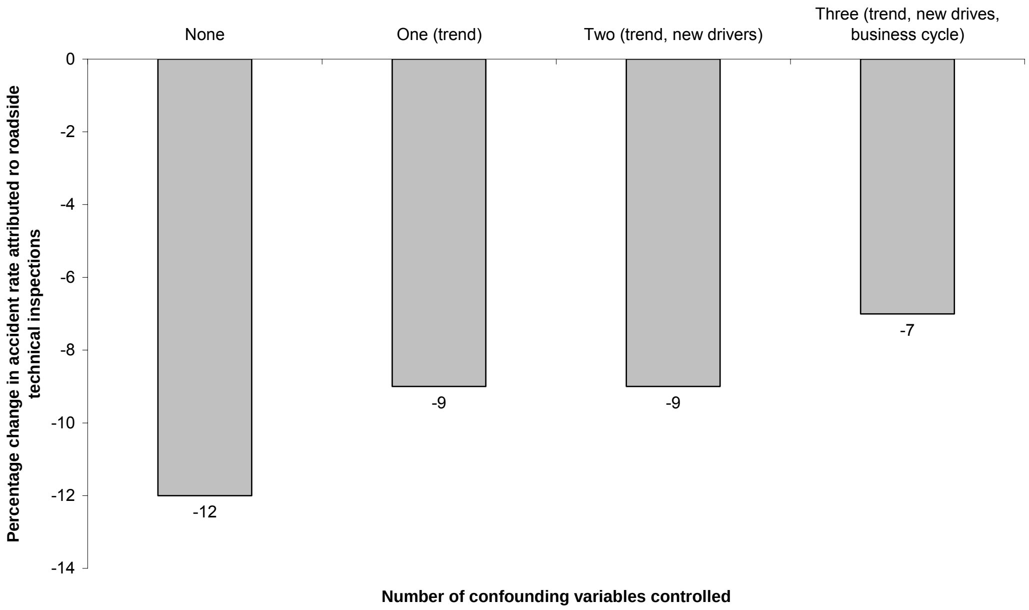

Figure 3: Effects attributed to technical inspections of heavy vehicles in Norway, depending on the number of confounding variables controlled for in the analyses. Based on Elvik 2002B.

This study illustrates the problems of achieving high quality in non-experimental road safety evaluation studies. In discussing study findings, the author states (Elvik 2002B, page 758):

"In the first place, the study was made when it was long time overdue. The National Highway Agency in Norway had carried out an extensive program of technical inspections of heavy vehicles for more than ten years, before any study was made to determine the effects on safety of these inspections. That study (Elvik 1996) was a very simple one, which has been updated and refined in this paper. The fact that it took so long before an evaluation study was commissioned means that the study had to rely on the data that happened to be available. It is, for example, impossible by now to obtain reliable data about driver behaviour in the 1980s and early 1990s when inspections were stepped up.

In the second place, the study is based on a small sample with limited variation. The effective sample size is twelve, that is twelve years of data. Increasing this sample size, for example by using data referring to months, rather than years, or by using data referring to each county in Norway, rather than the whole country, was not possible. Besides, even if this had been possible, it is highly likely that it would have lead to a loss of statistical power, by increasing the contribution of randomness in the counts of heavy vehicles involved in accidents.

In the third place, the study is limited to injury accidents, which are known to be incompletely reported in official accident statistics (Elvik and Mysen 1999). The insurance companies keep statistics of property-damage-only accidents, but these *statistics go back only to 1991. It is impossible to know if the reporting level for injury accidents has changed in the study period, since an accident record known to be complete does not exist. Indeed, if such a record existed, incomplete reporting would cease to be a problem.*

In the fourth place, the study did not control for very many confounding variables. The possibility of controlling for confounding variables was severely circumscribed by the small size of the sample and the limited availability of data going back to 1985 or 1986.

In the fifth place, the study did not uncover the causal mechanism through which technical inspections affect accident rate. There are, for example, no data to show whether technical inspections improve the technical condition of vehicles, or whether this in turn reduces accident rate.

Can the results of a study afflicted by such weaknesses be trusted at all? Or are the results merely the product of imperfect data, analysed by means of imperfect techniques?"

For the moment, this question will be left unanswered, inviting readers to reflect on it.

2.2.4 Case 4: Claiming to control for confounding factors that were actually not controlled for

In a paper published in Accident Analysis and Prevention in 1997, Ogden (1997) evaluated the safety effects of paved shoulders on rural highways. He used a matched pair design, intended to model as closely as possible a true experimental design. 36 sites at which shoulders had been paved were matched with 36 comparison sites that retained unpaved shoulders. Ogden describes the matching procedure in the following terms (1997, page 356):

"Control sections were generally adjacent or very close to the treatment site. The main criteria were that they were similar in design standard, alignment, terrain, roadside conditions, traffic flow etc. Importantly, they were all on the same highway, so that the traffic volume and road-user characteristics would be the same for both control and treatment sites."

The number of accidents was reduced from 73 to 44 for the treated sites. At comparison sites, the number of accidents increased from 58 to 61. Used the odds ratio as a measure of effect, this indicates that paving shoulders was associated with an accident reduction of 43 % (44/73)/(61/58). This estimate of effect does not control for regression to the mean, which could be a potentially confounding factor in this study, despite the matching procedure applied in selecting comparison sites. Traffic volume was the same at comparison sites as at the treated sites; despite this the number of accident in the before-period was 73 at treated sites and only 58 at comparison sites. In an appendix to the paper, Ogden (1997, page 362) claims to have controlled for regression-to-the-mean using "a sufficiently large sample to enable estimation of the parameters" (needed to apply the empirical Bayes method). He states that regression to the mean was estimated to 4 %, resulting in an adjusted estimate of effect of 41 % accident reduction.

It is, however, possible to control for regression to the mean by making use of the data presented in the study. Ogden presents the recorded number of accidents before and after treatment for each site, both the treated sites and the comparison sites. In Figure 4, these data have been plotted and regression lines fitted to them.

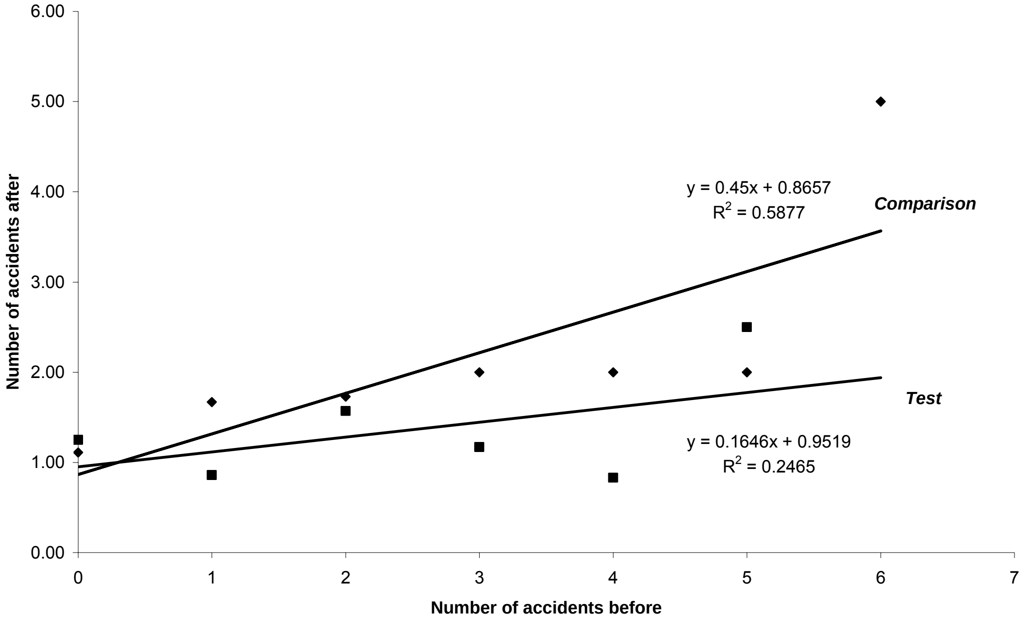

Figure 4: Regression to the mean in matched pairs of sites studied by Ogden (1997).

In the comparison group, 9 sites had 0 accidents in the before-period. The mean number of accidents at these sites during the after period was 1.11. A regression line has been fitted to the data points for the comparison group. This is the steepest of the two lines shown in Figure 4. It is a regression of the mean number of accidents in the after-period as a function of the number of accidents recorded (per site) in the before-period.

If accidents at treated sites had regressed to the mean at the same rate as observed in the comparison group, one can estimate, using the fitted line in Figure 4, a reduction from 73 accidents to 64.2 accidents, which is 12 %, rather more than the 4 % estimated by Ogden. Moreover, the matched comparison group used by Ogden is entirely too small to reliably capture long-term trends in the number of accidents. The mean year for paving shoulders was 1987. The period covered by the study was 1983-1991. If the years 1983-1986 are taken as representative of the before-period, and the years 1988-1991 are taken as representative of the after-period, the mean number of road accident fatalities in Victoria, Australia, declined from 668 per year to 632 per year. This leads to the following estimate of the expected number of accidents in the after-period: 64.2 × (632/668) = 60.7. The recorded number of accidents in the after-period was 44. The effect of paved shoulders on safety was an accident reduction of 28 %.

This example shows that studies claiming to control for certain confounding factors cannot always be trusted to do so. A careful reading of a study, possibly including a re-analysis of the data it presents, may sometimes be needed in order to ascertain whether the control for confounding factors was adequate or not. Unfortunately, most studies do not report the data as completely as Ogden did. The assessment of study quality then has to rely on the study report. If a report paints too rosy a picture of how a study was actually conducted, the study may be more favourably assessed than it deserves. According to a study by Huwiler-Müntener, Jüni, Junker and Egger (2002), there is a relationship between the quality of reporting and the quality of a study. The relationship is, however, not perfect. A study may have controlled for a source of error even if this is not reported. Conversely, studies may on rare occasions report that a certain source of error was controlled for, when in fact this was not done, or – more likely – done inappropriately (as in the study reported by Ogden). Hence, the completeness and honesty of study reports is one of the factors that limits the possibilities of accurately assessing study quality. However, as a general rule one should not give studies the benefit of doubt. If a study does not state that it controlled for a certain confounding factor, one is almost always correct in assuming that this was not done.

2.3 An ideal quality scoring system

Is it possible to outline what an ideal system for scoring studies for quality would look like? To fix ideas, an attempt has been made to describe exhaustively what an ideal formal quality scoring system would look like – what would be the desirable characteristics of such a system. Table 1 presents an attempt to describe an ideal quality scoring system. Ten characteristics are listed. A brief comment on these characteristics will be given.

A quality scoring system should be applicable to any study design (point 1). This is important, because a variety of study designs are used in road safety evaluation research. It is important to score studies employing different designs according to the same system. Ideally speaking, a scoring system should rank study designs from the best to the poorest. This objective is, however, difficult to realise, because studies employing a certain design may vary in quality. A good before-and-after study can be better than a poor multivariate analysis. Besides, developing an exhaustive list of study designs, in which the categories are mutually exclusive (to prevent the possibility that a given study can be put in two or more categories), is difficult. The problem is that there are so many variants of study designs, that a list of them quickly becomes unwieldy.

A quality scoring system should be comprehensive, which means that it should include everything that is generally recognised as an aspect of study quality (point 2). Again, this is an attractive ideal, but it quickly leads to problems. The chief problem is that there are so many aspects of study quality that are relevant, that a list of them all easily gets too long to be practical. Shadish, Cook and Campbell (2002) list 37 generic threats to validity. Not all of these threats will be relevant in all studies. It is nevertheless the case, as will become apparent in the next chapter, that checklists used in formal quality scoring systems can easily get very long. In short, to be comprehensive it may be necessary to be very detailed.

A quality scoring system should produce an overall quality score (point 3), on a scale that has a bounded range (point 4). This could be regarded as strong reductionism. Yet, if we maintain that it makes sense to speak about study quality as a general concept, then we should strive to measure that concept at the same level of generality as it is used when we discuss it in abstract terms. An objection that immediately comes to mind is this: How can, say, scores assigned to 10-20 items that represent different aspects of study quality be aggregated into an overall score in a non-arbitrary way? Some researchers, notably Greenland (1994), argue that this is impossible. An overall quality score is bound to be arbitrary, and, by the same token, uninformative.

Due to the rich variety of study designs found, it may be necessary to introduce optional items in a quality scoring system, although this is not desirable (point 5). To preserve consistency in scoring otherwise identical studies, it is suggested that optional items (i.e. item that are used when appropriate, but not for all studies) should not count towards the overall quality score. It is, however, better to avoid introducing optional items.

Table 1: Characteristics of an ideal formal quality scoring system

A quality scoring system should be/have	Explanation of the criteria
1: Exhaustive with respect to study designs	A quality scoring system ought to be applicable to the whole range of study designs used in a subject area, not just to a particular design. Moreover, it is desirable to be able to rank different study designs according to overall validity.
2: Comprehensive	A quality scoring system ought to include all factors that may affect study validity, that is all aspects of study quality recognised as such by the scientific community.
3: Produce an overall quality score	It should be possible to aggregate the scores assigned to each item in a quality scoring system into an overall quality score, by aggregating the scores assigned to each item.
4: Have a bounded range	There should be fixed lower and upper boundaries for the number of points assigned to studies. If there is a bounded range, a relative quality score (that is actual score/maximum possible score) can be assigned to each study.
5: Independent of the number of items scored	A quality scoring system may contain some items that are optional, that is used only for studies employing a specific study design. These optional items should not count in determining the overall quality score of a study for which they were not relevant.
6: Explicit	There should be clear and explicit rules for assigning scores to each item of the system. These rules should be easy to apply.
7: Reliable	Different individuals scoring the same study should get the same result. Rules for scoring studies should not leave a large room for personal judgement, which may differ greatly between individuals.
8: Sensitive	A quality scoring system should assign different quality scores to studies that differ in quality. An item for which all studies score the same value is insensitive and should not be retained.
9: Independent of the results of each study	It should be possible to score a study for quality without knowing anything about the results of the study. Quality scoring should rely on study methodology only.
10: Independent of the results of other studies in the area	A study whose results are different from those of other studies about the same subject should not be rated as lower in quality simply because of this fact, presuming everything else is equal.

Arbitrariness is perhaps inevitable at some stage of the process of scoring studies for quality, but it can be reduced by making the details of scoring explicit and by showing that the system is reliable (points 6 and 7). A quality scoring system is reliable if different individuals using the system assign the same scores when scoring the same studies (inter-rater reliability). Quality scoring systems that have acceptable reliability exist. An acceptable level of inter-rater agreement is at least around 0.70-0.80 (70-80 % of scores are identical).

Sensitivity (point 8) is an important requirement of a quality scoring system. Simply put, a system is sensitive if poor studies get low scores and good studies get high scores. If all studies get the same score, the system is insensitive. Sensitivity depends on at least two characteristics of a quality scoring system: (1) The number of items scored, and (2) The number of categories used to score each item. A system that has ten items each ranging from 1 to 10, can take on values from 10 to 100, whereas a system with three items, each scored from 1 to 3, only can take on values from 3 to 9. Adding items, and adding categories for each item can, however, reduce reliability. Hence, there may be a trade-off between sensitivity and reliability. On the other hand, fine-graded scores that approximate a continuous scale in principle permit a more precise assessment of study quality than coarser scales.

How about the results of a study? Should they in any way affect the score assigned to it? Ideally speaking not (points 9 and 10). It is sometimes tempting to discount studies whose results are at odds with all other studies that have been reported. Would this be appropriate? In some cases, it would. This might be the case in situations in which there are many studies of at least fair quality. All these studies have, within the bounds of randomness, reported identical findings. Moreover, these findings can be accounted for in theoretical terms. Then a study appears whose findings contradict all previous studies. Surely, most of us would find the results of such a study harder to believe than the results of all the other studies.

On the other hand, one should be wary about canonising theory. Studies that refute what appears to be a well-established theory may set in motion “scientific revolutions” that lead to important new insights. Scientists proceed cautiously, and science as an institution is conservative, despite the fact that its mission is to create new knowledge. One should, however, not dismiss new and contradictory findings simply because they are new and contradictory.

3 A review of scales for study quality

3.1 What can be learnt from existing quality scales?

A survey has been made of scales that have been proposed for the purpose of formally assessing study quality. The survey was mainly based on a paper by Jüni et al (1999), which listed 25 formal quality scales for clinical trials. Jüni et al (1999) found that these 25 scales resulted in very different relative quality scores for the same set of studies, and very different estimates of the relationship between study quality and summary estimates of effect in meta-analysis. While some of the scales indicated that bad studies were associated with large estimates of effect, other scales indicated exactly the opposite. The study reported by Jüni et al (1999) casts serious doubt on the validity and reliability of scales for study quality. If these scales reliably measure the same phenomenon, one would not expect their application in meta-analysis to give so inconsistent results as reported by Jüni et al. Their study shows that scales that have been proposed for assessing study quality need to be examined critically. The objective of this chapter is to shed light on the following questions:

1. Do scales designed to measure study quality formally define the concept of study quality? Are the definitions of study quality underlying different scales consistent?
2. Which aspects of study quality are covered by the scales that have been developed? Do all scales cover the same aspects of study quality?
3. Which aspects of study quality are most frequently addressed by scales designed to assess study quality?
4. How are summary scores for study quality derived according to the scales?
5. How and to what extent has the reliability of the different scales been tested? Is there evidence that some scales are more reliable than other scales?
6. How and to what extent has the validity of the different scales been tested? Is there evidence that some scales are more valid than other scales?
7. Is there evidence that the scales have discriminative power, i.e. does the quality scores assigned to studies vary or do all studies score more or less the same for quality?
8. Can elements of the scales be used to develop a formal quality scoring scale for road safety evaluation studies?

Following this review, a pilot quality scoring system that has been developed for road safety evaluation studies will be presented in Chapter 5.

3.2 Main findings of review

3.2.1 Scales that have been reviewed

Table 2 lists the scales for assessing study quality that have been reviewed. Scales are listed chronologically. For each scale, the country of origin as well as the intended area of application are stated. The review is limited to quality scales, i.e. assessment tools that are intended to produce a numerical score for study quality. Most quality assessment tools that have been reported are checklists. A checklist is a list of study characteristics that are ticked off by answering yes or no. Although a rudimentary scale can be formed by counting the number of items checked by answering yes, most checklists are not intended to be used as numerical scoring instruments. Checklists have therefore not been included in this review.

The quality scales included in this review are just a few of those that have been proposed. Deeks et al. (2003) identified a total of 194 study quality assessment tools. They reviewed these quality assessment tools, concluding that: “Most were poorly developed with scant attention paid to principles of scale development.” They concluded that only a minority of the quality assessment tools were suitable for assessing non-randomised trials.

In the field of road safety, almost all evaluation studies are non-randomised (Elvik 1998, Wentz et al. 2001). It is therefore of particular interest to identify quality scales that have been developed for non-experimental study designs. A total of 35 quality scales are listed in Table 2. These scales have been developed during the period from 1981 to 2003. Most of the scales have been developed in the United States, Great Britain, Norway or the Netherlands. Most of the scales have been developed within medicine. Rather few scales are intended to be applied to all study designs; ten of the scales have been developed for use in assessing randomised controlled trials only. Table 3 is a cross-tabulation of intended area of application versus study design covered by the scales.

Nearly all of the scales, 28 out of 35 have been developed within medicine. Only eight of these scales can be applied to any study design. Most of the medical scales have been developed for assessing randomised controlled trials (experiments) or other controlled trials (not necessarily involving randomisation). Only three scales have been found that have been developed specifically for use in assessing road safety evaluation studies. It is, by the way, not entirely clear if all the scales designed for “controlled trials” are intended for non-experimental studies. In some cases, it is likely that the term “controlled trials” is shorthand for “randomised controlled trials”.

Table 2: A list of scales for assessing study quality

Authors	Year	Country	Intended area of application	Study designs scale applies to
Chalmers et al	1981	United States	Medicine	Randomised controlled trials
Andrew	1984	Norway	Medicine	Controlled clinical trials
Evans et al	1985	Great Britain	Medicine	Randomised controlled trials
Barley	1988	United States	Education	All study designs
Poynard	1988	France	Medicine	Randomised controlled trials
Reisch et al	1989	United States	Medicine	Controlled clinical trials
Gøtsche	1989	Denmark	Medicine	Randomised controlled trials
Gibbs	1989	United States	Psychology	All study designs
Lösel and Köferl	1989	Germany	Criminology	All study designs
Chalmers et al	1990	United States	Medicine	Controlled clinical trials
Imperiale et al	1990	United States	Medicine	Controlled clinical trials
Spitzer et al	1990	United States	Medicine	All study designs
Ter Riet et al	1990	Netherlands	Medicine	All study designs
Kleinen et al	1991	Netherlands	Medicine	All study designs
Koes et al	1991	Netherlands	Medicine	Randomised controlled trials
Levine	1991	United States	Medicine	All study designs
Beckerman et al	1992	Netherlands	Medicine	Randomised controlled trials
Detsky et al	1992	United States	Medicine	Randomised controlled trials
Nurmohamed	1992	Netherlands	Medicine	Controlled clinical trials
Onghena et al	1992	Belgium	Medicine	Controlled clinical trials
Smith et al	1992	Canada	Medicine	Controlled clinical trials
Friedenreich et al	1994	France	Medicine	Case-control studies
Goodman et al	1994	United States	Medicine	All study designs
Margetts et al	1995	Great Britain	Medicine	Case-control studies
Sindhu et al	1997	Great Britain	Medicine	Randomised controlled trials
Downs and Black	1998	Great Britain	Medicine	All study designs
Moher et al	1998	Canada	Medicine	Randomised controlled trials
Elvik	1999	Norway	Road safety	All study designs
Shannon et al	1999	Canada	Workplace safety	All study designs
Greer et al	2000	United States	Medicine	All study designs
Zaza et al	2000	United States	Medicine	All study designs
Elvik	2001	Norway	Road safety	All study designs
Balk et al	2002	United States	Medicine	Randomised controlled trials
Egan et al	2003	Great Britain	Road safety	All study designs
Slim et al	2003	France	Medicine	Non-randomised designs

Table 3: Intended area of application and study designs covered by scales for assessing study quality

	Study designs to which scale can be applied				
Area of application	**Randomised trials**	**Controlled trials**	**Case-control studies**	**All study designs**	**Total**
Medicine	10	8	2	8	28
Education				1	1
Psychology				1	1
Criminology				1	1
Occupational safety				1	1
Road safety				3	3
Total	10	8	2	15	35

3.2.2 The concept of study quality underlying the scales – items covered

Few, if any, of the papers that present scales intended for assessing study quality provide a theoretical, or general definition of study quality. It is fair to say that study quality is implicitly defined in nearly all cases to mean: “quality is what this scale measures”.

All scales contain more than one item, recognising the fact that there are many aspects of study quality. The mean number of items on the 35 scales is 16.9. Some scales have a range from 0 to 100 points, thus expressing the quality score obtained by a study as a percentage of the maximum score. The minimum number of items is 3; the minimum number of points assigned to these items is 5. The maximum number of items is 34; the maximum number of points is 170.

There are large differences between scales with respect to the description of the items covered. Although these differences are in some cases just a difference of words, in most of the cases the differences reflect different concepts. In order to capture the richness of these concepts in as great detail as possible, a total of 158 variables identified by the scales have been coded. These variables have been classified into the following groups:

1. Quality of study report, 12 variables.
2. Description of study context, 4 variables
3. The development of a theoretical basis for a study, 16 variables
4. Choice and development of study design, 6 variables
5. Description of the treatment whose effects is evaluated, 10 variables
6. Procedures for data collection and description of data collected, 20 variables
7. Presentation of detailed data in study report, 3 variables
8. Method for sampling study subjects, 11 variables

9. Control for confounding factors by means of an experimental study design, 11 variables
10. Control for confounding by means of statistical analysis, 16 variables
11. Quality of statistical analysis in general, 18 variables
12. Miscellaneous other items, 31 variables

Table 4 lists the items that have been classified in each group. The table also states how many of the 35 quality scoring scales that include a certain item. Unfortunately, the table covers many pages.

The first observation that can be made, is that there is a very great diversity with respect to which study characteristics that are considered to be aspects of study quality. A second observation, is that many of the items of study quality listed in Table 4 appear to have been developed for use in a single study only. With few exceptions, criteria of study quality appear to be domain-specific, in some cases even specific to a particular study design in a particular domain (e.g, randomised controlled trials in heart surgery). A third observation, is that there are just a few items that have been proposed by a majority of the 35 scales that have been studied. Such items include description of the sampling method used, check of pre-trial equivalence of treatment and control groups, and whether an appropriate statistical analysis has been performed. The latter criterion is often vaguely stated and in need of further elaboration in many of the scales that include it.

Table 4: Items identified in 35 scales that have been developed for assessing study quality

Description of items designed to assess study quality	Number of scales
Category 1. Quality of study report (12 variables)	
Q1. Accuracy of the title of the report	2
Q2. Helpfulness of the abstract of the report	2
Q3. Conciseness of text	2
Q4. Organisation of manuscript (order of sections, subsections, etc)	1
Q5. Style of presentation	1
Q6. Clear-cut sections of manuscript	3
Q7. Level of detail of description of study	3
Q8. References correct	1
Q9. Type of publication (book, report, article, etc)	1
Q10. Description of study background	1
Q11. Description of aim of study	10
Q12. Study results appropriately reported	7
Category 2. Description of study context (4 variables)	
C1. Description of study setting	2
C2. Description of external review and monitoring of research	1
C3. Description of selection of study sites	3
C4. Appropriateness of staff for study	2

Table 4: Items identified in 35 scales that have been developed for assessing study quality, continued

Description of items designed to assess study quality	**Number of scales**
Category 3. Development of a theoretical basis for a study (16 variables)	
T1. Formulation of explicit study hypotheses	4
T2. Operational definitions of theoretical concepts	2
T3. Definition of outcome variables in study (“endpoints”)	16
T4. Description of causal process by which treatment is effective	2
T5. Clear determination of causal direction	2
T6. Findings of study lend support to theory	2
T7. Inadequate theoretical definitions of concepts to permit operationalisation	1
T8. Mono-operation bias	1
T9. Mono-method bias	1
T10. Hypothesis guessing within experimental settings (“looking at the data first”)	1
T11. Evaluation apprehension	1
T12. Experimenter expectancies	1
T13. Confounding constructs and levels of constructs	1
T14. Interaction of testing and treatments	1
T15. Restricted generalisability across constructs	1
T16. Interaction of selection and treatment	1
Category 4. Choice and development of study design (6 variables)	
D1. Use of methods that are easily replicable	3
D2. Study design chosen (i.e. reasons given for choice of design)	7
D3. Use of a contemporaneous comparison group	1
D4. Method appropriate for study (not further elaborated)	1
D5. Method validated (not further elaborated)	1
D6. Careful planning of study (not further elaborated)	4
Category 5. Description of the treatment whose effects is evaluated (10 variables)	
M1. Definition of treatment exposure	4
M2. Measurement of treatment exposure	8
M3. Reliability and validity of exposure measures	1
M4. Independent criteria of validity and reliability of treatment exposure	1
M5. Description of treatment implemented	16
M6. Treatment procedure	10
M7. Additional treatments (beside main treatment)	8
M8. Stability of treatment goals	1
M9. Description of placebo treatment	7
M10. Side effects of treatment discussed	9

Table 4: Items identified in 35 scales that have been developed for assessing study quality, continued

Description of items designed to assess study quality	Number of scales
Category 6. Procedures for data collection and description of data collected (20 variables)	
P1. Informed consent obtained from study participants	4
P2. Response rate (for questionnaires)	1
P3. Source of data regarding compliance with treatment	1
P4. Use of incident cases only (i.e. not those diagnosed before study started)	3
P5. Keeping of log of rejected cases	3
P6. Keeping of log of participants withdrawn from study	6
P7. Description of reasons for withdrawal from study	5
P8. Withdrawal from study below 10 percent	11
P9. Description of portion size (in study of nutrition)	1
P10. Number of quantitative methods used to describe nutrition	1
P11. Types of quantitative methods used to describe nutrition	1
P12. Frequency estimates for nutrition intake	1
P13. Number of food items specified	1
P14. Types of food items specified	1
P15. Quality of data on cooking of food items	1
P16. Manual transcription of food diary	1
P17. Season of food intake	1
P18. Type of food table used	2
P19. Description of data collection procedure	7
P20. Quality check of interview regarding food habits	1
Category 7. Presentation of detailed data in study report (3 variables)	
A1. All raw data reproduced in study report	3
A2. Data on statistical scores presented	1
A3. Data given in detail for all figures and tables in study report	2
Category 8. Method for sampling study subjects – sample size etc (11 variables)	
S1. Study population adequately described	2
S2. Sampling frame stated	1
S3. Sampling criteria/method stated (random or non-random)	22
S4. Sample size stated	11
S5. Low statistical power	1
S6. Sampling unit used (individuals or clusters)	3
S7. Representativeness of sample assessed	7
S8. Control sample obtained (in addition to treated sample)	2
S9. Description of sample (by basic demographic characteristics)	2
S10. Specification of exclusion criteria	10
S11. Use of power calculation to determine sample size	10

Table 4: Items identified in 35 scales that have been developed for assessing study quality, continued

Description of items designed to assess study quality	Number of scales
Category 9. Control for confounding by means of study design (use of experimental design) (11 variables)	
E1. Use of a control group in study	7
E2. Random allocation of study subjects to treatment and control conditions	12
E3. Allocation of study subject concealed (not further elaborated)	1
E4. Patients blinded to experimental condition	9
E5. Physicians blinded to experimental condition	13
E6. Physicians blinded to results of ongoing treatment (during study)	5
E7. Test of randomisation (absence of systematic differences between groups)	13
E8. Test of blinding to experimental condition	7
E9. Test of compliance with experimental protocol	7
E10. Blinding of statistician to experimental condition	6
E11. Check of pre-trial equivalence of treated and control groups	20
Category 10. Control for confounding by means of statistical analysis (16 variables)	
N1. Known confounders controlled for (not further elaborated)	11
N2. Long term trends controlled for	3
N3. Specific events controlled for	2
N4. Regression-to-the-mean controlled for	3
N5. Instrumentation effects controlled for	2
N6. Dose-response pattern analysed statistically	2
N7. Specificity of effects to target group determined statistically	2
N8. Subject reactivity to treatment tested	3
N9. Teacher reactivity to treatment tested	1
N10. Interaction with selection controlled for	1
N11. Testing effect controlled for	1
N12. Diffusion or imitation of treatment controlled for	2
N13. Compensatory equalisation of treatments controlled for	1
N14. Compensatory rivalry by subjects receiving less desirable treatments	1
N10. Statistical analysis not characterised as data mining	2
N11. Subjects stratified by risk factors	7

Table 4: Items identified in 35 scales that have been developed for assessing study quality, continued

Description of items designed to assess study quality	Number of scales
Category 11. Quality of statistical analyses in general (18 variables)	
G1. Posterior power calculation performed	4
G2. Formal statistical inference (test of significance) performed	5
G3. Appropriate method used for data synthesis (not further elaborated)	5
G4. Shape of distributions (skewness, etc) tested	1
G5. Robustness of mean tested	1
G6. Appropriate statistical analysis performed (not further elaborated)	20
G7. Analysis of proportions performed	1
G8. Analysis of numbers (needed to treat) performed	1
G9. Confidence intervals reported	5
G10. Exact P-values reported	2
G11. Probability of making type II error reported	2
G12. Handling of attrition described	10
G13. Violated assumptions of statistical tests	1
G14. Reliability of measures	1
G15. Reliability of measurement implementation	1
G16. Irrelevancies in the experimental setting	1
G17. Random heterogeneity of respondents	1
G18. Effect magnitude reported	1

Table 4: Items identified in 35 scales that have been developed for assessing study quality, continued

Description of items designed to assess study quality	Number of scales
Category 12. Miscellaneous other items (31 variables)	
O1. Are results of study credible? (not further elaborated)	2
O2. Does study add to knowledge? (not further elaborated)	1
O3. Are known risk factors described? (not further elaborated)	1
O4. Definitions of important concepts provided	2
O5. Is this a pragmatic study? (not further elaborated)	1
O6. Are diagnoses accurate?	4
O7. Are objective criteria used? (not further elaborated)	2
O8. Is analysis of endpoints performed?	4
O9. Observation bias controlled for	1
O10. Selection bias controlled for	3
O11. Results stable over time (external validity)	1
O12. Results stable in space (external validity)	1
O13. Results stable across other contextual variables (than space and time)	1
O14. Form of questionnaire administration (postal, etc)	1
O15. Pain measure used	1
O16. Retrospective analysis carried out (in case-control studies)	1
O17. Dates of study reported	2
O18. Specification of accident severity	2
O19. Questionnaire validated? (pre-tested)	1
O20. Systematic errors in data	1
O21. Are data valid and reliable?	5
O22. Subject sensitization tested	1
O23. Length of after period sufficient	7
O24. Outcome variables relevant	4
O25. Timing of events appropriate	7
O26. Dependent variable reliable	1
O27. Study limits clearly stated	2
O28. Can findings be generalised? (not further elaborated)	6
O29. Publication bias tested for	1
O30. Appropriate conclusions drawn	8
O31. Miscellaneous other items	6

Despite the great diversity of characteristics that have been proposed to describe and rate study quality, it is possible to reduce these characteristics to a few more general concepts. The first two categories, quality of the study report and description of study context, are not aspects of study quality as such – although, as noted above, the assessment of study quality must rely to a great extent on the study report. The third category, developing a theoretical basis for a study, is a relevant aspect of study quality, and will be discussed more extensively in chapter 7. The choice and development of study design (category 4) is an important determinant of study quality, whereas a detailed description of the treatment whose effects is evaluated is less important. The importance of study design in influencing study quality, rests on the fact that choice of design exerts a major influence on how well a study controls for confounding factors (Shadish, Cook and Campbell 2002).

Category 6, procedures for data collection and description of data collected, is in principle an important aspect of study quality. As an example, it is not difficult to think of examples of how incomplete accident reporting in data files used in road safety evaluation studies can produce misleading findings. Non-response in sample surveys can also bias study findings. Even routine measurements that one would think are simple, like measuring the speed of cars passing a certain point on the road, have been found to be very error prone (Ragnøy and Muskaug 2003).

Category 7, presentation of detailed data in study report, is again not an aspect of methodological quality, but an aspect of study presentation. It will there not be further discussed in this report.

Category 8, sampling method and sample size, is a relevant aspect of study quality. Very many road safety evaluation studies rely on convenience samples. Strictly speaking, statistical inference cannot be applied to these studies, as sampling theory applies only to samples that have been obtained by means of a known statistical sampling technique. Practice, however, is very different and inferential techniques, like significance tests, are widely applied in studies relying on non-statistical sampling techniques.

Categories 9 and 10 concern control for confounding factors by means of study design or statistical analysis. This is a very important aspect of study quality – clearly the most important as far as road safety evaluation studies are concerned. Poor control of potentially confounding factors is the most common weakness of these studies and should be emphasised in any system for assessing study quality.

Category 11, quality of statistical analysis is potentially important. To be useful for the purpose of assessing study quality, however, specific errors that can be made in statistical analyses need to be identified. It is not very instructive to employ a vague concept like "appropriate statistical analysis performed", as very many of the quality assessment tools included in Table 4 appear to do.

Finally category 12, is very heterogeneous and most items belonging to this category have only been employed in a few of the 35 scales included in Table 4. Some of the items refer to study findings, which is clearly not an aspect of study quality. Other items are too vague to be useful, like "are data valid and reliable?", which needs to be a lot more specific to be useful.

A preliminary conclusion is that categories 1, 2, 7 and 12 of those identified in Table 4 are of no interest in developing a scale for assessing study quality. Category 5 is of limited interest. The remaining categories, in particular categories 4, 9 and 10, are important and deserve a more careful examination.

3.3 Lessons to be learnt

What can be learnt from this with respect to developing a system for scoring road safety evaluation studies for quality?

Referring to the questions asked at the beginning of this chapter, it is clear that few of the scales refer to any formal definition of study quality. Moreover, the implicit notion of quality underlying the various scales does not appear to reflect any consensus about the meaning of the concept. On the contrary, the diversity of study characteristics that are treated as aspects of study quality is striking.

Very many aspects of study quality are covered by the 35 scales reviewed here. The aspects that are addressed by at least 40 % of the scales (i.e. at least 14 scales) include:

- Definition of outcome variables (“endpoints”)
- Description of treatment whose effects is evaluated
- Sampling method
- Check of pre-trial equivalence
- Appropriate techniques of statistical analysis

All these aspects are relevant for road safety evaluation studies. With respect to the first variable, it should be noted that the dependent variable in road safety evaluation studies can be defined in many ways: the number of accidents, the accident rate (accidents related to some measure of exposure) or a ratio of accident rates – to name but a few.

Description of treatment is relevant, but not so much in assessing the quality of a specific study as in meta-analyses seeking to combine the findings of several studies. In meta-analyses, it is important to make sure that the studies whose findings are pooled did evaluate the same road safety measure.

Sampling method is a relevant aspect of study quality for road safety evaluation studies. As noted above, convenience samples are very often used and few details are provided about how the samples were obtained.

Check of pre-trial equivalence is important in randomised controlled trials to ensure that randomisation was successful and that there are no systematic differences between the treatment group(s) and the control group(s). Since most road safety evaluation studies are non-experimental, one might think that this criterion is of little importance. It is, however, very important. Thus, as an example, regression-to-the-mean will not be controlled for if treated black spots are compared to roads that have a normal level of safety in a before-and-after study. The groups must be equivalent in the sense that the comparison group is subject to the same regression-to-the-mean effect as the treated group, or, alternatively, that regression-to-the-mean can be controlled for statistically. Hauer (1991, 1997) and Hauer, Ng and Papaioannou (1991) have proposed criteria for assessing the pre-treatment equivalence of treatment and comparison groups in before-and-after studies of road safety measures.

The use of appropriate techniques for statistical analyses is obviously an important aspect of study quality, but it needs to be made considerably more specific to become a useful item in a numerical scale for study quality.

The quality scales included in Table 4 all derive the summary score simply by adding the scores for each item. In some scales, all items count equally, in other scales different weights are attached to the different items. Differences in weight reflect the importance of the items; items regarded as important are assigned greater weight than less important items. However, the justification given for assigning different weights to different items is either highly subjective, i.e. it merely shows the opinion of the researchers who developed a scale, or completely missing.

Few of the scales refer to other scales. In fact, nearly all the reviewed scales appear to have been developed without reference to other work in the area. This means that research in this field is not very cumulative and that researchers do not learn from each other. This is very unlike science in general, in which exchange and accumulation of knowledge is a key activity.

For 23 of the 35 scales, reliability is not reported. For the 12 scales for which reliability is reported, the mean reliability score is 0.76. The median score is 0.83. This level of reliability is acceptable and shows that, in principle, reliable scales can be developed.

As far as validity is concerned, nearly all scales, 31 out of 35, fail to discuss the issue at all. 4 scales claim to have assessed validity, but it is not altogether clear that these tests are relevant, as they are not always based on a precise notion of validity. As an example, Downs and Black (1998) tested the “criterion validity” of their scale by correlating the scores assigned to specific studies to similar scores obtained by means of two other scales. This test would be relevant if the other scales could be treated as a “gold standard” for a quality scale. But then, if a scale that can be regarded as a gold standard has been developed, why would researchers want to develop another scale at all? Why not rely on the gold standard, if indeed such a standard makes sense?

Evidence of discriminative power is given for 22 of the 35 scales; not for the other 13. By discriminative power is meant the ability of the scales to discriminate between studies of different quality, i.e. assign low scores to bad studies and high scores to good studies (in terms of the operational definitions of good and bad according to each scale). Converted to a range between 0 (bad studies) and 1 (good studies), the mean difference between maximum and minimum scores was 0.53. The mean value of the maximum scores was 0.79. The range was between 1.00 and 0.55. The mean value of the minimum scores was 0.25. The range was between 0.00 and 0.60. Based on this review, it is concluded that quality scales with sufficient discriminative power can be developed.

Only 3 of the 35 scales were developed for the purpose of assessing the quality of road safety evaluation studies. One of these scales will be reviewed in detail in Chapter 5. One of the other two scales was a simple instrument consisting of nine items. These items address only a few of factors that may influence the quality of road safety evaluation studies.

To summarise, the main lessons learnt from the review of scales for assessing study quality were:

1. 35 scales that have been developed for the purpose of assessing study quality have been reviewed. Most of these scales (28) were developed in medical research. Only 3 scales designed to assess the quality of road safety evaluation studies were identified.
2. The scales reflect widely divergent views concerning what constitutes study quality. A total of 158 variables were coded based on the 35 scales. Many of these variables have nothing to do with study quality.
3. A small set of variables are common to more than 40 % of the scales. All these variables are relevant in assessing the quality of road safety evaluation studies, but some of them need redefinition to be applicable to road safety evaluation studies.
4. Very few of the scales refer to other work in the area of study quality assessment.
5. Reliability is not reported for all scales. For those that report reliability, it is at an acceptable level.
6. Tests of the validity of scales for assessing study quality are almost never performed and some of the few tests that have been reported make little sense.
7. Scales that have acceptable discriminative power can be found, but the scales reviewed differ greatly with respect to discriminative power.
8. Most of the variables included in the scales are irrelevant when assessing the quality of road safety evaluation studies.

In short, the findings of this review are discouraging. They show that researchers have not employed a scientific approach to the task of developing quality scoring systems. On the contrary, nearly all the scales reviewed are ad hoc instruments, reflecting little else than the preconceived notions each researcher has about study quality. Next to nothing useful can be learnt from these scales for the purpose of developing a quality scoring system for road safety evaluation studies.

4 What do the experts think?

4.1 A survey of ten prominent road safety experts

What do leading road safety experts around the world think about study quality? What do they think characterises a good study?

A small survey intended to shed light on this question was conducted in order to gain an impression of what leading road safety experts mean by the notion of study quality. The experts were asked to following four questions:

1. What do you think characterises a high quality road safety evaluation study? Please list up to ten aspects of study quality that you regard as important.
2. What do you think are the most commonly found weaknesses of road safety evaluation studies? Please list up to five flaws in evaluation studies that cast doubt on the validity of their conclusions.
3. Do you think some aspects of study quality are more important than others? Try to indicate the three most important aspects of study quality.
4. Do you think it is possible at all to measure study quality numerically? Please state briefly what you think are the most important arguments for and against trying to measure study quality numerically.

The first question was intended to elicit notions about research quality, that is about the characteristics of a study that are regarded as relevant in judging whether it is good or bad. The second and third questions were designed to investigate which aspects of study quality are regarded as the most and least important. The aspects listed in answers to questions two and three are interpreted as the most important, while any additional aspects listed in answer to question one, but not questions two and three, are interpreted as less important. The fourth question is intended to obtain opinions as to whether a formal, numerical quality scoring system for road safety evaluation studies is regarded as too arbitrary or sufficiently objective to make sense.

No pre-coded answers were provided. The experts had to formulate their own answers. The survey was sent to ten road safety experts. Eight experts answered the questions.

4.2 Results of the survey

A transcript of the questions asked and the answers given to them is found in Table 5. Respondents have been identified as A, B, C, D, E, F, G, and H. As can be seen from the transcript, the answers given varied greatly in terms both of their length and content. Not all researchers listed explicitly the characteristics of a good road safety evaluation study.

Table 5: Results of survey of leading road safety experts about the quality of road safety evaluation studies

Question 1: What do you think characterises a high-quality road safety evaluation study?	
Respondent	**Answer**
A	1 Objectivity and ability
B	1 If a before and after study: B1 Regression effects considered B2 Changes in traffic volumes accounted for B3 General trends accounted for if long-term evaluations B4 Migration effects considered B5 Data quality ascertained 2 If cross-section: B6 What confounding variables are there B7 How does traffic volumes vary and several more I believe before and after is easier to deal with
C	I do not think that I can tell you here something you do not already know. There is one aspect of quality that stems from what I said in answer to question 4 (listed below). If a weak study design comes up with an answer that seems to support other results, I am tempted to give it more weight.
D	D1 Produces an estimate of the safety effect of the measure D2 The estimate is unbiased i.e. controlled for confounding factors D3 The uncertainty of estimate is presented D4 Statistical and other methods are applied in a proper manner D5 The study is based on a theoretical framework D6 The effect is not a black box effect i.e. the study also sheds light on the effect mechanisms: e.g. effect on road user observation making → effect on behaviour → effect on crash occurrence → effect on crash consequences D7 The data is presented (or be made available) in the study in such a way that other researchers can also replicate the analyses D8 Proper reference is made to the work of other authors, when relevant D9 The authors of the study are properly indicated, with contact information
E	A renowned transport research institute should be responsible for the study (university departments do not always produce high quality research), researchers should be familiar with the subject. E1 The study should be subjected to peer review E2 A comprehensive report should be written up (but not an unnecessarily long one), to permit readers to replicate at least some of the analysis E3 Assumptions made, methods chosen and conclusions drawn should be clearly stated E4 Good quality of data, a representative and sufficiently large sample E5 Correct statistical methods are used E6 Regression-to-the-mean is controlled for E7 Changes in exposure and general changes in the number of accident should be accounted for E8 An estimate of the uncertainty of the results should be included E9 A discussion of whether results are "reasonable" or make sense in relation to other studies
F	F1 A comprehensive approach F2 Frank assessment of strengths and weaknesses of the relevant data F3 An unbiased approach – which includes a willingness to find the outcome 'not proven' F4 An appropriate analytical methodology F5 Full account and explanation of the analytical methods used F6 An approach which attempts to convince the reader of the validity of the conclusions, rather than simply presents the results and expects the 'significant' items to be accepted blindly F7 Where assumptions are made, supporting analyses should be made as far as possible to show that they are plausible F8 The minimum possible reliance upon control groups (based upon a personal distrust of controls that may not be widely shared!)

Table 5: Results of survey of leading road safety experts about the quality of road safety evaluation studies

G	G1 Exogenous sample: the set of observations must not itself be influenced by the phenomenon under study. Alternatively: rigorous control for the regression-to-the-mean bias. G2 Randomized experiment or, alternatively, multivariate data analysis. G3 If experimental, assessment of validity. (Needs to be representative of the situations where results are to be applied.) G4 If non-experimental, adequate multivariate method of analysis. Usually, this means (generalised) Poisson modelling or – if large accident counts – properly specified heteroskedastic regression models. G5 Awareness of distinction between random and systematic variation and of the fact that the former is always present in accident counts. A general idea about the size of the random variation or of the methods to assess it. G6 Awareness of the pitfalls of multiple regression analysis (and of simple bivariate comparisons) and on how (if possible) to avoid them. Alternatively: exhaustive discussion of omitted variable bias and of simultaneity (endogeneity) bias. G7 Large data set, i e large enough that random variation does not blur any systematic variation, and that omitted variables are unlikely to seriously distort the picture. Simple before-and-after studies are notoriously in violation of these rules. G8 Maximum transparency of analysis, especially in terms of various “cleansing” processes designed to “remove” “trends”, autocorrelation etc. Raw data are preferable to cooked data, in terms of interpretation and control. G9 Awareness and assessment of intermediate causal factors. G10Could’t think of more items right now!
H	H1 Detailed documentation of methodology H2: Explicit accounting for uncertainty H3: Proper specification of control group H4: Consideration of regression to the mean H5: Proper accounting of traffic volume and other factors that can cause changes in safety H6: How independent is the evaluation? H7: Would the study have been published if the results showed a negative safety effect? H8: What are the consequences to the road authority of finding a negative safety effect? H9: Representativeness and size of evaluation sample H10: Transferability of results
Question 2: What do you think are the most commonly found weaknesses of road safety evaluation studies?	
A	1 Advocacy and partisanship
B	B1 Not having considered what I listed in question 1 B2 Not having counted exposure, just assumed or used weak induced exposure measures B3 Not including sites with zero crashes in analysis of types of layouts
C	Again, you already know all there is to know. One flaw the importance of which came to my attention recently is the ‘assumed functional form’ in multivariate statistical modelling. It is common to use “variable$^{\text{regression parameter}}$’ or e$^{\text{variable}*\text{regression parameter}}$ to represent the effect of a variable. I think that to do so is very limiting. These functional forms cannot show effects that have peaks or valleys. Such models have contributed to the apparently incorrect conclusion that the wider the lane the safer the road. Another (more general) flaw is to attribute cause-effect virtues to multivariate model results. I do not know what are the important conditions that need to be met for a cause-effect interpretation to be plausible. One condition is that the addition of a new variable or parameter no more changes the values of the existing parameters.
D	D1 All confounding factors are not accounted for D2 Data is not presented properly D3 The study deals with accidents/crashes only with even no reference or no accompanying efforts made to study the effect mechanisms (all depends on effects on travel or traffic behaviour, usually) D4 No estimate is given for the safety effect, but rather just tested whether the effect is statistically significant or not D5 Researchers do not take the trouble to find out about earlier studies around the same subject, which results in duplicated efforts and even duplicated errors in experimental design

Table 5: Results of survey of leading road safety experts about the quality of road safety evaluation studies

E	E1 "Hobby researcher" (some people affiliated with universities seem to be) E2 No peer review E3 Research not independent E4 Simple description with no statistical analysis E5 Errors and weaknesses in data or methods
F	F1 Because of the generally good levels of road safety in Western countries, numbers of accidents, casualties etc in the units studied tend to be small, so that it is more difficult to be certain that apparent changes have not arisen by chance. F2 In "real world" studies, the data collected will be influenced by various factors outside the experimenter's control – including most importantly the way the police record accident data. F3 A naïve use of controls. F4 A preference by some analysts for elaboration, i. e. to build unduly complex models that do not show the link between cause and effect clearly.
G	Breaches with items 1, 2, 4, 5 and 7 above
H	H1: Not accounting for regression to the mean H2: No or improper control for changes due to other factors H3: No or improper accounting for traffic volume changes H4: No or improper accounting for uncertainty H5: Not accounting for spillover/migration effects
Question 3: Do you think some aspects of study quality are more important than others?	
A	Yes
B	If small numbers, then regression effects the most important
C	I cannot think of a general answer
D	D1 Produces an estimate of the safety effect of the measure D2 The estimate is unbiased, i. e. controlled for confounding factors D3 The uncertainty of estimate is presented
E	E1 The most important thing is that a well renowned research institute has done the research. E2 And that the study has undergone peer review
F	F1 A comprehensive approach that recognises the various potential confounding influences. F2 An empirical exploratory approach rather than a crude or unquestioning application of standard techniques.
G	Items 1, 5, 7 above (refers to numbers in answer to question 1 above)
H	1, 4, 6 (refers to numbers in answer to question 1 above)
Question 4: Do you think it is possible at all to measure study quality numerically?	
A	Perhaps. If anyone can do it, you can.
B	I think you have previously done that. But to add up all factors into one quality number seems impossible. To work with confidence intervals (or standard deviations) for all results seems good enough. But I see your point, in meta-analysis, it would be nice if we when weight together likelihoods could weight by the quality of the study in some way that include what was not captures in the likelihood. I hope you understand what I mean.
C	By now I have had considerable experience with attempting to interpret published findings. Had I done a review of, say, the safety effect of illumination, I might have concluded that the many studies fall into three of four prototype groups and that the numerical results can be combined quantitatively, at least within each prototype group. But the reviews I have done were about various geometric road features and about truck size and weight. These were more like detective work; trying to extract clues from studies that were simple minded (accident rate, single variable), before-after of various kinds, cross-section tabulations and multivariate models etc. I think, at this time, that the mechanistic combination of quality-weighed results is perhaps not as fruitful as the search for commonality and attempts to identify reasons for diversity.
D	I think that would not be very easy. I would go for a multicriteria analysis.

Table 5: Results of survey of leading road safety experts about the quality of road safety evaluation studies

E	Yes! It ought to be possible, although perfect measurement is of course impossible. However, something “half good” is better than nothing at all. One might assign points to the most important aspects, ranging, for example from –2 to +2. For less important items, scores ranging from –1 to +1 could be used. The overall score might range from –5 to +5. Such a scoring system would perhaps be “conservative”, but its advantages clearly outweigh the drawbacks.
F	I think it should be possible to measure certain aspects of study quality numerically. For example, if a panel of 10 experts were given 10 studies to rank, I imagine a reasonable consensus could be achieved on questions of whether the methods were appropriate, well described etc. It would be much more difficult for questions such as data quality and whether potential confounding factors had been controlled adequately.
G	I think numerical measurements can indeed make a positive contribution, but I am sceptical to the arbitrariness involved in setting the scores and in the method to weigh the scores together. If the study is faulty on one criterion, it may not help if all the other scores are top! This suggests a multiplicative aggregate measure of scores (such as the geometric mean) rather than an additive one (arithmetic mean). Research should be done to arrive at some conclusion here, possibly establishing an international standard. Which will still be arbitrary, but at least all meta-analyses will have the same (arbitrary) benchmark. Under the iron law of evaluation studies, faulty studies are biased, and in a systematic direction. If this law holds, it may therefore not be an optimal strategy to weigh together good and bad studies in an average. This question also needs elucidation.
H	It is possible to do this but I don’t have too much faith in the process. It takes an expert to make a proper judgement on study quality and there are very few of these around and they are all very busy people. Most studies do not report sufficient information to make an informed judgement. This can create a bias since good studies that provide detailed reporting of methodology and find little safety effect are likely to be judged more harshly than bad studies that find high safety effects but hide the important details of the methodology.

Those researchers who did list characteristics of a good road safety evaluation study, mentioned the following:

1. An *estimate of the effect* of the safety measure should be produced (D1).
2. The *uncertainty of the estimate of effect* should be estimated and presented (D3, E8, G5, H2).
3. A *large accident sample* should be used (E4, G7, H9).
4. A *representative sample* should be used (G1, G2, H9)
5. An *explicit theoretical framework* should be developed (D5).
6. The *quality of data* should be *checked* (B5, E4, F2).
7. *Appropriate statistical techniques* of analysis should be used (D4, E5, F4, G4, G6).
8. *Assumptions made* in analysing data should be made *explicit* and should be *plausible* (D7, E3, F4, F7, G6, G8, H1).
9. Various *confounding factors* should be *controlled* (B1, B2, B3, B4, E6, E7, G1, G4, H3, H4, H5)
10. The *causal mechanism* producing effects on safety should be *described* (D6, G9).
11. *Reference* should be made *to other research* (Possibly C, D8, E9).

12. *Interpretation of study findings* should be *honest and objective*, and allow for the possibility that the study is *inconclusive* (Possibly A, E9, F2, F3, H8, H9).
13. The evaluation should be *independent*, and preferably not be made by an institution with a vested interest in the results (Introduction to E, H6)

The other items listed by the respondents are not so easy to interpret (F1: "A comprehensive approach") or fit into a certain category (E1: "The study should be subjected to peer review").

Even in this small sample, it is striking to see the extent to which respondents emphasise different aspects of study quality. While some stress study design and statistical analysis (especially respondent G), others draw attention to the interpretation of study findings (especially respondents E and F). Some respondents list the most important confounding factors a study should control for (respondents B, E, H, and partly G). One lesson to be learnt from this is that there are many aspects of both study design and data analysis that influence the overall quality of a study. A formal quality scoring system ought ideally to include all these aspects. It cannot be reduced to just two or three items. But are all items equally important? From the answers to questions 2 and 3 it would seem that items 1, 2, 3, 6, 7, and 9 on the above list are regarded as the most important.

As to the prospects of developing a numerical score for study quality, which is not too arbitrary, opinions differ. The majority of the respondents appear to accept the idea that specific items of study quality can be scored numerically, but are more hesitant about the possibility of aggregating these scores into an overall quality score. The idea of trying to develop such a score is, however, not rejected outright.

4.3 Study quality assessment for the Highway Safety Manual

The development of a Highway Safety Manual is a major research project in the United States, funded by the Transportation Research Board. As part of this project, syntheses of evidence from a very large number of road safety evaluation studies are developed. To make these syntheses informative, a system for scoring studies according to study quality has been developed, mainly by Ezra Hauer. He was one of the ten leading road safety researchers contacted in order to collect expert opinion on the concept of study quality. The following is a description of this system (Hauer 2007).

The system for assessing study quality in the Highway Safety Manual consists of three main elements:

1. Classification of studies with respect to study design and level of control for potentially confounding factors.
2. Application of corrections for regression-to-the-mean bias and traffic volume bias (see below).
3. Application of method correction factors to adjust the statistical weights assigned to studies in meta-analysis (see below).

A distinction is made between three basic study designs:

1. Before-and-after studies, which includes empirical Bayes studies, simple before-and-after studies, before-and-after studies employing likelihood functions, before-and-after studies with a comparison group, expert panels and meta-analyses.
2. Cross-section studies not employing regression techniques, that is studies that compare sites that have some safety feature to sites that do not have the safety feature, but do not adjust for potentially confounding factors by means of multivariate analysis.
3. Cross-section studies employing multivariate statistical analysis (accident prediction models) to adjust for potentially confounding factors.

For each study design, a distinction is made between five levels of study quality with respect to control for confounding factors. Table 6 shows these levels and the method correction factor proposed for them.

Table 6: Levels of study quality and method correction factors for assessing study quality in the Highway Safety Manual

Study design	**Level of control for confounding**	**Method correction factor**
Before-and-after	All potential sources of bias accounted for	1.2
	Accounts for regression to the mean	1.8
	Regression to the mean not controlled for, but judged to be minor if any	2.2
	Regression to the mean not controlled for and judged to be likely	3
	Severe lack of information on study design and results	5
Non-regression cross-section	All potential confounders controlled for by matching	1.2
	Most potential confounders controlled for by matching	2
	Traffic volume is only factor controlled for	3
	No confounding factors controlled for	5
	Severe lack of information on study design and results	7
Regression cross-section	All potential confounding factors controlled for by means of regression in an appropriate functional form	1.2
	Most potential confounding factors controlled for by means of regression in an appropriate functional form	1.5
	Several confounding factors controlled for; functional form is conventional	2
	Few variables used; functional form is questionable	3
	Severe lack of information on study design and results	5

To show the use of this system, consider the following example. Suppose a before-and-after study employing the empirical Bayes design – which involves controlling statistically for regression to the mean – reports that conversion of junctions to roundabouts have reduced injury accidents by 26 %. This corresponds to an accident modification factor (AMF) of 0.74. Suppose further that the standard error (SE) associated with this estimate of effect is 0.13. The statistical weight assigned to this estimate of effect in a formal research synthesis would be:

$$\text{Statistical weight} = \frac{1}{\text{SE}^2}$$

This estimate would therefore be assigned a weight of 1/(0.13 · 0.13) = 56.28.

The method correction factor is applied to the standard error in order to account for the risk that methodologically inferior studies produce biased and misleading estimates of effect. In a perfectly controlled randomised trial, the method correction factor would be 1.0. However, the best quality any non-experimental study can attain has been given a method correction factor of 1.2.

Adjusting the standard error by a method correction factor of 1.2 results an adjusted estimate of 0.160. The adjusted statistical weight becomes 39.08.

Thus, the method correction factors can be converted to equivalent study quality scores by which the statistical weights of each estimate of effect is multiplied in order to obtain a "quality-adjusted" statistical weight. This approach is attractive for several reasons. In the first place, it is consistent with the idea of measuring study quality by means of a bounded scale ranging from 0 to 1. The smallest method correction factor (1.2) is equivalent to dividing the un-adjusted statistical weight by 1.44 (1.2 · 1.2), which is equivalent to multiplying it by 1/1.44 = 0.694. In other words, if a well-controlled randomised trial has a quality score of 1.00, the highest attainable score for a non-experimental study is 0.694. In the second place, adjusting the statistical weights assigned to studies included in a meta-analysis in this way will always result in larger standard errors, the more so the poorer a study. This is consistent with the idea that poor studies should count for less than good studies. In the third place, the values proposed for the method correction factors will assign considerably less weight to poor studies than to good studies. A method correction factor of 7 is equivalent to a quality score of 0.02 on a scale in the (0, 1) range. In the fourth place, the system recognises the fact that the threats to study quality are different for different study designs.

The system does have a number of limitations that introduce an element of arbitrariness into it. Studies employing a multivariate cross-section design are rated for quality depending on whether they have controlled for "all", "most", "several" or "few" potentially confounding variables. These are rather vague categories, leaving a large room for disagreement in coding studies. To claim that a study has controlled for "all" potentially confounding factors, one should be able to list all these factors. This is never possible; at best only currently known potential confounding factors can be listed. Moreover, not all of these are equally important. Ideally speaking the potential confounding factors should not merely be listed, but rated for importance.

The system refrains from ranking study designs. Thus a good before-and-after study gets the same quality score as a good cross-section study. One may question this. In a different context, Hauer (2005A) has argued that observational cross-section studies tend to be inconclusive as far as determining causal relationships are concerned, and that the potentially most important confounding factors are better known in before-and-after studies than in cross-section studies. If this point of view is accepted, a case can be made for rating good before-and-after studies higher in terms of study quality than good cross-section studies. Trying to rank study designs is, however, a thorny problem.

The quality rating system developed for the Highway Safety Manual allows for adjusting study findings in order to remove bias attributable to not controlling for regression-to-the-mean and not adjusting for changes in traffic volume. The procedure proposed for this will not be described in detail in this report, but it does seem to be somewhat arbitrary. It is intended to be applied only to studies that, for example, failed to control for regression to the mean, and this failure is judged to be a likely source of bias in the study. However, not all studies report the information needed to judge whether lack of control for regression to the mean is likely to have biased study findings. Besides, even for studies that do report this information, the adjustment needed to remove the bias will be unknown. In some cases, the unknown regression-to-the-mean effect was merely 5 %, in other cases it could be as high as 40 or 50 %.

Despite these misgivings, the idea of adjusting for known errors in studies is attractive, at least if a non-arbitrary basis for these adjustments can be found and evidence can be produced that post-hoc adjustments actually remove bias from studies. A simulation study reported by Deeks et al. (2003), probing whether statistically adjusting for known confounders in studies of the effects of various medical treatments produced unbiased estimates of effect, concluded that the adjustments were not always successful and in some instances actually made things worse.

The main elements of the quality rating system developed for the Highway Safety Manual are promising and will be used as part of the basis for developing a more sophisticated quality scoring system for road safety evaluation studies in Chapter 8 of this report.

4.4 Lessons from listening to the experts and studying a quality rating system proposed by one of them

The main lessons learnt by asking a sample of leading road safety researchers how they understand the concept of study quality and by studying a quality assessment tool developed by one of these experts can be summarised as follows:

1. There is no consensus among leading road safety researchers about the concept of study quality. All the leading researchers treat the concept as multidimensional, but opinions differ regarding the relative importance of the various dimensions.
2. Aspects of study quality that are highlighted by a majority of the leading researchers include the use of appropriate statistical techniques, making assumptions made in analysis explicit and ensuring that they are plausible, and controlling for confounding factors.

3. A quality assessment tool has been developed for the Highway Safety Manual in the United States by Ezra Hauer. Compared to some of the highly detailed quality scoring systems that were included in the review in Chapter 3, the assessment tool must be regarded as rather crude. Nevertheless, this tool has a number of attractive features that are worth taking into account when developing a more sophisticated and detailed study quality assessment tool.

5 Testing a pilot scoring system for study quality

5.1 The scoring system

Based in part on the study assessment form proposed by Ezra Hauer and discussed in Chapter 4, and in part on previous attempts to develop an instrument for scoring road safety evaluation studies for quality (Elvik 1999), a pilot version of a formal quality scoring system was proposed in 2000 and tested in early 2001. This chapter presents this pilot quality scoring system.

Table 7 shows the items that were included in the pilot version of the quality scoring system for road safety evaluation studies. There are ten items altogether. The quality scoring system was to a great extent based on the threats-to-validity approach proposed by Cook and Campbell (1979). Cook and Campbell distinguish between four types of validity:

1. Statistical conclusion validity, which refers, among other things, to sampling techniques and types of statistical analyses employed in a study.
2. Construct validity, which refers to the way theoretical concepts and propositions have been defined operationally in an empirical study.
3. Internal validity, which refers to the basis for inferring a causal relationship between a treatment and the effects of the treatment.
4. External validity, which refers to the possibility of generalising the results of a study, or a set of studies, to other settings than those in which the studies were performed.

The validity system of Cook and Campbell has subsequently been updated in a book by Shadish, Cook and Campbell (2002).

The pilot quality scoring system presented here was meant for use in meta-analyses of road safety evaluation studies. This means that items of study quality that can be formally addressed as part of a meta-analysis were not included in the quality scoring system. There are at least three aspects of study quality that can be addressed by means of meta-analysis:

1. Sample size: In meta-analysis, study results are usually weighted in proportion to the inverse of their sampling variance. With respect to road safety evaluation studies, this means that the size of the accident sample is used as a weight.
2. Effect size: It is sometimes argued that a large effect is less likely to have been caused entirely by chance variation or confounding factors than a

small effect. In meta-analysis, the size of an effect is recorded for every study included, and is therefore included in the analysis.

3. Heterogeneity of effects: It is sometimes argued that a summary estimate of effect does not make sense when the individual estimates of effect are highly heterogeneous. Heterogeneity of effects can be tested formally in meta-analysis. This provides a basis for assessing the external validity of study results: Homogeneous results arising from heterogeneous study contexts are an indication that external validity is high.

Road safety evaluation studies are often purely empirical studies, with only a weak or no reference to theoretical concepts and statements. Construct validity is therefore, in general, not relevant. External validity was also omitted from the pilot quality scoring system, because it is difficult to assess for an individual study. External validity is more easily assessed as part of a meta-analysis, as indicated in the remarks to point 3 above.

Table 7: Main items of pilot quality scoring system for road safety evaluation studies to be used together with meta-anlysis

Main items of quality scoring system	**Specific items subsumed under each main item of quality scoring system**
1: Sample, quality of data, and statistical analysis	1A: Technique used to sample study units
(Statistical validity)	1B: Data referring to individual study units or aggregates of study units
	1C: Data specifying accident or injury severity
	1D: Reporting of statistical uncertainty of study results
2: Assessment of causal relationship between treatment and effect	2A: Direction of causality clear or not
(Internal validity)	2B: Degree of control of confounding variables
	2C: Knowledge of causal mechanism
	2D: Existence of dose-response pattern in results (optional)
	2E: Specificity of effects to target groups (optional)
	2F: Results conform with well established theory (optional)

Four items of the pilot quality scoring system refer to statistical validity; six items refer to internal validity. Table 8 proposes a set of scores for each item of the quality scoring system. The scores are assigned on an ordinal scale. A score of 1 represents the lowest quality. Scores of 2, 3 or higher numbers represent higher levels of quality.

Although a precise measurement of study quality is desirable, it was concluded that trying to measure study quality on an interval or ratio scale is not possible. The best that can be done is to score each item on an ordinal scale, going from “best” through “middle” to “worst”. In estimating an overall quality score, the scale is, however, in effect treated as if it is an interval scale.

Table 8: Scoring of items in pilot quality scoring system for road safety evaluation studies – ordinal scale approximating interval scale

Item of quality scoring system	Scores assigned (ordinal scale)
1A: Technique used to sample study units	3: Probability sample of study units from known sampling frame
	2: Sample chosen according to stated criteria (not a probability sample)
	1: Convenience sample or self-selected sample
1B: Data referring to elementary units of analysis or not	2: Data refer to elementary units, or these data can be retrieved
	1: Data refer to aggregates of elementary units only
1C: Specification of accident or injury severity	2: Level of accident or injury severity stated
	1: Level of accident or injury severity not stated
1D: Reporting of statistical uncertainty of study results	3: Confidence intervals and exact significance levels estimated or possible to estimate from information provided
	2: Only simple tests of significance at a set level reported
	1: No information provided on statistical significance or other quantitative measures of uncertainty; impossible to estimate
2A: Direction of causality	2: Can be determined to go from treatment to effect either a priori, or according to study design or other information
	1: Cannot be clearly determined, or may be reversed
2B: Degree of control of confounding factors	5: Full experimental control of confounding factors
	4: Statistical control of multiple confounding factors by means of multivariate techniques of analysis
	3: Partial control of confounding factors by means of well conducted quasi-experimental studies
	2: Inadequate control of confounding factors by means of observational or quasi-experimental studies
	1: No explicit control of any confounding factors
2C: Knowledge of causal mechanism(s)	2: Evidence of causal mechanisms provided
	1: No evidence of causal mechanisms provided
2D: Testing for dose-response pattern in results (optional)	2: A test for a dose-response pattern was included
	1: Testing for a dose-response pattern was possible, but not included in the study
2E: Specificity of effects to target groups (optional)	2: Testing for specificity of effects was included
	1: Testing for specificity of effects was possible, but was not included
2F: Results conform with well established theory (optional)	2: Study results are explicable in terms of well established theory
	1: Study results are not explicable in terms of well established theory

The range of scores assigned to each item varies from two (2 = best, 1 = worst) to five. Most of the items are scored on a dichotomous scale or a scale with three levels. A five-point scale is used just for one item, the degree of control of confounding factors. This item is very important. A somewhat more detailed scale was therefore developed for this item than for the other items. The possibility was considered of scoring studies by checking if specifically listed confounding factors were controlled for or not. This approach was rejected, because the number of potentially relevant confounding factors is very large, and varies depending on study design.

Most of the items proposed for the quality scoring system are more or less self-explanatory. Short comments will nevertheless be given to some of the items. Seven of the items should be checked for all studies. Three items are optional, that is a study can be scored on these items if applying them makes sense.

With respect to statistical validity, studies were scored according to:

1. How study units were sampled,
2. If data refer to elementary study units, or aggregates of these units,
3. If accident or injury severity is specified or not,
4. The types of statistical tests or estimates of uncertainty reported.

The six criteria proposed for internal validity were:

1. It should be possible to determine the direction of causality,
2. The effects of the most important potentially confounding factors should be controlled,
3. One or more causal mechanisms should be known,
4. If relevant, it should be possible to detect a dose-response relationship between treatment and effects,
5. Effects should only be found in the target group for the treatment (applicable if a sufficiently clear definition of the target group can be given),
6. Study findings should be explicable in terms of well-established theory or well-controlled empirical studies (applicable if relevant).

One of the major problems of any quality scoring system is to derive an overall quality score. Does it make sense to add or multiply scores on an ordinal scale? Does an arithmetic mean of ordinal scores make sense? Neither adding nor multiplying ordinal scores makes sense. Taking an average of ordinal scores also makes no sense, except if the purpose is to produce an average ranking. The absolute value of a mean of ordinal scores is meaningless, but a study whose mean score is 3.2 can be ranked as better than another study whose mean score is 2.4, provided both studies were scored on the same set of items on the same ordinal scale.

Given the fact that neither adding nor multiplying the scores assigned to each item make much sense, an alternative way of obtaining an overall quality score is proposed. The approach that is proposed involves converting the rating scale for each item to a relative scale, going from 0 to 1. For the dichotomous items, the possible scores are then 0.0 (1) or 1.0 (2). For items scored as 1, 2 or 3 on the ordinal scale, the corresponding relative scores become 0.0, 0.5, and 1.0. The five-point scale used to score for degree of control of confounding factors gets relative scores of 0.0, 0.25, 0.50, 0.75, and 1.0. The scale is, in other words, treated as an approximation to an interval scale.

To obtain an overall quality score, the mean of the relative scores is computed for statistical and internal validity. The mean score for statistical validity is then multiplied by the relative weight of 0.3, and the mean score for internal validity multiplied by the relative weight of 0.7. The weighted mean scores are then added. The resulting overall score will always take on values in the range from 0 to 1. The relative weights proposed for statistical and internal validity are, of course, arbitrary, and others may disagree with them. However, it is easy to select a different set of weights and examine how quality scores are affected by different weighting schemes.

5.2 Testing reliability and validity

5.2.1 Reliability

In order to test the reliability and validity of the pilot quality scoring system, a test of the system using five road safety evaluation studies was carried out. The following five studies were coded by means of the pilot quality scoring system:

1. Steven M. Rock. Impact of the 65 MPH speed limit on accidents, deaths, and injuries in Illinois. Accident Analysis and Prevention, 27, 207-214, 1995.
2. Kenneth W. Ogden. The effects of paved shoulders on accidents on rural highways. Accident Analysis and Prevention, 29, 535-362, 1997.
3. Lars Leden, Olli Hämäläinen, Esa Manninen. The effect of resurfacing on friction, speeds and safety on main roads in Finland. Accident Analysis and Prevention, 30, 75-85, 1998.
4. Robert G. Ulmer, David F. Preusser, Susan A. Ferguson, Allan F. Williams. Teenage crash reduction associated with delayed licensure in Louisiana. Journal of Safety Research, 30, 31-38, 1999.
5. Michael S. Griffith. Safety evaluation of rolled-in continuous shoulder rumble-strips installed on freeways. Transportation Research Record, 1665, 28-34, 1999.

These studies were selected because they were recently published in peer-reviewed journals and were easy to retrieve. Moreover, an informal impression had been formed to the effect that these studies were of a better than average quality.

A measuring instrument is reliable if it gives the same result when the same phenomenon is measured repeatedly in identical conditions. Applied to a formal quality scoring system for evaluation studies, the concept of reliability can be defined as the extent of agreement between different individuals scoring the same study for quality by means of the same scoring system. Reliability was tested by giving a sample of five researchers the five studies listed above for coding by means of the pilot quality scoring system. The scores assigned were then compared. If the scores assigned to a given study are identical, the scoring system is highly reliable. If the scores differ greatly, the scoring system is not very reliable. Two types of reliability were estimated:

1. Inter-rater reliability, which is the extent to which two or more raters agree on the scores assigned to a specific study, and
2. Item-specific reliability, which is the extent to which a specific item on the quality scoring system is scored identically by a group of raters.

In practice, a reliability of one hundred percent cannot be expected. The reliability of a quality scoring system will be regarded as satisfactory if more than about 75 of 100 scores assigned agree. The following measures of reliability were computed:

$$\text{Percent agreement} = \frac{\text{Number of identical scores}}{\text{Total number of scores assigned}}$$

$$\text{Kappa index} = \frac{\text{Proportion of identical scores - Expected by chance}}{\text{1 - Expected by chance}}$$

Percent agreement is the simplest measure of the reliability of an instrument. However, it does not account for the fact that a certain proportion of scores will agree by chance. If, for example, two raters score an item as “yes” or “no”, they will agree 50% of the time even scores are assigned entirely at random. The Kappa index adjusts for the percentage of identical scores expected by chance. It takes on values between –1 and +1. A general formula for the proportion of scores expected to be identical by chance is:

$$\text{Proportion of chance agreement} = \left(\frac{1}{g}\right)$$

In which g is the number of categories used to score a specific item and n is the number of people scoring it. For an item having three categories, the proportion of chance agreement between a pair of raters is (1/3).

The set-up and results of the reliability test is shown in Appendix 1. The ten items of the quality scoring system form ten lines. The scores assigned by each rater are listed next to each other. The next set of columns shows the pairwise agreement between the raters. The two columns to the right shows item-specific agreement and the level of item-specific agreement expected by chance.

As an example, raters 1 and 2 scored study 1 identically except for items 2 (level of aggregation of data) and 9 (test for specificity of effect). The proportion of agreement was between raters 1 and 2 with respect to study 1 was therefore 0.8. Five raters scored study 1. They agreed perfectly on the score assigned to item 3 (accident or injury severity stated), which therefore had an item-specific reliability score of 1.0 for study 1.

When all studies and all raters are combined, the inter-rater and item-specific agreement were both 0.67. There was, however, more variability in scoring between items than between raters. The Kappa-index varied from 0.04 for item 10 to 1.00 for item 5. Only items 4, 5, 6 and 7 had a Kappa-index of more than 0.60. The reliability of the scoring system was therefore lower than ideal.

In the pilot quality scoring system, six items can take on two values, three can take on three values and one has five values. This may be too crude and will not make the scoring system as sensitive as desirable. The points of view obtained from the international sample of road safety researchers, see Chapter 4, also suggest the need for developing a more sensitive scoring system. If all ten items are treated as relevant, the pilot quality scoring system can take on values between 10 (for a study that gets the bottom score on all item) and 26 (for a study getting top score on all items). Study 1 got a total score varying from 17 to 19, with a mean score of 18.2. For study 2, the range was 13-20, and the mean score 16.4. For study 3, the range was 15-22, the mean score 17.2. Study 4 got a total score varying from 14 to 18, with a mean of 16.4. Finally, for study 5, the range was 15-19, and the mean 17.4. Within the range spanning from 10 to 26, all studies got a mean score close to the midpoint of the range (18), indicating that their quality was “half way” from really bad to outstanding. The differences between studies with respect to the mean score are quite small. This could either indicate that all these studies were of nearly the same quality, or that the quality scoring system is insensitive.

5.2.2 Validity

When the pilot quality scoring system was developed, the idea was to test the validity of the system by using the opinions of the sample of experts, see Chapter 4, as a criterion of validity. This was based on the hope that leading international experts on road safety research would agree on the concept of study quality and would be able to articulate this concept in sufficient clarity and detail to form the basis for developing a formal instrument designed to measure quality.

The opinions of the experts regarding study quality were presented in Chapter 4. The sample is very small – in fact only eight replies were obtained. Moreover, not all of the experts seemed to be inclined to offer an extensive description of their notion of study quality. Some of the answers were quite brief and did not give any clues as to how best to make the concept of quality operational. On top of this, the answers were surprisingly diverse, suggesting that there may indeed not be much agreement in depth about the meaning of the concept of study quality, or, at the very least, that different experts disagree about which aspects of study quality are the most important.

If, despite these limitations of the survey of experts, one tries to use the results of this survey as an indication of the meaning of the concept of study quality, the following conclusions appear reasonable:

1. It makes sense to talk about the quality of road safety evaluation studies, and to try to assess study quality in a standardised manner.

2. Study quality is a multidimensional concept; it refers to several aspects of study design and conduct and must be assessed in terms of each of the relevant dimensions.
3. Relevant dimensions of study quality include (but is not limited to):

(a) Technique used to obtain a study sample (random sampling or other)

(b) Sample size

(c) Numerical estimate of effect of the measure being evaluated

(d) Estimate of the uncertainty of the effect

(e) Control of at least major potential confounding factors (preferably to be listed explicitly)

(f) A check on data quality

(g) An explicit statement of the assumptions made in analysis of the data

(h) Use of appropriate statistical techniques of analysis

(i) An attempt to uncover causal mechanisms

(j) A discussion of study findings in view of other studies, and

(k) An honest interpretation of study findings, allowing for the possibility that these are inconclusive.

4. The possibility is not ruled out that an overall numerical score for study quality can be developed.

The fact that the pilot quality scoring system includes some of the items listed above does not by itself constitute any test of the validity of the system. It merely shows that the pilot quality scoring system contains some elements that a few people have indicated that they regard as aspects of study quality. In fact, how best to test the validity of a formal quality scoring system is a complex issue, which it is fair to say has so far not been fully resolved (see, for example, the discussion by Verhagen et al 2001).

We will return to the approaches that may be taken to testing the validity of a formal quality scoring system in chapter 6. For the moment, the experiences gained in developing the pilot quality scoring system will be summarised.

5.3 Lessons from the pilot testing

The pilot quality scoring system for road safety evaluation studies presented above was not a success. There are two reasons why this system is still presented and discussed in this report:

1. Development of the system shows that it is, in principle, possible to develop a quite simple formal quality scoring system, and that it is possible to test the reliability of such a system.
2. There are important lessons to be learnt from the pilot quality scoring system, although the system was not very successful.

The lessons that were learnt, can be summarised as follows:

1. It is evident, both from the many quality scoring systems presented in Chapter 3, and from the survey of leading road safety experts, that study quality is a very complex and multi-dimensional concept. There is no agreement about a general definition of the concept of study quality, nor is there any agreement regarding an operational definition of the concept. A tentative definition was proposed in Chapter 1.
2. A gold standard for assessing study quality, that can be used as an external point of reference in evaluating a formal quality scoring system, does not exist. Study quality is a theoretical construct only; it does not have any objective empirical reference in the same sense as observations of physical quantities or chemical processes.
3. Since no objectively observable empirical reference to the concept of study quality exists, testing the validity of any formal quality scoring system – any numerical instrument designed to measure quality – will be difficult. While some systematic assessment of validity may be possible, its essential elements will be quite different from standard models for testing empirical hypotheses in science (i.e. by comparing predictions derived from the hypotheses to data reflecting the underlying reality). How can you falsify the statement that: "This study is bad and is correctly scored at 0.2 on a quality scale ranging from 0 to 1".
4. It is difficult to derive a summary quality score from a set of items in a non-arbitrary manner. The best that can be accomplished is to derive summary quality scores in a transparent and reproducible manner. Nearly all methods for deriving summary quality scores that have been proposed treat items that have been scored on an ordinal scale as approximations to an interval scale. One rarely sees any discussion of whether this makes sense from a statistical point of view.
5. It was not possible to meaningfully test the validity of the pilot quality scoring system for road safety studies. The reliability of the system was not satisfactory with respect to scoring studies for control for confounding factors. The scale used for rating studies according to control for confounding factors was too crude and left too much room for judgement on the part of raters.
6. The pilot quality scoring system did not seem to have sufficient discriminative power. The scores assigned to studies that were initially believed to be of different quality were too close. The range of the scores was too small.
7. The pilot quality scoring system contained some items that are likely to vary too little between studies to be useful. Sampling technique is one example: nearly all road safety evaluation studies rely on convenience samples. Generalisation of the findings of these studies is nearly always non-statistical, and has to rely on a grounded theory of causal inferences, such as the theory proposed by Shadish, Cook and Campbell (2002).

Based on these lessons, it was concluded that substantially more work needed to be put into the development of a formal quality scoring system for road safety evaluation studies. Part of that work was the survey of existing formal quality scoring systems presented in Chapter 3. Very little came out of that survey. It merely confirmed that most of the work that has been done with respect to formal quality assessment of empirical studies has been sloppy and thoroughly non-scientific.

6 A typology of study designs and threats to validity

6.1 The validity of study quality assessments

In order to develop a comprehensive quality scoring system for road safety evaluation studies, it is necessary to identify all study designs used in road safety evaluation studies and all threats to validity that are relevant for each design. At this point, it may be useful to propose a formal definition of validity for a quality scoring system:

A numerical scale measuring study quality is valid if it includes all elements of study design and analysis that may influence study results and rates as most important those elements of study methods (design and/or analysis) that may have the greatest influence on study results.

To be valid in this sense, a quality scoring system should first and foremost be comprehensive, that is it should include all study designs that are applied in road safety evaluation studies and all methodological problems that may influence the results of studies employing a certain design. This is a tall order. Yet, the definition of validity proposed here corresponds closely to the definition of study validity given by Shadish, Cook and Campbell (2002): A study is valid to the extent that its findings approximate the truth. A statement or finding is true if it corresponds to reality.

Reality in the present context is an unbiased estimate of the effects of a road safety measure, i.e. an estimate that is not affected by any known sources of bias and for which no reasons can be given for suspecting that unknown or unnamed sources of error could have influenced the estimate. An unbiased estimate of effect shows the true effect of a road safety measure.

What we are looking for when assessing study quality is therefore methodological artefacts and weaknesses that may influence the estimate of effect. The extent to which various methodological shortcomings influence estimates of effect may, in principle, be determined by conducting methodological studies that evaluate how study findings are influenced by, for example, not controlling for regression to the mean.

Study quality is, at least ideally speaking, an objective characteristic of a study. It depends strictly on whether a study has applied appropriate methods or not. While different researchers may have different opinions regarding which aspects of study methods that are most important, this may in principle be determined by evidence.

The most important basis for developing a comprehensive system for assessing study quality is therefore an equally comprehensive programme of methodological research, designed to assess the effects of as many aspects of study methodology as possible. A review of relevant methodological research will be presented in the next chapter. In this chapter, a framework for identifying relevant methodological studies is proposed by developing a classification of study designs and a list potentially confounding factors inherent in each study design.

6.2 Study designs used in road safety evaluation studies

Table 9 lists study designs employed in road safety evaluation studies.

Table 9: Study designs employed in road safety evaluation studies

Study design (main group)	Version of study design
Experiments	Full random assignment
	Clusters assigned at random
	Matched pairs assigned at random
Before-and-after	Model-based empirical Bayes studies
	Simpler empirical Bayes designs
	Before-and-after with matched comparison group
	Before-and-after with non-equivalent comparison group
	Before-and-after with treatment reversal
	Before-and-after with internal comparison group
	Before-and-after with data on some potentially confounding variables
	Simple before-and-after studies
Cross-section studies	Comparative studies controlling for confounding by stratification
	Comparative studies applying multiple classification schemes
	Simple comparative studies with/without safety measure
Case-control studies	Case-control studies statistically adjusting for confounding factors
	Case-control studies controlling for confounding by stratification
	Double pair comparison method
	Simple case-control studies
Multivariate models	Generalised Poisson regression models (e.g. negative binomial)
	Multinomial or ordered logit models; mixed logit models
	Logistic regression models
	Ordinary least squares linear regression models
Time-series analysis	Structural time-series models including explanatory variables
	Time-series with a comparison series
	Analysis of a single time-series

The designs listed in Table 9 are not exhaustive, but include those that are most commonly applied in road safety evaluation studies. For each main type of design, different versions of the design have been listed. The best versions are listed first, then simpler versions – more at risk of confounding – are listed. It is beyond the scope of this report to describe each design in detail. Fairly detailed descriptions can be found in, for example, Shadish, Cook and Campbell (2002) or Lund (2002).

6.3 Important threats to validity in road safety evaluation studies

For any study design, threats to study validity exist. For the purpose of assessing study validity, threats to internal validity will be regarded as the most important. Reasons for focusing on threats to internal validity will be given in Chapter 7. Table 10 lists major threats to internal validity for each main type of study design.

Table 10: Threats to internal validity relevant to different study designs in road safety evaluation studies

Study design (main group)	Major threats to internal validity
Experiments	Unsuccessful randomisation; pre-trial equivalence violated
	Diffusion of treatment to control group
	Differential attrition between groups
	Hawthorne effects
Before-and-after	Regression-to-the-mean
	Long-term trends
	Exogenous changes in traffic volume
	Co-incident events
	Introduction of multiple measures
	Accident migration
Cross-section studies	Self-selection of subjects to treatment
	Endogeneity of treatment
	Differences in traffic volume
	Differences in traffic composition
	Differences with respect to any other relevant risk factor
Case-control studies	Non-equivalence of cases and controls with respect to accident severity
	Non-equivalence of cases and controls with respect to prognostic factors
	Heterogeneity of treatment
Multivariate models	Endogeneity of treatment
	Wrong functional form of explanatory variables
	Collinearity among explanatory variables
	Omitted variable bias
	Erroneous specification of residual terms
	Mixing levels of accident severity
	Inappropriate model form
	Inappropriate choice of dependent variable
Time-series analysis	Inadequate adjustment for explanatory variables
	Co-incident events
	Erroneous specification of residual terms

The threats to validity listed in Table 10 does not include all conceivable threats, but is confined to those that have been found, or are believed, to be the most important. It is impossible for any practical system for assessing study quality to assess studies in terms of more than a few important potentially confounding variables.

The meaning of each potential source of confounding will be discussed in Chapter 7. For the moment, a preliminary classification will be made of study designs in terms of how well they control for the various confounding factors.

6.4 How well do different designs control for threats to internal validity?

Experimental study designs are sometimes believed to control perfectly for all potentially confounding factors. Unfortunately, this is not true. Although the experimental design as such is methodologically strong, a badly executed experiment may introduce confounding that a more rigorously executed experiment would avoid. It is therefore not possible to state in general whether an experimental study design successfully controls for the confounding factors listed in Table 10. It may do so; then again it may not. The presence of the potential confounding factors must be assessed for each experiment.

Before-and-after exist in many versions. The empirical Bayes (EB) design, based on an accident prediction model, is generally regarded as the best form of before-and-after study (Persaud and Lyon 2007). This study design will normally control for regression-to-the-mean, long-term trend and exogenous changes in traffic volume. It may, depending on the details of the design and the data collected, also control for the use of other road safety measures than those that are the focus of the evaluation and for accident migration. Simpler versions of the EB design, not based on an accident prediction model, will normally also control at least for regression-to-the-mean and long-term trends.

Before-and-after studies with a matched comparison group may – depending on the criteria used for matching groups and on how successful matching was – control for regression-to-the-mean, long-term trends and changes in traffic volume. However, this design will not necessarily control for these confounding factors, in particular not if matching was based on a small sample or was unsuccessful. Before-and-after studies with a non-equivalent comparison group may control for long-term trends, but will normally not control for regression-to-the-mean. In some before-and-after studies, a safety measure is introduced, then removed and then sometimes re-introduced. While this design is rarely applicable, it can be very informative when used and may control for regression-to-the-mean and long-term trends. For a case of this design, see Stewart (1988).

A before-and-after study with an internal comparison group is a study in which some accidents are believed to be influenced by the road safety measure and other accidents, believed not to be influenced by the road safety measure, are used as comparison group. This study design has been widely applied to evaluate the effects of road lighting, relying on the assumption that only accidents in darkness are affected by road lighting, permitting the use of accidents in daylight as a comparison group. The design is weak and will normally not control for any confounding factors.

In some before-and-after studies, data are collected on traffic volume or other variables that may influence the number of accidents. It may then be possible to control for changes in these variables – albeit often in very simple minded way, since the changes are often simultaneous to the introduction of the road safety measure. Simple before-and-after studies do not control for any confounding factors and their findings should never be trusted.

Cross-section studies are also found in many versions, but all of them are quite complex to assess as far as control for confounding factors is concerned. These studies tend to use accident rates as the dependent variable, often the number of accidents per million vehicle kilometres. This choice is the source of much trouble. In the first place, accident rates are not independent of traffic volume (Hauer 1995). Thus, to be comparable, accident rates with and without a certain safety measures ought to refer to roads with identical traffic volume. In practice, this is rarely the case, as traffic volume tends to be one of the criteria for introducing a road safety measure. In the second place, accident rates vary as a function of the composition of traffic, not just traffic volume. Ideally speaking, the relative contributions of large and small vehicles, pedestrians and cyclists to the traffic stream should be the same at locations that are compared. This can rarely be ascertained, as pedestrian and cyclist volume is often unknown. In the third place, accident rates are influenced by a host of risk factors. One cannot even hope to enumerate all these factors.

The quality of cross-section studies is very difficult to assess. There will always be potentially confounding factors these studies did not control for. It is tempting to conclude that cross-section studies are un-interpretable and that this study design should never be used. However, since a large number of such studies have been reported, they must somehow be included in the quality scoring system.

Case-control studies are also notoriously prone to error (Crombie 1996). The prospects for assessing their quality systematically are nevertheless slightly better than for cross-section studies, as case-control studies have been extensively applied in epidemiology and many potential sources of error are known. One may therefore draw upon the extensive literature in epidemiology to assess the quality of case-control studies (see, for example, Elwood 1998).

In recent years, multivariate accident models have increasingly been used to evaluate the effects of road safety measures. This is currently a very active field of research. Less work has been done to critically assess multivariate accident models, but some methodological aspects of such models are discussed by Elvik (2007A). Based on that discussion, some potential threats to validity are listed in Table 10. Illustrations of how they can influence study findings are given in Chapter 7.

Time-series models are discussed in some detail by Cook and Campbell (1979). However, techniques have developed considerably since then. Quddus (2008) discusses recent innovations in time-series analysis and their application to road safety studies. His chief conclusion is that for time-series characterised by a low mean value, integer-valued autoregressive Poisson models perform better than the conventional ARIMA time-series models. The comparison refers to model fit and does not address the question of whether one model controls better for confounding variables than another.

6.5 A hierarchy of study designs

Is it possible to form a hierarchy of study designs with respect to how well they control for confounding factors? This is difficult, as there are many versions of each type of design. A well-conducted before-and-after study can be better than a poorly conducted multivariate analysis. For each study design, however, a state-of-the-art version of the design exists. For before-and-after studies, for example, this would be the model-based empirical Bayes design. For an experiment, it would be a study that controlled pre-trial equivalence, that monitored attrition rates and that tested for Hawthorne effects or other unintended effects of the experiment and adjusted for such effects statistically if they were detected.

If the assumption is made that each study design is implemented in a state-of-the-art version, a rough hierarchy of study designs can be proposed as follows:

Level	*Designs included*
Best	Experiments (randomised controlled trials)
Second best	Before-and-after studies; multivariate analyses
Third best	Case-control studies; time-series analyses
Fourth best	Cross-section studies (not based on a multivariate model)

It is emphasised once more that each design comes in different versions and that a badly conducted study employing one design could be worse than a well-conducted study employing a design that would normally be rated as inferior.

6.6 Summary of lessons learnt

The chief lessons learnt in this chapter can be summarised as follows:

1. The concept of validity as applied to a quality scoring system can be defined in terms of the inclusiveness and weighting scheme of the system. A quality scoring system is valid if it includes everything that influences study quality and assigns a weight to each factor that reflects the potential size of the bias it could generate in study findings. Methodological research is needed in order to establish a more formal criterion of validity.

2. To develop a quality scoring system which is valid in this sense, it is necessary to include all study designs that are employed or likely to be employed in a field and identify all threats to the internal validity of these study designs. Different threats to internal validity will be associated with different study designs.

3. In road safety evaluation studies, six main types of study design are used: (a) Experimental designs (randomised controlled trials), (b) Before-and-after studies, (c) Cross-section studies, (d) Case-control studies, (e)

Multivariate analyses, (f) Time-series analysis. For each of these designs, several versions exist.

4. A valid quality scale must be tailored to each study design. This means that the principal threats to internal validity must be identified for each study design. A preliminary list of these threats is provided. Threats to internal validity have been extensively studied for experiments, before-and-after studies and multivariate accident models. Less is known about potential sources of error for the other study designs.
5. Since scoring studies for quality needs to be based on a scale tailored to each study design, there is a need for forming a hierarchy of study designs, as studies evaluating a certain road safety measure will often have employed different designs. In a meta-analysis, it will often be useful to try to combine evidence from studies that used different designs. A preliminary ranking of study designs is proposed, but this ranking needs to be converted to a numerical scale.

7 Elements of a rationally justified quality scoring system for road safety evaluation studies

7.1 The problem of arbitrariness in quality scoring systems

The problem of arbitrariness in study quality assessment is very severe. Unfortunately, this problem has not been properly addressed in previous research. On the contrary, the review of quality assessment tools in Chapter 3 confirmed the relevance of the criticism by Greenland (1994), quoted in Chapter 1. The development of quality assessment tools has not been cumulative; very few of the tools refer to other work in the area; an explicit definition of study quality is rarely given; very few attempts have been made to assess the validity of the proposed scales; the diversity of items regarded as relevant for study quality is staggering; there is no consensus on the relative importance of the items included in the tools; reliability is not always reported; not all scales have sufficient discriminative power. In short, almost all previous attempts at developing a numerical scale for assessing study quality have been based on a subjective, ad hoc approach – not on a rigorous scientific method. The resulting scales are therefore highly arbitrary and seem to reflect mainly the personal opinions of each researcher regarding study quality.

Elvik (2007B), in discussing operational criteria of causality for observational road safety evaluation studies, offers the following comments on this issue:

“Is it possible to develop an overall score indicating study quality based on the criteria? Developing an overall score is no problem. It can be done simply by assigning numbers to the criteria and converting the verbal assessment to a numerical score. The trouble, as pointed out by Greenland, is that any such numerical score will be arbitrary. Different researchers may assign different weights to the criteria, resulting in different scores for the same study. Trying to get a consensus on a numerical scoring system is difficult. Researchers may agree that, for example, control for confounding is very important. But how important is it? Should it carry 60 % of the sum of weights given to the criteria or 80 %? It is hard to give a good justification for choosing one or the other. Even if a widely accepted scoring system could be developed, there might still be room for disagreement. Does a study that gets 60 % of the maximum score support a causal inference, or does it not? Some researchers may be reluctant to infer causality unless the score in favour of doing so is at least 80 %; others may be willing to do so if the score favouring it is only 60 %.”

It is important to minimise the contribution of arbitrariness to a numerical quality scoring system. This chapter will discuss how to reduce arbitrariness. While it may be impossible to avoid arbitrariness altogether, each step taken in developing a quality scoring system should be carefully justified. The main source for justifying the choices made in developing a quality scoring system will be methodological studies that show that a certain aspect of study design or analysis may influence study findings. This chapter will review some of this methodological research.

7.2 Elements of a rational (non-arbitrary) quality scoring system

In a series of papers and reports, Elvik (1998, 2002A, 2002C, 2003A, 2004, 2007A, 2007B) has discussed some elements that can form the basis of a rationally justified quality scoring system for road safety evaluation studies. These elements include:

- Theory
- Causal modelling
- Causal criteria
- A validity framework
- A typology of study designs
- A typology of potentially confounding factors
- Methodological research

The role of each of these elements in constituting a rational foundation for the development of a numerical scale for assessing the quality of road safety evaluation studies will be briefly reviewed.

Road safety evaluation research is applied research that does not have the development or testing of theories as its main objective. Hence, much of this research tends to be atheoretical. This means that well-established theory, supported by law-like relationships between variables can rarely support the interpretation of road safety evaluation studies to any great extent. As an example, one would normally expect road lighting to reduce accidents at night, since visibility is improved. However, it is not possible to rule out adverse effects, resulting from an increase in traffic at night, higher speed and reduced alertness. These and other factors may in principle eliminate any favourable effect of road lighting on accidents.

Elvik (2004) has proposed a theoretical framework for interpreting road safety evaluation studies. This theoretical framework is not a theory, but it helps in identifying relevant variables in road safety evaluation studies. Once relevant variables have been identified, the relationship between these variables can be modelled in terms of a generic causal diagram, examples of which are given by Elvik (2003A, 2004). Figure 5 shows a generic causal diagram including all classes of variables that are judged to be relevant for assessing the quality of road safety evaluation studies.

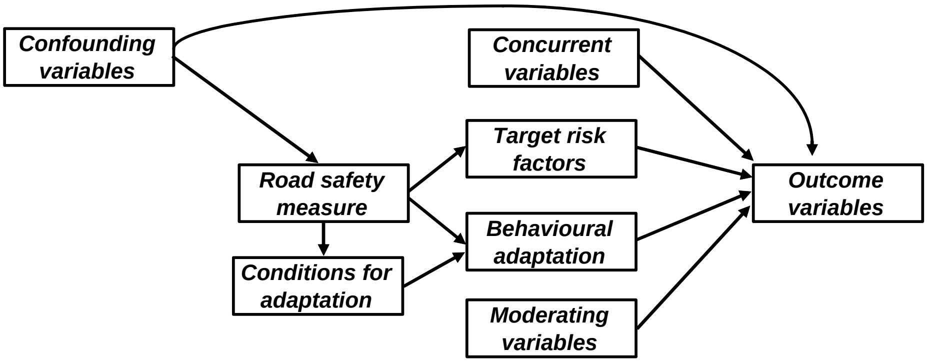

Figure 5: Classes of variables relevant in road safety evaluation studies

This causal diagram may serve as a reference point for the criteria of causality proposed by Elvik (2007B). These criteria are listed in Table 11. Criteria of causality refer to internal validity. According to the validity framework of Shadish, Cook and Campbell (2002), a distinction is made between four types of validity:

1. Statistical conclusion validity
2. Construct validity (theoretical validity)
3. Internal validity
4. External validity

The criteria of causality listed in Table 11 refer partly to all these types of validity. The first three criteria may be regarded as aspects both of statistical conclusion validity and internal validity. These criteria are relevant to internal validity, because the existence of a statistical relationship between variables is normally regarded as a necessary (but not sufficient) condition for a causal relationship. Moreover a strong statistical relationship is regarded as more likely to be causal than a weak relationship. Finally, consistency and regularity is an important characteristic of a causal relationship. A given cause should always produce the same effect, within the bounds of random variation, when acting within a given context. The latter clause ("within a given context") is needed because sometimes the size of an effect depends on contextual factors that are not directly influenced by the causal variable.

Clarity of causal direction (criterion 4) is related specifically to internal validity. The direction of causality is clear when, of two variables A and B, it can be determined which is the cause and which is the effect. Criteria for determining causal direction include: (a) Temporal order of variables: causes come before effects in time; (b) A priori considerations: age and sex may be determinants of accidents, but not the other way around; (c) Reversal of effects: when a cause is removed, the effect will be reversed.

Table 11: Criteria of causality for road safety evaluation studies. Based on Elvik 2007B

Criterion of causality	Theoretical definition	Operational definition
1. Statistical association	There should be a statistical association between cause and effect	A statistically significant change in variables measuring safety associated with safety treatment
2. Strength of association	A strong association is more likely to be causal than a weak association	Treatment effect stated in terms of effect size compared to effect sizes for other variables present in the data
3. Consistency of association	A consistent association is more likely to causal than an inconsistent association	The consistency in direction and size of effect attributed to safety treatment across subsets of the data or different model specifications, assessed by means of a consistency score (see text)
4. Clear causal direction	It should be clear which of two variables is the cause and which is the effect	The temporal order between variables; a priori considerations; reversal of effect when treatment is removed
5. Control for confounders	The association between cause and effect should not vanish when confounding variables are controlled for	The identification of potentially confounding variables; invariance of the effect attributed to treatment with respect to potentially confounding variables controlled for; completeness of the control for confounding variables
6. Causal mechanism	The mechanism generating an effect should be identified and measured	Changes in target risk factors influenced by a road safety treatment and changes in risk factors representing behavioural adaptation to the treatment
7. Theoretical explanation	A plausible theoretical explanation of the findings of a study should be given	Findings should not contradict well established laws of physics or laws of human perception and information processing
8. Dose-response pattern	Treatments administered in large dose should have larger effects than treatments administered in small doses	Treatments that are intense or have large effects on target risk factors should be associated with larger changes in safety than less intense treatments or treatments with small effects on target risk factors
9. Specificity of effect	Effects of a cause operating only in a certain clearly defined group should only be found within that group	An effect of safety treatments targeted at clearly defined groups should only be found in those groups and not in other groups

Control for confounding (criterion 5) is by far the single most important criterion of causality for road safety evaluation studies. As will be shown later in this chapter, poor control for confounding factors can seriously distort the findings of road safety evaluation studies and make them completely worthless. It is therefore very valuable to conduct methodological studies designed to assess the effects of confounding factors on the results of road safety evaluation studies. Such studies constitute a key element in justifying a scheme of weights to items of a numerical scale for study quality. The greater the potential effects of a confounding factor, the greater should be the weight assigned to controlling for this factor when assessing study quality.

If it is possible to identify the causal mechanism producing a statistical association between cause and effect (criterion 6), this may strengthen a causal inference. Strictly speaking, however, it is not necessary to know the causal mechanism if a study has controlled for all important confounding factors. This is very clearly illustrated in the study reported by Elvik (2003A). In that study, salting of roads during winter was one of the cases studied. The causal mechanism generating the effect had been studied in great detail in this case. Spreading salt on the road surface was associated with a reduction of percentage of traffic taking place on a snow- or ice-covered road surface. This was in turn associated with increased road surface friction and a shorter stopping distance. However, due to behavioural adaptation among road users, the net effect on stopping distance was very small. Despite this, the study estimated a large effect on accidents. Closer examination of the data revealed that the salted roads differed systematically from the unsalted roads in terms of a number of confounding factors that were not controlled for. This was a cross-section study, not a before-and-after study. Re-analysis of the data found that all the effect attributed to salting roads was more likely to be attributable to confounding factors that were not controlled for in the study. Thus, studying causal mechanisms may be useful by revealing anomalies in study findings that may in turn alert the researcher to the presence of confounding factors not controlled for.

Although road safety evaluation research is not based on strong theory, causal inferences can be strengthened if a theoretical explanation of study findings can be given (criterion 7). Thus, as an example, the effects of guardrails in reducing accident severity can be explained in terms of well-known relationships between the stiffness of objects and their energy-absorbing properties. A rockside is unyielding and does not absorb any energy. A vehicle striking it will absorb all the energy; hence rocks close to the road will be more hazardous to strike than a guardrail. By comparing the outcomes of crashes into guardrails with the outcomes of crashes into other objects – controlling for confounding factors (e. g. impact speed, size of vehicle, first impact point, wearing of seat belts, etc, etc) – one may determine if the severity of outcomes is systematically related to the energy-absorbing characteristics of the objects, thus explaining findings in theoretical terms.

The presence of a dose-response pattern between a treatment and an effect has long been regarded as an important criterion of causality in epidemiology (criterion 8). This applies to road safety studies as well. Two forms of dose-response patterns are relevant in road safety evaluation studies. The first form is related to the effects of a road safety measure on target risk factors. If, for example, the measure is designed to reduce speed, the greater the reduction in speed, the greater the effect on accidents. The second form is related to characteristics of the measure. For example: the more intense police enforcement, the greater the reduction of violations and the greater the effect on accidents. The dose-response criterion is useful when it can be applied. However, in assessing study quality, it is of course not study findings that are important. Dose-response as an aspect of study quality is related to whether a study was designed so that a dose-response pattern could be detected if it was present – not to whether such a pattern was found or not.

Specificity of an effect to a target group (criterion 9) is also a useful criterion when applicable. Again, in assessing study quality, it is unimportant whether an effect was found in the target group for an intervention and not outside the target group. What matters for the purpose of assessing study quality is whether a clearly designated target group was defined, allowing the specificity of an effect to be tested for.

The main purpose of road safety evaluation studies is to estimate the effects of road safety measures. Hence, the criteria for assessing study quality should focus on whether a causal relationship between the road safety measure and changes in road safety can be inferred or not. Are more criteria of study quality, related to statistical conclusion validity, construct validity or external validity needed in assessing the quality of such studies?

As far as statistical conclusion validity is concerned, three aspects of it (presence, strength and consistency of a statistical relationship) are included among the criteria of causality. However, three additional aspects are relevant:

1. Sampling technique
2. Source of accident data (it is assumed that a variable based on accident data is the dependent variable in all studies)
3. Specification of accident severity

As already mentioned, convenience samples are very common in road safety evaluation studies. There are a number of good reasons for that. Traffic engineering measures, like converting junctions to roundabouts, are not implemented by drawing a random sample from an inventory of all junctions. Junctions are selected for conversion based on more or less explicit criteria, which may include traffic volume, accident history, cost of conversion, proximity to other junctions allowing for economies of scale in engineering works, etc, etc. It will rarely be the case that all criteria are stated explicitly and almost never the case that their relative importance in selecting junctions for conversion can be determined. So, what the researcher is usually left with, is a list of junctions that have already been converted to roundabouts or, more unusually, a list of junctions for which conversion to a roundabout is planned. This is the sample that can be used in the evaluation study and it will only rarely be possible to describe in very precise terms how it was obtained.

A similar logic applies to the process leading to the introduction of very many road safety measures. Police enforcement, for example, is targeted at certain roads and certain groups of road users. It is not introduced by selecting the times and places at random, although theoretical considerations suggest that doing so may actually be superior in the long term to selecting the targets of enforcement according to, for example, traffic volume, accident history or violation rate (Bjørnskau and Elvik 1992).

In short, it is very rarely the case that road safety measures are introduced by means of random sampling from a known sampling frame. Despite this, it is not satisfactory that the sampling procedure is very rarely described in road safety evaluation studies. It is therefore proposed that a quality scoring system should at the very least note if any information is given at all regarding how the sample was obtained.

By far the most common source of data in road safety evaluation studies is the official accident record for a country or state. This is subject to incomplete and biased reporting (Elvik and Mysen 1999). It is easy to construct numerical examples showing how incomplete and biased accident reporting may introduce bias in road safety evaluation studies. However, as long as an accident record known to be complete does not exist, there is no way of adjusting for this bias. The choice facing researchers is either not to perform an evaluation study at all, since there is a risk that incomplete accident reporting could bias the findings, or use the accident data that are available and accept the risk of an unknown bias that cannot be estimated as no data are available for that purpose. Official accident data are known to be more incomplete for pedestrians and cyclists than for other groups of road users. A supplementary collection of data, for example self-reported accidents, may sometimes be needed to reduce bias.

In view of this, it makes little sense to assign a low quality score to studies that are based on official accident statistics, since better alternatives are hard to come by. Data recorded by hospitals are sometimes available, but even hospital records are not complete. Self-reported accident data are sometimes used, but the reliability of these data can be questioned, as not everybody will have the same understanding of what counts as an accident, and past accidents may be forgotten.

Specification of accident or injury severity is important. It is known that the effects of many road safety measures vary according to accident or injury severity. Hence, estimates of effect that mix different levels of severity can be misleading, or at least less informative than estimates of effect that are specified according to accident or injury severity. This is a relevant aspect of study quality.

As far as construct validity is concerned, its relevance to road safety evaluation studies is limited. It is, however, not entirely irrelevant. Theoretical concepts, like driving skills, are relevant in some evaluation studies. The adequacy of the operational definition of the concept is then one aspect of study quality. This aspect of study quality can be included as an element of the possibility of giving a theoretical explanation of study findings (criterion 7 for causality). A specific assessment of construct validity beyond this is not regarded as necessary.

External validity refers to the possibility of generalising the findings of a study to other contexts than those in which the study was performed. This is certainly highly relevant when applying meta-analysis to summarise studies reported in different countries during an extended period of time. External validity can then be tested for statistically, and it is present when the findings of studies reported at different times in different countries do not vary significantly. Assessing the external validity of a single study does not make sense. This aspect of validity is therefore not relevant in developing a quality scoring system intended for use in assessing individual studies.

As a preliminary summary, this leaves the following list of aspects of study quality that ought to be included in a quality assessment tool for road safety evaluation studies:

1. Process by which the study sample was obtained
2. Specification of accident or injury severity
3. Power to detect the presence of a statistical association between road safety measure and outcome variables
4. Strength of statistical association between road safety measure and outcome variables
5. Within-study consistency of statistical association between road safety measure and outcome variables
6. Clarity of causal direction
7. Control for confounding factors
8. Specification of causal mechanism

9. Theoretical explanation of study findings
10. Possibility of detecting the presence of a dose-response pattern
11. Specificity of effect to target group

The next section reviews methodological research that can serve as a basis for rating the importance of these aspects of study quality.

7.3 A review of methodological research regarding road safety evaluation studies

The objective of this section is to determine, as exhaustively as possible, how various confounding factors may influence the results of road safety studies employing different study designs. This will be done by presenting studies that have estimated the effects of confounding factors in a way that enables comparison of the effects attributed to these factors to the effects attributed to road safety measures.

A general approach to such studies is presented by Hirst, Mountain and Maher (2004). They show how an observed change in the count of accidents in before-and-after studies can be decomposed into the following contributing factors:

1. Regression to the mean
2. Long-term trends
3. Local exogenous changes in traffic volume
4. Road safety measure

This approach can only be applied in its most rigorous form when a before-and-after study employs the empirical Bayes design. It is, however, a very useful approach, because it can address multiple potential confounding factors at the same time. Ideally speaking, a methodological study ought to identify the contributions of all relevant confounding factors, which, for before-and-after studies, includes the introduction of other road safety measures and accident migration in addition to the factors listed above. Unfortunately, most methodological studies have addressed just a single or a few of the confounding factors that are relevant to a certain study design. This makes any conclusion regarding the relative importance of the confounding factors very uncertain. It is important to bear this caveat in mind when reading the following review. The review will be structured in terms of study design, as different confounding factors are relevant for different study designs.

7.3.1 Experimental study designs

As mentioned before, few road safety evaluation studies are experimental. The few there are, tend to rely on small samples. A matched pair study relying on a small sample does not guarantee that the test group and the control are equivalent. A study by Basile (1962) is a case in point.

Basile evaluated the effects of pavement edge markings. To set up the study, 29 matched pairs of road sections were formed. The sections forming each pair were matched by:

1. Geographical proximity: the sections were adjacent to each other
2. Length: sections had the same, or nearly the same length
3. Traffic volume: AADT was the same or nearly so
4. Road width: paved width between 20 and 26 feet on all sections
5. Shoulder width: turf shoulders between 1 and 6 feet on all sections
6. Surroundings: should be uniformly rural

Within each pair one section was chosen by chance for edge markings, the other section remained unmarked. Basile provides a table showing that the sections were successfully matched by road width and shoulder width. However, there were differences with respect to traffic volume. Table 12 summarises these differences and the main results of the experiment.

As can be seen from Table 12, both traffic volume (million vehicle miles of travel), the number of accidents, and accident rate (accidents per million vehicle miles) were higher on control sections than on treated sections. Basile regarded the two groups as successfully matched and estimated the effect of edge lines in terms of the odds ratio:

Table 12: Matched pair experiment with edge lines in Kansas. Source: Basile 1962

	Treated sections			Control sections		
Data	**Before**	**After**	**Change (%)**	**Before**	**After**	**Change (%)**
Million vehicle miles	99.65	99.65		108.38	108.38	
Accidents	166	146	−12 %	200	173	−14 %
Accident rate	1.67	1.47	−12 %	1.85	1.60	−14 %

$$\text{Effect of edge lines} = \frac{\left(\frac{146}{166}\right)}{\left(\frac{173}{200}\right)} = 1.017 = 2\ \%\ \text{accident increase}$$

This, however, is not the only reasonable use and interpretation of the data. Since the control sections had a higher accident rate than the treated sections, one could argue that the observed reduction for the control sections is attributable to regression-to-the-mean. If the accident rates observed in the after-period are regarded as unbiased in both groups, an alternative estimate of effect would be the ratio of accident rates: 1.47/1.60 = 0.918 = 8 % accident reduction. A third option would be to disregard the control group altogether and use the accident reduction in the treated group as an estimate of effect. In that case, the effect is estimated to 12 % accident reduction.

The point is that when matching is not successful, an experimental study design does not necessarily offer any advantage in terms of control for potentially confounding factors compared to a non-experimental design. In the study by Basile, it cannot be ruled out that the higher accident rate in the control group was associated with regression-to-the-mean that introduced bias in the estimate of effect, despite the fact that allocation to the two experimental conditions was random.

An even starker example is given by Peltola (2000). 147 matched pairs of road sections were formed and one section in each pair selected randomly to have a winter speed limit of 80 km/h. The other section retained a speed limit of 100 km/h. Despite the large number of pairs, matching was unsuccessful. For sections that had winter speed limits from 1989, the mean accident rate in the before-period was 56 % higher than the accident rate on control roads, leading to a large, uncontrolled regression-to-the-mean effect.

A good example of a very successful experiment is an evaluation of periodic motor vehicle inspection in Norway (Fosser 1992). A total of 204,000 cars were randomly assigned to three experimental conditions: annual inspection, inspection once during three years and no inspection. Tests found complete pre-trial equivalence between these groups with respect to insurance coverage, annual driving distance, age of car owner, gender of car owner and percentage of cars that changed owner.

The few experimental studies that have been made do not provide an adequate basis for quantifying the importance of the potential confounding factors that have been identified. Based on the examples given above, it is clear that unsuccessful matching in studies employing a matched pair design can introduce substantial bias.

7.3.2 Before-and-after studies

As noted in Table 10, the most important confounding factors in before-and-after studies include:

1. Regression-to-the-mean
2. Long-term trends
3. Exogenous changes in traffic volume
4. Co-incident events (simultaneous to the introduction of the measure)
5. Introduction of multiple measures
6. Accident migration

A number of studies have tried to determine the effects of these potentially confounding factors, including Persaud (1987), Levine, Golob and Recker (1988), Mountain Jarrett and Fawaz (1995), Odberg (1996; re-analysed by Elvik), Elvik (1997, 2002A, 2008A), Grendstad et al. (2003; re-analysed by Elvik), Amundsen and Elvik (2004), Hirst, Mountain and Maher (2004), Mountain, Hirst and Maher (2005) and Persaud and Lyon (2007).

The results are highly mixed. With respect to regression-to-the-mean, often believed to be the most important potentially confounding factor in before-and-after studies, the estimated mean effect is an accident reduction of 6 %

(percentage points). The maximum estimate is an accident reduction of 30 %. The direction of the bias is not consistent. In some studies, regression upwards from abnormally low accident counts has been found. It is clear that lack of control for regression-to-the-mean can seriously bias a study, but neither the magnitude nor the direction of the bias can be reliably predicted based on the experience gained in the studies quoted above.

The mean effect attributed to long-term trends is an accident reduction of 4 %. The maximum effect attributed to this potentially confounding factor is a 17 % accident reduction. Neither the size nor the direction of bias can be reliably predicted on the basis of historical experience.

The potential bias attributable to changes in traffic volume is smaller. The mean estimate is an accident reduction of 3 %, the maximum bias observed in the studies quoted above is 5 %.

As far as the use of multiple measures is concerned, the review by Elvik (2008A) suggests that not controlling for this inflates the estimate of effect by about 13 %, on the average, increasing the effect attributed to a road safety measure from an accident reduction of 24 % to an accident reduction of 37 %. The direction of the bias is unambiguous, but the magnitude depends on how many road safety measures that have been introduced. If the main measure has been supplemented by one other measure, the bias is 6 %. If more than four other measures have been introduced, the bias is 26 %.

Accident migration was a hotly debated topic around 1985-1990, when there was a spate of papers dealing with it. Since then, accident migration appears to have been more or less forgotten and it is rarely discussed in recently published road safety evaluation studies. In the studies quoted above, the mean bias attributable to not controlling for accident migration was to exaggerate the effect of the road safety measure by 15 %. The maximum bias was 27 %.

Thus, it has been found that all the potentially confounding factors may actually confound a study. The bias can be substantial. The maximum bias found in the studies quoted above was 30 % for regression-to-the-mean, 17 % for long-term trends, 5 % for changes in traffic volume, 26 % for the use of multiple measures and 27 % for accident migration. But this does not mean that every study that fails to control for any of these confounding factors is therefore biased to this extent. The average bias is considerably smaller and its direction is not consistent. However, the effects of the confounding factors are sometimes larger than the effects of the road safety measure.

It is not clear how best to apply the results of this review to answer the question: which confounding factor is the most important and should therefore carry the greatest weight in assessing study quality? The answer to this question is as uninformative as one could imagine: Sometimes A is the most important confounding factor, sometimes it is B, and at other times it is C. Sometimes it is neither A, B or C, but D. Sometimes the bias implies an exaggeration of the effect attributed to a road safety measure, sometimes it implies an underestimate. In short, the guidance it was hoped that the methodological studies would give, is almost nil. The results of these studies cannot readily be invoked to justify a particular set of weights assigned to the confounding factors when assessing study quality.

7.3.3 Cross-section studies

The term cross-section study refers to studies that compare the safety of various elements of the transport system, relying on cross-section data and not employing a multivariate accident model. This study design is rarely used today, but used to be very common. The two most common areas of application were in evaluating highway design elements and driver training. In general, cross-section studies compare roads or drivers in terms of the accident rate and study how the accident rate is influenced by a number of characteristics of the road or the drivers.

There are numerous problems associated with this study design, rendering the findings of such studies almost un-interpretable. In fact, the problems are so severe that continued use of the design should be discouraged. In table 10, the following potentially confounding factors associated with cross-section studies were identified:

1. Self-selection to treatment
2. Endogeneity of treatment
3. Differences in traffic volume or annual driving distance
4. Differences in traffic composition
5. Differences in other relevant risk factors

In this section, examples will be given of how each of these factors may confound study findings.

Bias due to self-selection has been found in studies of driver training to which students volunteer and in studies of driver health regulations. An example is given by Harrington (1972). He compared the accident records of drivers who chose to take high-school driver education to those who did not choose to take high-school driver education. Some of his main findings are reproduced in Table 13.

Table 13: Self-selection bias in driver training. Based on Harrington (1972)

	Males			Females		
Outcomes	**Training**	**No training**	**Difference**	**Training**	**No training**	**Difference**
Accidents per driver; not adjusted	0.135	0.176	− 23 %	0.086	0.108	− 20 %
Accidents per million miles; not adjusted	31.834	37.862	− 16 %	17.901	22.805	− 22 %
Accidents per million miles; adjusted for self-selection bias	34.273	33.862	+ 1 %	19.167	20.877	− 8 %

It will be seen that adjusting for self-selection bias removes the entire difference in accident rates among males, and most of it among females. Harrington comments on this finding as follows:

> *"It will be noted that the biographical characteristics of those taking driver training are those associated with better accident and conviction records. … The question remains as to whether or not the better driver record was caused by the driver training, or was merely a consequence of the pre-existing personal differences between the groups. Such a question could only be given a definitive, conclusive answer by repeated experiments in which subjects were randomly assigned to take or not take driver training. The present research is quasi-experimental, or ex-post-facto type of research in which naturally occurring groups are studied. … Analysis of covariance is used here to adjust the driver record differences by taking into account the volunteer bias, so that the resulting adjusted means represent the effect of the driver training with the volunteer bias removed."*

Unfortunately, not all researchers have shown the same keen awareness of this problem as Harrington did, and many studies that failed to control for self-selection bias can be found in the literature.

While self-selection bias is a threat in studies using human subjects, endogeneity bias is a threat in studies comparing accident rates on different types of road. As an example, Muskaug (1985) found the following injury accident rates per million vehicle kilometres of driving on roads that had different speed limits:

Urban centres (speed limit 50 km/h)	0.61
Speed limit 50 km/h	0.51
Speed limit 60 km/h	0.40
Speed limit 70 km/h	0.28
Speed limit 80 or 90 km/h	0.25

In other words, the higher the speed limit, the lower the accident rate. Actual driving speed is closely related to the speed limit. If taken at face value, these data suggest that the higher the speed, the safer the traffic. However, exactly the opposite is true (Elvik, Christensen and Amundsen 2004). The accident rate is endogenous to the speed limit, i.e. speed limits have been lowered on roads that have a high accident rate, but the accident rate continues to be high despite the lowering of speed limits.

In general, endogeneity bias is rampant in cross-section studies and is, possibly, the single most important factor that make the results of these studies close to un-interpretable. As noted above, the most common dependent variable in cross-section studies is the accident rate, which is not an unbiased estimator of safety if the number of accidents is not a linear function of traffic volume or traffic composition (Hauer 1995). A study that may be affected by bias attributable both to non-linearity with respect to traffic volume and traffic composition is an early evaluation of bicycle tracks by Jørgensen and Rabani (1969). Table 14 gives some key data from the study.

Table 14: Comparison of safety on urban streets in Copenhagen with and without bicycle tracks. Derived from Jørgensen and Rabani 1969

Characteristics	With bicycle track	Without bicycle track	Difference
Road length (km)	5.4	3.9	
Road width (m)	15.6	15.4	
Number of side roads	85	83	
Side roads per km	15.7	21.3	
AADT – motor vehicles (1966)	23,000	22,500	
AADT – bicycles (1966)	6,500	10,000	
Motor vehicle accidents (1965-67)	409	411	
Bicycle accidents (1965-67)	107	200	
Motor vehicle accident rate	3.01	4.28	−30 %
Bicycle accident rate	2.78	4.68	−41 %
Total accident rate	2.96	4.40	−33 %

The accident rate – estimated per million motor vehicle kilometres for motor vehicle accidents and per million bicycle kilometres for bicycle accidents – was lower on roads that have bicycle tracks than on roads that do not have such tracks. On the basis of this comparison, Jørgensen and Rabani (1969) concluded that bicycle tracks contribute to improving road safety.

It is not obvious that this conclusion is correct. Subsequent research, reviewed by Elvik (2008C), shows that the risk of bicycle accidents is highly non-linear. Thus, the difference in cycle volume between roads that have bicycle tracks and roads that do not have them would, by itself, be expected to generate a difference in accident rate. If the expected rate of bicycle accidents is estimated using an exponent of 0.7 for motor vehicle volume and an exponent of 0.5 for cycle volume, the expected accident rate on roads with a bicycle track is found to be about 18 % lower than on roads without such a track. This alone accounts for almost half the observed difference in the rate of bicycle accidents between the two groups of roads.

Roads with bicycle track had fewer junctions per kilometre than roads without bicycle track. If bicycle accident rate is estimated separately for junctions and sections between junctions, it is found to be 1.87 per million cycle kilometres in roads with bicycle tracks and 2.37 per million cycle kilometres on roads without bicycle tracks. If roads without bicycle tracks had the same number of junctions per kilometre as roads with such tracks, the number of bicycle accidents would have been reduced by 13.5 %. Thus, differences with respect to traffic composition and the number of junctions are found to reduce bicycle accident rate by 29 % on roads with bicycle tracks. Thus, the true effect of these tracks is at most 16 %, rather than the 41 % presented in the study.

In a similar vein, consider the results presented in Figure 6. Figure 6 presents the relationship between paved road width and injury accident rate for national roads in Norway (Nordtyp-projektgruppen 1980). In the report, the thick curve in the middle of the figure is highlighted and presented as the main result. It shows a consistent decline in accident rate as road width increases.

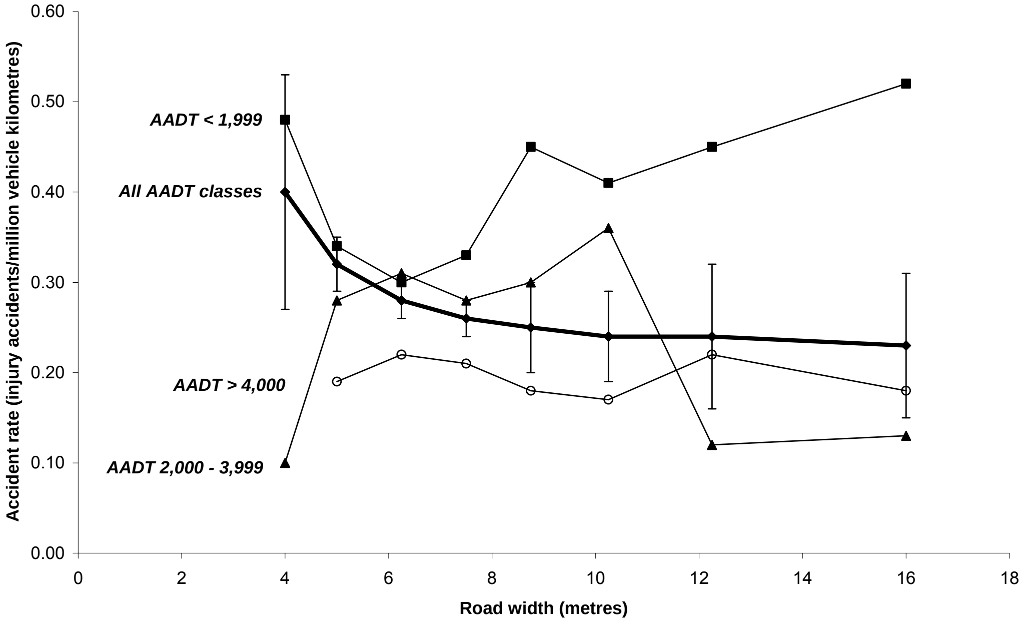

Figure 6: Relationship between paved road width and injury accident rate. Based on Nordtyp-projektgruppen 1980

The thick curve shows the simple bivariate relationship between road width and accident rate. This relationship is confounded by numerous confounding factors. The three other curves shown in Figure 6 refer to roads with AADT less than 2,000, roads with AADT between 2,000 and 3,999 and roads with AADT 4,000 or more. As can be seen, even the very crude control for traffic volume introduced by these three classes makes the relationship between road width and accident rate vanish completely. The three curves representing roads with different traffic volume in Figure 6 show no consistent relationship between road width and accident rate. These curves fluctuate erratically around a flat line. In other words, the relationship indicated by the thick curve is entirely spurious and can be fully explained in terms of the correlation between traffic volume and road width and the fact that the number of accidents does not increase in strict proportion to traffic volume.

As a final example of the pitfalls of cross-section studies, consider the study of salting of roads by Vaa (1995), analysed in greater detail by Elvik (2003A). Vaa estimated that salting roads during winter reduced the accident rate by 26 %. He used the following accident rate ratio to obtain this effect:

$$\text{Accident rate ratio} = \frac{\left(\dfrac{\text{Accident rate on salted roads in winter}}{\text{Accident rate on unsalted toads in winter}}\right)}{\left(\dfrac{\text{Accident rate on salted roads in summer}}{\text{Accident rate on unsalted roads in summer}}\right)} = \frac{\left(\dfrac{0.163}{0.204}\right)}{\left(\dfrac{0.158}{0.147}\right)}$$

As noted by Elvik (2003A), this simple comparison of salted and unsalted roads is misleading, as the two groups differed substantially in terms of a number of characteristics that influence safety. Mean AADT for salted roads was 8,000; mean AADT for unsalted roads was 2,900. The distribution of roads by speed limit was very different. If these differences are adjusted for, the effect attributed to salting vanishes completely.

7.3.4 Case-control studies

Case-control studies in road safety evaluation have mostly been applied to study the effects of measures designed to reduce injury severity. Examples include guard rails, seat belts and crash helmets. In these studies cases will be killed or injured road users; controls will often also be injured road users, and the groups will be compared with respect to injury severity as a function of their exposure to the road safety measure. The general layout of results will be a 2 x 2 table as shown below:

		Seat belt worn	
		Yes	No
Injury	Yes	A	C
	No	B	D

Effects will be stated in terms of the odds ratio: (A/B)/(C/D). For this to indicate the effects of a measure, all other factors influencing injury severity should ideally speaking be controlled for. Thus, the most important potentially confounding factors in case-control studies will be:

1. Differences between cases and controls with respect to factors influencing injury severity
2. Differences between cases and controls with respect to prognostic factors (i.e. factors influencing the probability of survival, given an injury)
3. Differences in the treatments applied to various groups among the controls.

There has not been much methodological research trying to determine how the level of control for these confounding factors may influence the results of case-control studies. A recent discussion in Accident Analysis and Prevention (Curnow 2003, 2005, 2006, 2007; Hagel and Pless 2006; Cummings et al. 2006) raised issues concerning the use of case-control studies to evaluate the effects of bicycle helmets. Curnow argued that a number of case-control studies that have evaluated the effects of bicycle helmets were flawed by: (1) Not using outcome variables that show how bicycle helmets influence serious brain injury, i.e. the outcome variables used were too general to identify the specific injury outcomes that Curnow argued ought to be the main targets for prevention by means of helmets; (2) Not using an appropriate sample of controls, i.e. controls were in most studies injured bicyclists seeking treatment for other injuries (not head injuries), whereas Curnow argued that controls ought to have been a sample of bicyclists representative of the population of bicyclists on the road; (3) Not distinguishing between different types of helmets, as hard shell helmets are likely to be more effective than soft shell helmets.

Hagel and Pless (2006) and Cummings et al. (2006) responded to the criticism put forward by Curnow. The most detailed response was given by Cummings et al. (2006). They reject Curnow's criticism as relying on a misconception of case-control studies and argue that using injured bicyclists as control is entirely appropriate. They do, however, admit that perfect control for confounding factors is difficult in case-control studies. It is not the purpose of this report to take sides in the debate, but it does illustrate that the results of case-control studies may be influenced by choice of outcome variable, choice of controls, method for controlling for confounding factors, and the definition of the road safety measure whose effects is evaluated.

It can be argued that the most stringent version of case-control studies applied in road safety is the double-pair comparison method developed by Evans (1986A) and applied by him to evaluate the effects of seat belts (Evans 1986B) and helmets for motorcyclists (Evans and Frick 1988). Evans argues that the double-pair comparison method essentially controls for all confounding factors, since cases and controls were involved in the same accident – hence, factors like car size, impact speed, impact angle, etc, are identical for cases and controls. Relevant differences that need to be controlled for may, however, remain. Case subjects and control subjects may, for example, differ in terms of age and gender. In most applications of the double-pair comparison method, differences in age and gender have been controlled for by stratifying the data into homogeneous subsets. The disadvantage of this approach is that it rapidly exhausts data.

In studies evaluating the effects of seat belts relying on a case-control design, there are very many potentially confounding variables, for example:

1. Seating position (driver, front seat passenger, rear seat passenger)
2. Impact speed
3. Car size (large cars offer an advantage)
4. Age of car (new cars protect occupants better)
5. Type of accident
6. Age of car occupant
7. Presence of other people in the car

Table 15 compares the mean effect attributed to seat belts for drivers in studies that controlled for less than 3, between 3 and 5 and more than 5 potentially confounding variables.

Table 15: Association between number of potentially confounding variables controlled for and effects attributed to seat belts in case-control studies

	Percentage change in risk of injury by number of potentially confounding variables controlled for		
Injury severity	**Less than 3**	**3 to 5**	**More than 5**
Fatal	−50 %	−54 %	−51 %
Serious	−52 %	−48 %	− 3 %
Slight	−11 %	−25 %	+55 %

While the estimates of effect with respect to fatal injury are hardly influenced by the number of potentially confounding variables controlled for, this is not the case with respect to serious and slight injury. Large differences are found, and there is a tendency for studies controlling for a large number of potentially confounding variables to attribute smaller effects to seat belts than studies controlling for fewer potentially confounding variables.

For some road safety measures, differences in prognostic factors between cases and controls may confound study results. A case in point is a study by Cunningham et al. (1997) comparing survival among injury victims transported to hospital by ground ambulance or helicopter. The objective of the study was to evaluate the effects of helicopter transport on patient survival. Among patients transported by means of ground ambulance, 960 out of 11,765 patients died (8.2 % mortality rate). Among patients transported by means of helicopter, 297 out of 1856 died (16.0 % mortality rate). Thus based on crude mortality rates, transport by means of helicopter is associated with an odds ratio of 1.96 of dying, suggesting that transport by means of helicopter has no benefit, but is, on the contrary, associated with almost a doubling of the probability of dying.

Closer examination of the data reveal that the effects of helicopter transport are moderated by baseline survival probability. If survival among those transported by ground ambulance is taken as baseline, the relationship between baseline probability of survival and the effect associated with helicopter transport is shown in Figure 7.

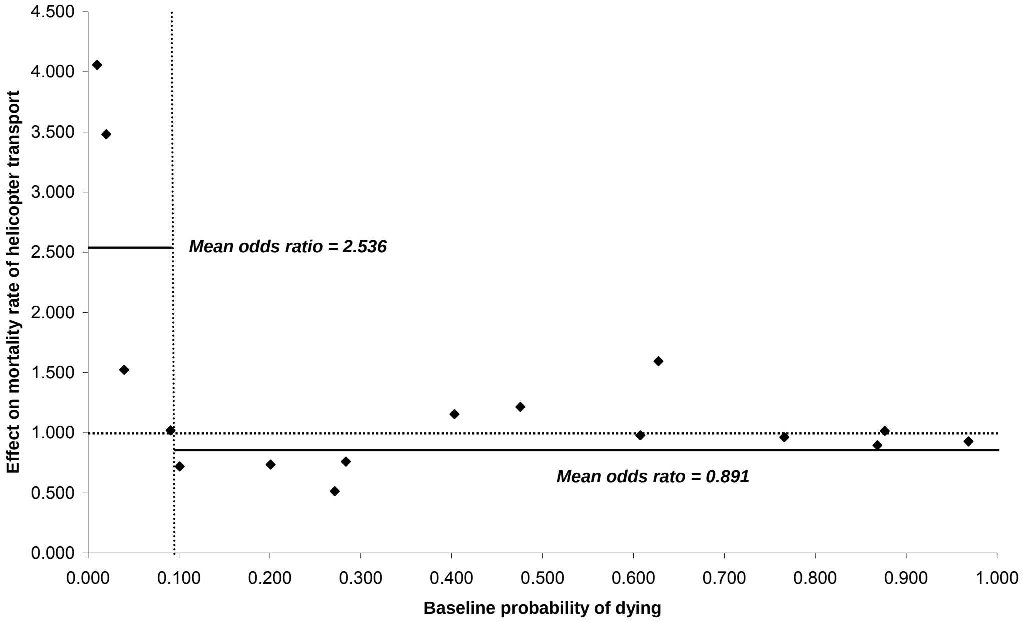

Figure 7: Relationship between baseline probability of survival and effects of helicopter transport in trauma patients. Based on Cunningham et al. 1997

If helicopter transport did not have an effect, all data points would be aligned along the dotted horizontal line at the value of 1.0. However, if baseline mortality is higher than about 0.10, there is a small net benefit of helicopter transport, corresponding to an odds ratio of 0.891. If baseline mortality is less than 0.10, helicopter transport appears to have an adverse effect. Reasons for this are not known. The overall odds ratio, adjusted for baseline probability of survival, is 1.19. Even this estimate is, however, misleading as most of the data points, representing 72.4 % of all observations, indicate a net benefit of helicopter transport.

The final potentially confounding factor to be discussed is the definition of the safety measure. Cummings et al. (2006) present data that indicate that the effect of any helmet worn by a bicyclist on any head injury is 0.314 (odds ratio = 69 % injury reduction). The corresponding odds ratios for different types of helmets were 0.293 for hard shell helmets, 0.304 for soft shell helmets and 0.386 for helmets with no shell. Thus, mixing different types of helmets may produce somewhat misleading estimates of effect.

7.3.5 Multivariate accident models

The following potentially confounding variables have been identified for studies employing multivariate accident models:

1. Endogeneity of treatment
2. Wrong functional form of independent variables
3. Collinearity among explanatory variables
4. Omitted variable bias

5. Wrong specification of residual terms
6. Mixing levels of accident severity
7. Inappropriate model form
8. Inappropriate choice of dependent variable

The potential influence of these confounding variables on study findings has been discussed at length by Elvik (2007A). The full details of that discussion will not be repeated here. The main findings were as follows.

Endogeneity bias is a potentially very serious source of bias in studies that include one or more road safety measures as explanatory variables. This bias arises because the dependent variable of the model – usually the number of accidents – is one of the factors, sometimes the most important factor, that influences introduction of the road safety measure. To give a hypothetical, but by no means entirely unrealistic example: suppose road lighting is introduced at roads that have a high number of accidents in darkness. Suppose further, that on such a road, the number of accidents is reduced from 16 to 12. On an otherwise identical road without lighting, the number of accidents in darkness is 10. In a cross-section study employing multivariate modelling, the model estimate may well indicate an adverse effect of road lighting, when in fact the opposite is true. For a striking example of such bias, see a paper by Kim and Washington (2006).

A wrong functional form describing the effect of an explanatory variable can also introduce bias. It is not always clear if a specific functional form is correct or not. Figure 8 shows an example of a case in which at least one functional form can be considered erroneous.

The size of the dots indicate the number of accidents underlying each data point. A linear function has been fitted to the data as well as a curvilinear function. If the linear function is extrapolated, it will produce nonsensical estimates of accident reductions exceeding 100 %. This does not apply to the curvilinear function, which is therefore the appropriate functional form in this case. In general, choosing the appropriate functional form in accident modelling is a topic where more research is needed (see Hauer 2004).

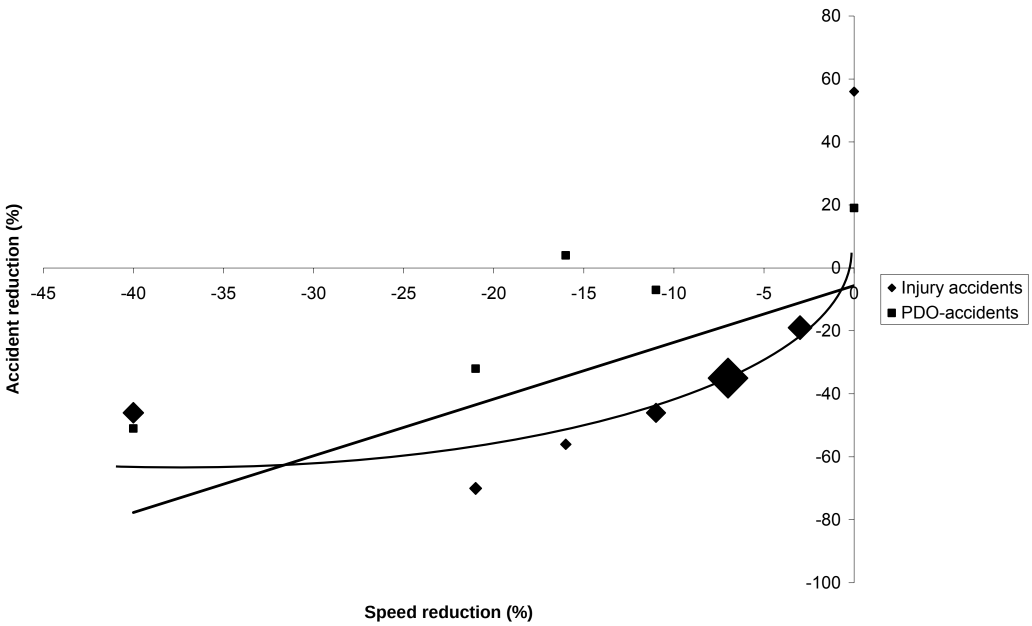

Figure 8: Relationship between changes in mean speed and changes in the number of accidents in environmental streets

Collinearity among explanatory variables is a thorny problem in multivariate analysis. Collinearity denotes a high degree of correlation, which makes it difficult to reliably estimate the effect of a particular variable, controlling for all other variables included in the model. It is tempting to try to solve this problem by omitting one of the highly correlated variables, but this can often be a pseudo-solution, giving rise to a different kind of bias, which is omitted variable bias. A very striking example of collinearity is found in a study of the effects of quantified road safety targets (Elvik 2001C). A dummy variable identifying the United States correlated 0.989 with the dependent variable in the model, which was the number of road accident fatalities. Thus, inclusion of the United States in the model made it virtually impossible to estimate the effects of any other variables. The only possible solution was to omit the United States from the study. This meant not just dropping an explanatory variable, but leaving out all records pertaining to the United States.

An illustration of how omitting a variable may influence study results and generate bias is given by Jonsson (2005). He developed models to predict the number of accidents involving pedestrians or cyclists. He compared model coefficients for models that included bicyclist or pedestrian volume and models that did not include these variables. In the model that included both motor vehicle and bicyclist volume, the coefficient for motor vehicles was 0.76 and the coefficient for bicyclist volume was 0.35. When bicyclist volume was omitted, the coefficient for motor vehicle volume changed to 0.93. Similarly, for pedestrian volume, the coefficients when it was included were 0.83 for motor vehicle volume and 0.38 for pedestrian volume. When pedestrian volume was omitted, the coefficient for motor vehicle volume changed to 0.92.

The bias created by omitting a relevant variable can be substantial. Using the coefficients estimated by Jonsson, one can estimate that an increase in the volume of motor vehicles from 5,000 to 10,000 on a road that has 500 bicyclists per day will be associated with an increase in the number of accidents involving bicyclists of 69 % if the coefficients of 0.76 and 0.35 are applied. If, however, only the coefficient of 0.93 for motor vehicle volume is applied, the increase in the number of accidents can be estimated to 91 %.

How can we know if a model is afflicted by omitted variable bias? The answer is that we can never know this for certain. Even a model that has a very high explanatory power may be biased due to omitted variables, since any omitted variables could be correlated both with the variables included in the model and the residual term of the model.

Possibly the most common form of omitted variable bias in current accident prediction models is the incompleteness of exposure data. These data rarely include pedestrian or cyclist exposure.

As far as the residual terms of accident models is concerned, these are often specified as either normal, Poisson or negative binomial. Few models have compared these specifications. Khan, Shanmugam and Hoeschen (1999) compared Poisson, negative binomial and log-normal models for Interstate-25 in Colorado. In most cases, the Poisson models performed best, as assessed by the authors. However, the difference between Poisson and negative binomial models was in most cases very small. The errors made in choosing between these very common specifications of residual terms were minor. It should be added, however, that the authors of this study did not compare model fit in terms of conventional criteria, like the over-dispersion parameter.

Mixing levels of accident severity can introduce bias in accident models. A case in point is a model developed to identify hazardous road sections in Norway (Ragnøy, Christensen and Elvik 2002). Unlike most other models, this model used the number of injured road users as dependent variable. Table 16 shows estimated coefficients for traffic volume (AADT) and a speed limit of 90 km/h for various levels of injury severity as well as for all injury accidents.

Table 16: Coefficients for traffic volume and speed limit 90 km/h in accident prediction model fitted for national roads in Norway. Based on Ragnøy et al. 2002

Variables	Fatal	Very serious	Serious	Slight	All injury
AADT	0.842	0.829	0.809	0.972	0.923
Speed limit 90 km/h	0.090	0.025	-0.850	-0.743	-1.105

As can be seen, the coefficients vary depending on injury severity. Using a coefficient that mixes different levels of severity would therefore produce misleading results.

The choice of model form may influence study results. In recent years, dual state models have become more common in accident modelling. A dual state model is a model based on the idea that the accident generating process has two states: a normal state and a deviant state, in which the expected number of accidents differs from that produced in the normal state. A common form of dual state model has been zero-inflated Poisson regression models, or zero-inflated negative binomial regression models. These models posit a zero-risk state, in which the expected number of accidents is zero. However, as shown by Lord, Washington and Ivan (2005), zero-inflated models can produce results that are entirely artefacts. Use of such models must therefore be regarded as a methodological weakness of a study.

Finally, with respect to the choice of dependent variable in accident modelling, the count of accidents is to be preferred to the accident rate. By using the accident rate as dependent variable, one is assuming that there is a linear relationship between traffic volume and the number of accidents. This is in general not a valid assumption and may generate misleading findings.

7.3.6 Time-series analysis

Very few methodological studies of time-series analysis have been found. The study of Quddus (2008) was mentioned above. Despite the lack of methodological studies, it is not difficult to find examples of how different interpretations of a time-series can give rise to grossly different estimates of effect for road safety measures. A striking example is given in Figure 9. The figure is based on a paper by Phillips and McCutchen (1991), evaluating the effects of deregulation of the motor carrier industry in the United States in 1980.

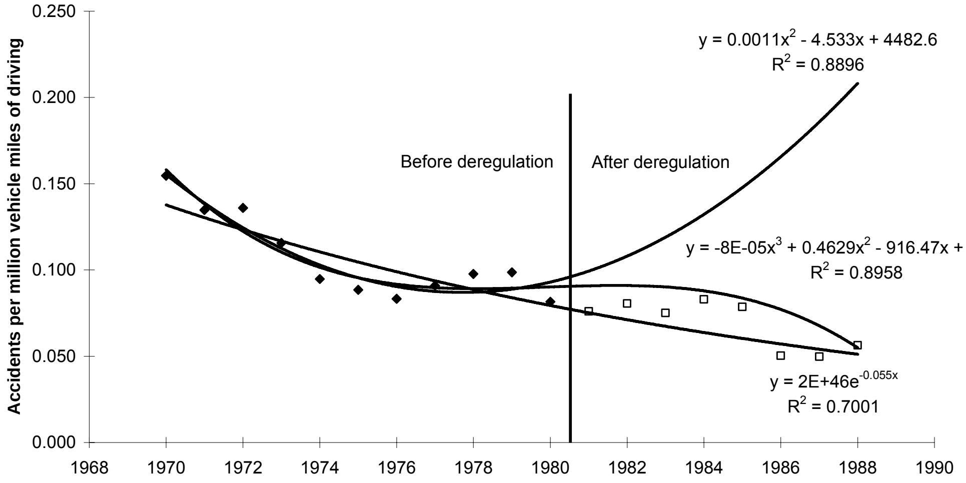

Figure 9: Three projections of future accident rates for motor carriers in the United States. Based on Phillips and McCutchen 1991

The three projections are very different, although they all fit the data quite well and track each other closely during the before-period. It may be objected that in most time-series analyses, there will be considerably more data points in the before-period than in Figure 9, and the analyses will not simply consist of fitting a polynomial function to the data, but accounting for trends, seasonality, lagged effects, step functions, etc. This may well be correct, but it does not refute the basic points that can be made by reference to Figure 9, which are that:

1. For any time series, it is likely that more than one model will adequately fit the time series. Projecting one model to the after-period may give very different predictions from those obtained by projecting another model to the after-period.
2. Projections based on a function fitted to a time series for the before-period for the treated units (persons, locations) fail to establish the counterfactual, i.e. they do not answer the question: what would have happened if the road safety measure had not been introduced.

To establish the counterfactual, it is necessary to show that projections based on the before-period accurately predict what would have happened if the road safety measure had not been introduced. This can be done in three ways:

1. One may divide the before-period into two periods, fit a function to the first of these and test how well that function predicts observations in the remaining part of the before-period.
2. One may use a comparison time series, shown to be equivalent to the case time series, and use projections made for the comparison time series to establish the counterfactual for the case time series.
3. One may apply a multivariate technique of analysis, trying to account for as many factors influencing the time series as possible. This approach is sometimes referred to as structural time series modelling (Harvey and Durbin 1986).

Comparatively few time-series try to establish the counterfactual. Most such analyses erroneously believe that projecting the before-period series amounts to showing what would have happened if the road safety measure had not been introduced.

7.4 Lessons from methodological research

The objective of this chapter was to develop a rational foundation for quality scoring of studies, thereby avoiding the arbitrariness that characterises most quality scales that have been proposed so far. A rational foundation for quality scoring exists if there is knowledge showing how and to what extent various sources of error and bias can distort study findings. If sources of bias A, B, and C are relevant for a specific study design, and it is known that the potential bias attributable to source A is considerably greater than the potential bias attributable to source B, which is in turn greater than the potential bias attributable to source C, then one could justify a quality scoring system assigning greater weight to item A than to item B and greater weight to item B than to item C.

Unfortunately, the methodological research reviewed in this chapter is as inconclusive as all other research designed to support the development of quality scoring systems that has been reviewed in other chapters of this report. It was hoped, for example, that studies evaluating how various confounding factors can distort before-and-after studies would consistently find that, say, regression-to-the-mean is the most important source of bias and should therefore be assigned the largest weight in a quality scoring system.

The results of methodological studies are not at all consistent. In some cases, not controlling for a potential confounding factor did not influence study findings at all. In other cases, it influenced study findings greatly, but the direction of bias was not consistent. Even such a well-known potentially confounding factor as regression-to-the-mean does not always influence before-and-after studies, and when it does, the direction of bias could go both ways. It is therefore totally unjustified to use previous studies as evidence for stating, for example, that: "Not controlling for regression-to-the-mean has been found to be the greatest source of bias in before-and-after studies. It always inflates the estimate of effect, and the bias is always between 20 % and 30 %."

The main lessons learnt in this review of methodological studies can be summarised in the following points:

1. Arbitrariness in quality scoring systems is a very serious problem. It is important to reduce the element of arbitrariness in study quality scoring. One way of doing so, is to study the extent to which various sources of error and bias may distort study findings. It is reasonable to assign greater weight to factors that may seriously bias a study than to factors that have less potential for introducing bias.
2. A review has been made of studies that have assessed the importance of various sources of error and bias in road safety evaluation studies. These studies are incomplete and have not covered all relevant study designs exhaustively. The designs best covered by methodological studies are experimental designs, before-and-after studies and multivariate accident modelling.
3. The findings of methodological studies are inconsistent and do not provide an adequate basis for assessing the relative importance of various potential sources of error and bias in road safety evaluation studies. Moreover, this research has not addressed all potentially confounding factors that were identified in Chapter 6.
4. It is not possible to use methodological studies as a basis for selecting items in a quality scoring system and assigning weights to these items.

8 A formal quality scoring system for road safety evaluation studies

8.1 Framework for the system

This chapter proposes a formal quality scoring system for road safety evaluation studies. The framework for the system is the typology of study designs and potentially confounding factors relevant to each study design developed in Chapter 6. The quality scoring system needs to be customised to each study design, as the threats to validity differ between study designs.

The aim of Chapters 3 through 7 of this report was to develop a scientific basis for study quality scoring. This research effort has by and large been unsuccessful. More specifically, it has been found that:

- Previously published quality scoring instruments were mostly developed in medicine and contain few, if any, items that are relevant for assessing the quality of road safety evaluation studies. Moreover, the validity of most of these instruments has not been meaningfully tested and there is an extremely great diversity of items included.
- There is no consensus among leading road safety researchers about the concept of study quality as applied to road safety evaluation studies. Different researchers emphasise different aspects of study design and analysis when asked to specify the characteristics of a good study.
- A pilot version of a quality scoring system for road safety evaluation studies was developed and its reliability tested. Although the system had an acceptable degree of reliability, no way of testing the validity of the system was found. Besides, the system was quite crude, did not cover all relevant items and appeared to have poor discriminative power, tending to assign very similar quality scores to studies that were informally judged to be of different quality.
- The validity of a quality scoring system can be defined as its inclusiveness, meaning that a valid system is one that covers all aspects of study quality. In order to identify relevant items for a valid quality scoring system, a typology of study designs and threats to validity of these designs was developed. This typology is quite extensive as many varieties of study design are used and the lists of potential sources of bias are long. In fact, the lists of potential sources of bias had to be limited to those that, on the basis of previous research, are judged to be the most important.
- In an attempt to justify assigning weights to the different items of a quality scoring system, methodological studies assessing the bias various

> confounding factors can generate were reviewed. The review was inconclusive. There are no regularities with respect to the effects on study results of uncontrolled confounding factors. Therefore, studying how confounding factors may influence study results does not provide an adequate basis for determining the weights to be given to these factors when assessing study quality.

It would therefore seem that a large element of arbitrariness is unavoidable in any formal quality scoring system. This is a serious shortcoming of any such system, and some would regard it as a decisive objection. The position taken in this report is that arbitrariness should not be regarded as a decisive problem as long as any quality scale is treated as provisional and subject to revision. Moreover, it must be recognised that no global quality scale, applicable to any study design, can be developed. Any quality scale has limited general value – indeed it will sometimes be necessary to develop a quality scale for use in a single literature review only.

8.2 The scoring system

8.2.1 Is there a need for ranking study designs?

A preliminary ranking of study designs was proposed in Chapter 6. The challenge is to assign numerical values to this ranking. This clearly cannot be done in a non-arbitrary way. There is no way of knowing how much better an experimental study is compared to a non-experimental study. Twice as good? Three times as good? Ten times as good? Any answer is subjective and impossible to justify by reference to research.

To keep the element of arbitrariness in quality scoring at a minimum, it is therefore desirable not to rank study designs. As mentioned before, a study employing a certain study design is not always of better quality than a well-conducted study employing a different design that would normally be regarded as inferior. A matched-pair experiment, in which matching was unsuccessful and resulted in uncontrolled regression-to-the-mean, should be rated lower for quality than a well-conducted empirical Bayes before-and-after study.

Hence, no attempt is made to rank study designs. However, whenever scoring a study for quality, the first item scored is study design. Once study designs have been chosen, the subsequent steps are: (1) to score a study according to items that are common to all study designs, and (2) to score a study according to items that are specific to a certain study design.

8.2.2 Items common to all study designs

In chapter 7, eleven items that ought to be included in a quality scoring system were listed. Ten of these items are common to all study designs. The eleventh item, control for confounding factors, is specific to each study design, as the relevant potentially confounding factors differ from one design to the other.

Table 17 lists items that are common to all study designs and proposes how to score these items. The first item is sampling technique. A distinction is made between four techniques: (1) Studies based on the entire population of interest; (2) Studies relying on random sampling from a known sampling frame, which also includes cluster sampling, (3) Studies using non-random sampling according to criteria that are stated explicitly, and (4) Convenience samples. Ordinal scores have been assigned to these four techniques. As convenience samples are very common in road safety evaluation studies, this sampling technique has been assigned a score of 1, rather than 0, so as not to omit otherwise well-conducted studies. Depending on whether scores are added or multiplied – an issue to which we will return – a score of 0 for a single item may lead to an overall quality score of 0, effectively assigning zero weight to a study.

The second item is specification of accident or injury severity. This item is relevant, because the effects of many road safety measures have been found to vary according to accident or injury severity. A study that does not specify effects according to accident or injury severity will fail to detect such variation, if present. A distinction is made between five levels for the specification of accident or injury severity, and ordinal scores have been assigned to these five levels.

The presence of a statistical association between a road safety measure and one or more outcome variables is the first condition for inferring a causal relationship. The direction of the effect (increase or reduction of the number of accidents) and its denomination (accidents, odds ratios, correlations, etc.) are irrelevant for assessing study quality. Thus, an effect can be stated as a percentage change in the number of accidents, an accident rate ratio, an odds ratio, a regression coefficient, or any other measure of statistical association. A statistical association is deemed to be present if a statistically significant change in road safety is associated with the introduction of the road safety measure. However, a study should not be rated lower for quality simply because it fails to find an effect of a measure. Quality scoring should not depend on study findings. The criterion is therefore stated as the possibility of detecting an effect in a study, not the presence of an effect.

This criterion is related to study power. It might therefore be regarded as superfluous, since statistical weights that depend on the number of accidents are assigned to all study findings in meta-analyses, and the power to detect an effect depends on the number of accidents. Thus, statistical weight can be interpreted as an indicator of study power. It is, however, the possibility of detecting an effect of practical interest that counts. Very often, even small effects are of practical interest and may decide the cost-effectiveness of a measure. Thus, as an example, many road markings are cost-effective even if their effects are as small as 2-5 % reduction of the number of accidents, lasting for only a few years. Clearly, detecting an effect of this size is a challenge, and many studies will therefore be inconclusive as to whether there really is an effect of practical interest or not.

Thus, to assess whether a study can detect an effect of practical interest, it is necessary to know at least roughly what the costs of a measure are and on that basis try to assess what the smallest effect of practical interest would be. Case illustrations given later in the report will show the application of this concept.

Table 17: Scoring road safety evaluation studies with respect to items common to all study designs

Item	Levels of item	Score assigned
Sampling technique	Study of entire population of interest	4
	Random sampling from known sampling frame	3
	Non-random sampling; criteria stated	2
	Convenience sample	1
Specification of accident or injury severity	Fatal, serious, slight injury and property damage only specified	4
	Fatal, injury and property damage only specified	3
	Fatal, serious injury and slight injury specified	3
	Injury (including fatal) and property damage only specified	2
	Study limited to injury accidents; severity not further specified	1
	Accident or injury severity not specified; mixing of levels probable	0
Statistical association between measure and outcome variable	Detecting an effect of practical interest possible	1
	Detecting an effect of practical interest not possible	0
Strength of statistical association between measure and outcome variable	Comparison of effect size with other effects present in data is possible	1
	Comparison of effect size with other effects not possible	0
Consistency of statistical association between measure and outcome variable	Consistency of association across subsets of data can be assessed	1
	Consistency of association across subsets cannot be assessed	0
Clarity of causal direction between measure and outcome variable	Causal direction can be determined and is in the right direction	1
	Causal direction cannot be determined or is in wrong direction	0
Specification of causal mechanism (causal chain) generating effects	Causal mechanism fully specified and evaluated empirically	3
	Causal mechanism partly specified and evaluated empirically	2
	Causal mechanism specified, but not evaluated empirically	1
	No causal mechanism discussed or identified	0
Possibility of explaining study findings theoretically	A well-established theory exists that may explain study findings	1
	No well-established theory exists that may explain study findings	0
Presence of dose-response pattern	Study design allows for assessing a dose-response pattern	1
	Study design does not allow for assessing a dose-response pattern	0
Specificity of effect to target group for intervention	Study design allows for assessing specificity of effect	1
	Study design does not allow for assessing specificity of effect	0

It is conceivable, however, as noted by Elvik (2007B), that a causal relationship may exist even if it fails to produce a direct statistical association between a road safety measure and the number of accidents. The presence of a statistical association is, by itself, therefore not a decisive criterion of causality. However, when a statistical association is not found, this indicates that other causal factors may be more important than the road safety measure whose effects a study is intended to evaluate.

A strong statistical association is often regarded as more likely to be causal than a weak statistical association. To assess the strength of a statistical association, it can be compared to other effects that have been estimated in a study. When assessing study quality, it is not feasible to compare the strength of all statistical associations estimated in a study.

Besides, the number of effects that have been estimated in a study will vary considerably. In a simple before-and-after study, it may be that the only effect that has been estimated statistically is the effect attributed to the road safety measure. In a multivariate analysis, it may be that more 20 effects have been estimated. As far as scoring studies for quality is concerned, comparing the strengths of statistical associations found in a study is therefore a dichotomous variable. Studies are coded as either permitting such comparisons to be made or not permitting it.

Causal relationships are generally held to be stable and consistent. A given cause always produces the same effect in the same context, within the bounds of random variation and the precision of measurement instruments. Thus, a study permitting an assessment of the consistency of the statistical association between a road safety measure and outcome variables is of better quality than an otherwise identical study not permitting such an assessment. For the purpose of quality scoring, this characteristic has been coded as a binary variable: either a study permits the consistency of a statistical relationship in subsets of the data, or across model specifications, to be assessed, or it does not permit such an assessment.

The possibility of determining causal direction unequivocally depends partly on study design. In particular, one of the criteria – that the cause should precede the effect in time – will not be applicable to cross-section studies in which the data do not refer to changes over time. In these studies, the possibility of determining causal direction will depend on whether the study has controlled adequately for endogeneity bias or not. However, the temporal order between variables may not necessarily be clear even in before-and-after studies. Suppose that at time T a high number of accidents is recorded in a junction. This lead to safety treatment at time T+1. Following this, the number of accidents goes down. The reduction of the number of accidents may, however, be purely regression-to-the-mean, which means that causality is, in a way, reversed: it was the abnormal number of accidents before treatment that caused the treatment to be applied and that caused the number of accidents to go down, not the treatment that was applied after observing the high number of accidents.

In practice, therefore, it is rarely possible to determine causal direction without considering how well a study has controlled for potentially confounding factors. For assessing quality, a simple binary assessment is proposed: either causal direction can be determined, or it cannot.

The specification of a causal mechanism that generates – or conceals – a statistical association between treatment and effect is often regarded as useful when trying to answer the question: why did this treatment have an effect, or why did it not have an effect? All road safety measures are intended to influence road safety by modifying or one more risk factors that are associated with accident occurrence or injury severity. To specify a causal mechanism is to identify these factors and measure changes in them. It must be recognised, however, that specifying and measuring a causal mechanism will not always make the findings of a study easier to interpret. Knowing a causal mechanism may help in interpreting a study, but will not necessarily do so. For the purpose of assessing study quality, a distinction has been made between three levels: (2) A causal mechanism has been specified and evaluated empirically, (1) A causal mechanism has been suggested, but not evaluated empirically, and (0) No causal mechanism has been identified or discussed.

The possibility of explaining study findings theoretically is related to how well known causal mechanisms are, but in principle it is not necessary to know causal mechanisms in detail in every road safety evaluation study if these mechanisms are sufficiently known in terms of well-established theory. It is, unfortunately, rarely the case that well-established theory can be applied to explain the findings of road safety evaluation studies, but some relationships may be regarded as sufficiently well-established to serve this function. Examples include:

- Biomechanics explains why properly fastened seat belts reduce injury severity. A finding to the contrary is theoretically very implausible.
- Biomechanics and laws of physics explain why properly installed guardrails reduce injury severity. Again, a finding to the contrary would be hard to explain theoretically.
- Laws of physics and energy dissipation explain why lower speed is associated with less serious accidents, possibly also fewer accidents. While findings that are inconsistent with this can be imagined (such as a high number of accidents occurring at very low speeds in parking lots and parking garages), these findings would not imply that the laws of physics and energy dissipation are false.
- Optics and ophthalmology can explain why road lighting improves safety at night. Findings that are inconsistent with this cannot be ruled out, but such findings cannot be explained in terms of human visual functions.

In short, while none of these relationships are sufficiently well-established to entirely rule out findings that appear to counter them, they are sufficiently well-established to add plausibility to the findings of evaluation studies, thus increasing the likelihood that these findings do represent causal relationships. For assessing study quality, a distinction is made between cases in which a well-established theory exists that can explain study findings, and cases in which no such theory exists.

In some studies, it is possible to probe for a dose-response pattern, long regarded as a key indicator of causality in epidemiology. When it is possible to probe for a dose-response pattern, the study earns an additional point for quality. The same applies when it is possible to test for the specificity of an effect to a target group. Studies that permit such a test earn an extra point for quality.

8.2.3 Items specific to each study design

The lists of potentially confounding factors are specific to each study design, although some of the potentially confounding factors are common to several study designs. Table 18 lists the confounding factors by study design and shows how studies may be assessed with respect to their level of control for potentially confounding factors.

Table 18 is very extensive, reflecting the large number of potentially confounding factors that may influence the results of road safety evaluation studies. It is proposed to score each item by means of an ordinal scale (later to be converted to the 0,1 range; see below) that has at least two levels – in some cases up to four levels. For all items, the following principle has been applied in defining levels of the ordinal scale:

1. The highest score is assigned when *it can be positively shown* that a potentially confounding factor *did not confound a study*. This requirement will be regarded as fulfilled when a study has controlled for a potentially confounding factor by means of techniques generally recognised as appropriate by the scientific community.
2. The lowest score is assigned when a potentially confounding factor was not controlled for. In some cases, additional information may be available to show whether confounding actually did occur or not. In general, however, *the potential for confounding is a sufficient reason for controlling for the potentially confounding factor*.

In some cases, a more fine-grained assessment is possible. Thus, as an example, there are several techniques for controlling for regression-to-the-mean, and some of these are better than others (Elvik 2008B). Short comments will be given with respect to the scoring proposed for some of the items listed in Table 18.

As noted in the examples discussed in Chapter 7, violation of pre-trial equivalence threatens the validity of experiments. Ideally speaking, an experiment ought to demonstrate pre-trial equivalence. However, this is rarely done, and in its absence, the assumption can be made that if the samples assigned to different experimental conditions were all large, systematic differences between them are unlikely to be present. Reliance on sampling theory is nevertheless rated lower than an actual demonstration of equivalence between samples assigned to different experimental conditions.

Table 18: Scoring road safety evaluation studies with respect to control for confounding factors

Study design	Potentially confounding factors	Level of control for confounding factors	Score assigned
Experimental designs	Pre-trial non-equivalence	Pre-trial equivalence tested for and confirmed	4
		Large sample makes pre-trial equivalence likely (not tested for)	3
		Pre-trial equivalence violated; differences adjusted for statistically	2
		Pre-trial equivalence violated; no adjustment for differences	1
	Diffusion of treatment to control group	Treatment implementation monitored and no diffusion found	3
		Diffusion present; statistical adjustment implemented	2
		Diffusion present; no adjustment possible	1
	Differential attrition between groups	Attrition monitored; no differences found	3
		Differential attrition present; statistical adjustment made	2
		Differential attrition present; no adjustment possible	1
	Unintended side-effects of experiment (e.g. Hawthorne effects)	No evidence of unintended effects and no reason to suspect them	4
		Evidence of unintended effects of experiment; adjustment possible	3
		No evidence of unintended effects, but reason to suspect them	2
		Evidence of unintended effects; adjustment not possible	1
Before-and-after studies	Regression-to-the-mean	Empirical Bayes model-based control for regression-to-the-mean	3
		Control for regression-to-the-mean by means of simpler techniques	2
		No control for regression-to-the-mean	1
	Long-term trends	Long-term trends controlled for (comparison group or time-series)	2
		Long-term trends not controlled for	1
	Exogenous changes in traffic volume	Local exogenous changes in traffic volume controlled for	2
		Local exogenous changes in traffic volume not controlled for	1
	Co-incident events	No co-incident events known to have occurred	2
		Co-incident events occurred	1
	Introduction of multiple measures	The use of multiple measures known and controlled for	2
		Use of multiple measures not known or not controlled for	1
	Accident migration	Accident migration identified and controlled for	3
		Accident migration is judged not likely to occur	2
		Accident migration is judged likely, but was not controlled for	1

Table 18: Scoring road safety evaluation studies with respect to control for confounding factors, continued

Study design	Potentially confounding factors	Level of control for confounding factors	Score assigned
Cross-section studies	Self-selection of subjects to treatment	Self-selection found not to be present; selection is close to random	4
		Self-selection known and adjusted for statistically or by matching	3
		Self-selection not positively known, but suspected	2
		Self-selection known to occur; no adjustment for it	1
	Endogeneity of treatment	Endogeneity of treatment assessed and found not to be present	4
		Endogeneity present and adjusted for statistically	3
		Endogeneity suspected; inconclusive evidence; no adjustment	2
		Treatment known to be endogenous; no control for it	1
	Differences in traffic volume	Differences in traffic volume adjusted for by multivariate model	3
		Differences in traffic volume adjusted for using accident rates	2
		Differences in traffic volume not controlled for	1
	Differences in traffic composition	Differences in traffic composition controlled for	2
		Differences in traffic composition not controlled for	1
	Differences with respect to any other relevant risk factor	Control for multiple risk factors known to influence safety	3
		Control for some (but not all) risk factors known to influence safety	2
		Control for few or no risk factors known to influence safety	1
Case-control studies	Non-equivalence of cases and controls with respect to accident severity	Matched pair design ensuring equivalence with respect to risk factors	4
		Statistical control for multiple risk factors	3
		Stratification or statistical control for some, but not all, risk factors	2
		Control for few or no risk factors	1
	Non-equivalence of cases and controls with respect to prognostic factors	Matched pair design ensuring equivalence with respect to risk factors	4
		Statistical control for multiple risk factors	3
		Stratification or statistical control for some, but not all, risk factors	2
		Control for few or no risk factors	1
	Heterogeneity of treatment	Cases receiving different treatments are not mixed	2
		Cases receiving different treatments mixed	1

Table 18: Scoring road safety evaluation studies with respect to control for confounding factors, continued

Study design	**Potentially confounding factors**	**Level of control for confounding factors**	**Score assigned**
Multivariate analyses	Endogeneity of treatment	Treatment shown not to be endogenous; selection close to random	4
		Treatment is endogenous; appropriate statistical control applied	3
		Endogeneity suspected; no conclusive evidence provided	2
		Endogeneity documented; no correction for it applied	1
	Wrong functional form of explanatory variables	Functional form explicitly chosen and shown to be best	3
		Functional form chosen by default; by standard model specification	2
		Implausible functional form; strange or non-logical implications	1
	Collinearity among explanatory variables	Collinearity shown not to be a problem	3
		Insufficient information to assess if collinearity is a problem	2
		Collinearity suspected or shown to be a problem	1
	Omitted variable bias	No omitted variables can be identified	3
		Insufficient information to assess potential omitted variable bias	2
		Omitted variables bias suspected or shown to be present	1
	Erroneous specification of residual terms	Reasonable specification of residual terms adopted	2
		Questionable specification of residual terns	1
	Inappropriate model form	Single-state or plausible dual-state model used	2
		Theoretically implausible dual-state model used	1
	Inappropriate choice of dependent variable	Number of accidents used as dependent variable	2
		Accident rate (linear) used as dependent variable	1
Time-series analysis	Inadequate adjustment for explanatory variables	Factors influencing time-series identified and adjusted for	2
		Intervention analysis only	1
	Co-incident events	No such events; or co-incident events identified and controlled for	2
		Intervention analysis only	1
	Erroneous specification of residual terms	Residual terms appropriately specified	2
		Inappropriate specification of residual terms	1

Accident migration, a potentially confounding factor in before-and-after studies (akin to treatment diffusion in experiments), is scored differently from other potentially confounding factors. The reason for this is that considerably less is known about how often and to what extent accident migration actually introduces confounding in before-and-after studies than what is known regarding the other potentially confounding factors. While confounding by regression-to-the-mean and long-term trends has often been found in before-and-after studies, accident migration remains of a more hypothetical nature. Hence, the assessment is more conditional than for the other potentially confounding factors and includes an assessment of the likelihood of confounding, not just the potential for it.

As far as multivariate analyses are concerned, one of the potentially confounding factors identified in Chapter 6 – mixing levels of accident severity – is omitted from Table 18 as it is a potentially confounding factor relevant to all study designs. This factor is therefore listed in Table 17.

Having identified the items to be included in quality scoring, the next issue to be discussed is the relative importance of these items.

8.2.4 Assigning weights to items included in quality scoring

The items listed in Tables 17 and 18 can be placed in four groups with respect to their contributions in assessing whether a road safety measure is causally related to outcome variables:

1. Group 1 consists of items that describe the statistical association between a road safety measure and its effects (presence, strength and consistency of statistical association),
2. Group 2 consists of other general criteria of causality (direction of causality, causal mechanism known, theoretical foundation, dose-response pattern, specificity of effect),
3. Group 3 consists of the item-specific potentially confounding factors, listed in Table 15 (control for confounding),
4. Group 4 consists of items describing the potential for generalisation and application of study findings (sampling technique, specification of accident or injury severity).

It has been argued – and numerous examples can be given in support – that control for confounding is the single most important aspect of study quality for road safety evaluation studies. Based on this, the following weights are proposed for items in the four groups (weights sum to 1):

1. Items in group 1 (statistical association) are assigned a weight of: 0.12
2. Items in group 2 (criteria of causality) are assigned a weight of: 0.30
3. Items in group 3 (control for confounding) are assigned a weight of: 0.50
4. Items in group 4 (external validity) are assigned a weight of: 0.08

Thus, control for confounding alone counts as much as all the other items combined. The next question to be asked is if it possible to assign weights to the potential confounding factors listed in Table 18. It was hoped that methodological studies would have produced knowledge with respect to the size of bias that may occur as a result of not controlling for a specific potentially confounding factor. The idea was that factors with a potential for generating a large bias if left uncontrolled should carry greater weight than factors with a smaller potential for bias. However, methodological studies turned out to provide little guidance in this respect. Hence, the simplest solution would be to give all potentially confounding factors the same weight.

On the other hand, it is not likely to be the case that all potentially confounding factors are equally important. There are, for example, clear indications that endogeneity bias in multivariate analyses is a considerably more potent source of error than erroneous specification of residual terms. It would be wrong to assign the same weight to two items, when one of them has been found to have a potential for generating a very much greater bias than the other. Therefore, despite the lack of complete and unambiguous evidence from methodological studies, the various potentially confounding factors have been weighted differentially. The weights proposed are listed in Table 19.

Comments will be given regarding the proposed weighting of the potentially confounding factors. Lack of pre-trial equivalence and unintended side effects have been rated as the most important potentially confounding factors in experiments. The bias that may result from lack of pre-trial equivalence is well documented. As far as unintended effects of an experiment are concerned, there are fewer examples in road safety evaluation studies. Some studies using driving simulators have induced unintended simulator sickness.

Regression-to-the-mean and long-term trends have been classified as the most important potential confounding variables in before-and-after studies. In particular, regression-to-the-mean can be very large in driver accident data (Hauer and Persaud 1983).

As far as cross-section studies are concerned, all potentially confounding factors have been rated as equally important, as no basis has been found to justify a different assessment. In case-control studies, lack of control of factors influencing injury severity has been rated as the most important potentially confounding factor. Endogeneity of treatment can give very misleading results in multivariate studies and is therefore regarded as more important than the other potentially confounding factors. In time-series, the basic flaw of most analyses is failure to establish the counterfactual condition.

Table 19: Weights assigned to potentially confounding factors when assessing the quality of road safety evaluation studies

Study design	Confounding factors	Weight assigned
Experiments	Pre-trial equivalence violated	0.40
	Diffusion of treatment to control group	0.20
	Differential attrition between groups	0.10
	Unintended side-effects of experiment	0.30
Before-and-after	Regression-to-the-mean	0.40
	Long-term trends	0.30
	Exogenous changes in traffic volume	0.10
	Co-incident events	0.05
	Introduction of multiple measures	0.10
	Accident migration	0.05
Cross-section studies	Self-selection of subjects to treatment	0.20
	Endogeneity of treatment	0.20
	Differences in traffic volume	0.20
	Differences in traffic composition	0.20
	Differences with respect to any other relevant risk factor	0.20
Case-control studies	Non-equivalence of cases and controls with respect to accident severity	0.60
	Non-equivalence of cases and controls with respect to prognostic factors	0.20
	Non-equivalence of treatment	0.20
Multivariate models	Endogeneity of treatment	0.40
	Wrong functional form of explanatory variables	0.10
	Collinearity among explanatory variables	0.10
	Omitted variable bias	0.10
	Erroneous specification of residual terms	0.05
	Inappropriate model form	0.10
	Inappropriate choice of dependent variable	0.15
Time-series analysis	Inadequate adjustment for explanatory variables	0.45
	Co-incident events	0.45
	Erroneous specification of residual terms	0.10

8.3 Application of the system – test cases

In this section, the proposed system for assessing study quality will be applied to a number of test cases in order to explore its discriminative power. Testing reliability and validity is not possible at this time. The proposed system should be regarded as a pilot version only; tests of reliability and validity must await more extensive use of the system.

The following studies have been coded by means of the system:

Fosser (1992), an *experimental evaluation* of periodic motor vehicle inspection, which was compared to Christensen and Elvik (2007), a non-experimental evaluation of periodic motor vehicle inspection, employing multivariate analysis.

The purpose of the comparison was to determine if an experimental study scores better for quality than a non-experimental study.

A comparison was made between different *before-and-after studies* that have evaluated the effects of converting junctions to roundabouts: Nygaard (1988); Oslo veivesen (1995); Odberg (1996; re-analysed by Elvik).

Three *cross-section studies* dealing with horizontal curve radius have been assessed (Brüde, Larsson and Thulin 1980; Matthews and Barnes 1988; Sakshaug 1998). Three *case-control studies* evaluating bicycle helmets have been compared (Maimaris et al. 1994; Thompson et al. 1996; Schrøder Hansen et al. 2003). Three *multivariate studies* evaluating protective measures at highway-railroad grade crossings have been compared (Hauer and Persaud 1987; Austin and Carson 2002; Park and Saccomanno 2005). Finally, three *time series analyses* were selected for scoring (Holder and Wagenaar 1994; Hagge and Romanowicz 1996; Wong et al. 2004).

8.3.1 Converting scores to a bounded scale

Before presenting the scores assigned to the studies selected, it is necessary to explain how the ordinal scores assigned to each item have been converted to a bounded scale ranging from 0 to 1. This bounded scale is treated as an approximation to a ratio scale. A perfect study will score 1, a worthless study will score 0. In tables 17 and 18, ordinal scores are assigned to each item. Thus, for example with respect to pre-trial equivalence in experiments, scores are assigned as:

Pre-trial equivalence shown:	4 points
Pre-trial equivalence presumed as samples were large:	3 points
Pre-trial equivalence violated, but statistical adjustment applied:	2 points
Pre-trial equivalence violated, no adjustment for this:	1 point

For each such ordinal scale – with one exception – the highest score has been converted to the value of 1, the lowest score has been converted to the value of 0. Values in-between have been assigned scores by linear interpolation. Thus, the ordinal scale above was converted to: 1.00 (4) – 0.67 (3) – 0.33 (2) – 0.00 (1).

The only exception from this rule applies to sampling technique. Due to the very common use of convenience samples in road safety evaluation studies, the lowest score for sampling technique is 0.25, not 0.00.

Item scores are multiplied by the weight assigned to each item. Then scores are added. The resulting scale has a range between 0 and 1.

8.3.2 Experiment versus multivariate study

The first case concerns the studies of Fosser (1992) and Christensen and Elvik (2007). The study reported by Fosser was scored as follows:

Item	Score assigned	Justification
Sampling technique	Entire population studied	All cars aged 6-8 years included; older cars not included to keep attrition low
Accident severity	Injury and PDO identified	Other levels of accident severity not identified
Detection of effect of practical interest	No	Smaller effects than the smallest effect the study could detect would make the measure cost-effective
Strength of association	Comparable	Several other relationships were assessed in addition to inspections
Consistency of association	Comparable	Different versions of data analysis could be compared
Causal direction	Can be determined	Assignment was at random before first inspection
Causal mechanism	Partly evaluated empirically	Effect of technical condition was assessed; behavioural adaptation not
Item	Score assigned	Justification
Relevant theory	No well-established theory	Findings cannot be explained in terms of well-established theory
Dose-response	Can be assessed	There were three levels for the inspection variable
Specificity of effect	Cannot be assessed	No attempt was made to assess if effects were greater for the most defective cars
Pre-trial equivalence	Documented	Was tested and confirmed
Diffusion of treatment	Not present	Was tested and not found
Differential attrition	Not present	Was tested and not found
Unintended side-effects	Not found	Was tested and not found

Converted to the (0, 1) scale, this study scored 0.835. It got the full score (0.500) for control for confounding and a score of 0.335 (maximum 0.500) for the standard items. Factors that reduced the score from a perfect 1.00 included:

Accident severity was stated only in terms of two levels: injury or property-damage-only. This lead to a loss of 0.025 points. The smallest effects the study could have detected was a reduction of injury accidents by 19 % and a reduction of property-damage-only accidents by 3.7 %. These reductions amount to an annual societal benefit per car of about 1,600 NOK. The cost of one periodic inspection is about 550 NOK. Hence, even smaller effects that those this study could have detected would have been of practical interest. The study was therefore scored as not being able to detect the smallest effect of practical interest, thus losing 0.060 points The causal mechanism was only partly evaluated. This lead to a loss of 0.020 points. There was no well-established theory to explain study findings. This lead to a loss of 0.030 points. The specificity of effect – which in this study would be a larger effect for the most defective cars – was not assessed. This lead to a loss of 0.030 points. Still, this study scored very high. As will be shown later in this chapter, it beats most of the other studies that have been selected for testing the quality scoring system.

The study by Christensen and Elvik (2007) was scored as follows:

Item	Score assigned	Justification
Sampling technique	Convenience sample	Data from an insurance company willing to share them were used
Accident severity	All levels mixed	Accident severity was not stated; most accidents were probably property-damage-only
Detection of effect of practical interest	No	The smallest detectable effect in the study was larger than the smallest effect of practical interest
Strength of association	Comparable	Several other relationships were assessed in addition to inspections
Consistency of association	Comparable	Different versions of data analysis could be compared
Causal direction	Can be determined	Assignment was at random before first inspection
Causal mechanism	Partly evaluated empirically	Effect of technical condition was assessed; behavioural adaptation not
Relevant theory	No well-established theory	Findings cannot be explained in terms of well-established theory
Dose-response	Can be assessed	There were three levels for the inspection variable
Specificity of effect	Cannot be assessed	No attempt was made to assess if effects were greater for the most defective cars
Endogeneity bias	Not present	Cars are selected by age only, not previous accident record
Functional form	Chosen explicitly	Several forms were tested for technical defects
Collinearity	Not clear	Co-variance matrix was inspected, but such an inspection is often inconclusive
Omitted variable bias	Likely to be present	Data referred to car owner, not driver; poor control for driver characteristics
Residual terms	Correctly specified	The negative binomial assumption was reasonable
Model form	Plausible	A single state model is plausible
Dependent variable	Correct	The number of accidents was used as dependent variable

The study by Christensen and Elvik scored 0.713. This is surprisingly high for a non-experimental study. Still, the study scored lower than the study by Fosser with respect both to control for confounding factors (0.425 versus 0.500) and with respect to the standard items (0.288 versus 0.335). The study is, however, an example of a fairly successful multivariate model-based study, which can reasonably be assumed to be free of endogeneity bias and some of the other methodological problems that often plague such studies.

8.3.3 Before-and-after studies

A detailed explanation of how the three before-and-after studies were scored will not be given. Appendix 2 shows the scores assigned to each study. The study by Nygaard (1988) employed a comparison group to control for long-term trends. It did not control for regression-to-the-mean. The study identified the number of legs in junctions and the size of the roundabout, thus permitting an analysis of a potential dose-response pattern related to these variables.

The study by Oslo veivesen (1995) employed the same design as the study by Nygaard, but did not record the number of legs in the junctions nor the size of the roundabout. Finally, the study by Odberg (1996) provided detailed data, permitting a re-analysis by Elvik, employing the empirical Bayes design. In re-analysed form, the study controlled for regression-to-the-mean, long-term trends, local changes in traffic volume and the introduction of other safety measures.

Accident migration was not judged to be a likely impact of converting junctions to roundabouts. It was judged that a theoretical explanation of findings would be possible for studies that identified characteristics of roundabouts that one would expect to be associated with the size of their effect on accidents – in particular number of legs and diameter of the central traffic island.

The quality score was 0.613 for Nygaard (1988), 0.533 for Oslo veivesen (1995) and 0.863 for Odberg (1996) (re-analysed). Nygaard scored 0.375 for standard items and 0.238 for control for confounding factors. Oslo veivesen scored 0.295 for standard items and 0.238 for control for confounding factors. Odberg (re-analysed) scored 0.375 for standard items and 0.488 for control for confounding factors. Thus, the largest differences between the studies were found with respect to control for confounding factors.

8.3.4 Cross-section studies

Appendix 2 shows in detail the scoring of the cross-section studies. Brüde, Larsson and Thulin (1980) evaluated the effects of horizontal curve radius on accident rate. The data were a mixture of injury accidents and property-damage-only accidents. Confounding factors were controlled by a mixture of stratification and restriction (i.e. restricting the study to certain types of accident, omitting other types). The data were stratified by speed limit, road width and vertical alignment. Moreover, accidents at junctions, pedestrian accidents and accidents involving animals were omitted. It was argued that these accidents were unlikely to be influenced by horizontal curve radius. Self-selection and endogeneity was judged not to be relevant for this study.

The study of Matthews and Barnes (1988) was very similar to the study of Brüde, Larsson and Thulin (1980). The data were stratified by various variables, although mostly one at a time, due to the limited amount of data. A single road was studied. It was not clear if accidents included only injury accidents or property-damage-only accidents as well.

Finally, Sakshaug (1998) made a preliminary study, which was subsequently continued as a multivariate modelling study. The initial study, however, relied on a simple design with little or no control for confounding variables.

Brüde, Larsson and Thulin scored 0.615, Matthews and Barnes 0.608 and Sakshaug 0.540. These scores show that all studies were of moderate quality. While closer to 1 than to 0, it is easy to point out methodological weaknesses in all studies. On the other hand, the studies were not so poor as to be altogether inconclusive.

8.3.5 Case-control studies

Appendix 2 shows in detail the scoring of the three case-control studies selected for testing. All three studies evaluated the effects of bicycle helmets. Maimaris et al. (1994) compared the incidence of head injury and other types of injury to cyclists wearing helmets (cases) and cyclists not wearing helmets (controls). Potentially confounding factors were controlled for by multivariate analysis. No distinction was made between different types of helmet.

Thompson et al. (1996) studied the effects of helmets by comparing cyclists presenting at an emergency department for head injury (cases) to cyclists presenting for other types of injury (controls). One might think that this design would generate bias (Curnow 2003, 2005, 2006), since the probability of head injury is correlated with helmet wearing, and the selection of cases is, in a sense, endogenous to the outcome of interest. However, this does not engender bias as long as the selection of controls is unrelated to the road safety measure. If helmets have no effect on the incidence of other injuries than head injuries, this criterion will be fulfilled. According to Maimaris et al. (1994), it is reasonable to assume that bicycle helmets have no effect on other injuries than head (and face) injuries. A drawback in the design adopted by Thompson et al. (1996) is that it does not permit an investigation of the effect of helmets with respect to other injuries than head injuries. Thompson et al. controlled for a number of confounding factors by means of multivariate analysis. They also made a distinction between three types of helmet.

Schrøder-Hansen et al. (2003) also studied a sample of cyclists presenting at an emergency department. Effects of two different types of helmets were compared and confounders were controlled for by means of multivariate analysis. Cases were cyclists who sustained head or face injuries. Two controls were used: either other emergency room patients, or a sample of cyclists in the city of Bergen.

Maimaris et al. scored 0.524, Thompson et al. 0.599 and Schrøder-Hansen et al. 0.604. Thus, all studies got nearly the same score indicating a moderate quality. This is not surprising, as the three studies were quite similar with respect to design and analysis. Besides, getting a very high score for study quality is difficult when employing a case-control design.

8.3.6 Multivariate analyses

Three multivariate analyses have been compared. These studies are remarkable, because unlike very many other multivariate analyses, they show a keen awareness of the potential for endogeneity bias and adopt different approaches for avoiding it. The first, Hauer and Persaud (1987), built on previous work and introduced the model-based empirical Bayes method for evaluating the effects of safety measures. To avoid endogeneity, variables describing safety devices at rail-highway grade crossings were not included in the accident prediction model. A risk of endogeneity bias would otherwise be present to the extent that selection for treatment is based on accident history.

The second study, Austin and Carson (2002) employed the instrumental variable method to control for endogeneity. The accident prediction included considerably more variables than the models fitted by Hauer and Persaud, but the correction for endogeneity appears to have been only partly successful, as at least one safety treatment appeared to be associated with an increased number of accidents, which is somewhat implausible in view of what other studies have found. The third study, by Park and Saccomanno (2005) generated classifying variables by means of a regression tree analysis. Some of the classifying variables were safety devices at rail-highway grade crossings.

Hauer and Persaud scored 0.735 for quality, Austin and Carson 0.669 and Park and Saccomanno 0.724. These scores are similar and are, on the whole, better than the scores obtained by the case-control studies reviewed above. A score of around 0.7 or more shows that the study has an acceptable quality and is more likely to show true effects than mere methodological artefacts.

8.3.7 Time series analyses

Three time series analyses have been selected for quality scoring. Details of the scores assigned are in Appendix 2. Holder and Wagenaar (1994) evaluated the effects of mandated server training on alcohol-involved crashes in Oregon. Server training is training in detecting when a bar patron should not be served more alcohol. A comparison time series was used in addition to the case series. Covariates were included in the analysis. Hagge and Romanowicz (1996) evaluated the effects of California’s commercial driver license program. A comparison time series was used and covariates were included in the analysis. Finally, Wong et al. (2004) tried to assess road safety policies in Hong Kong. No comparison time series was used and the variables intended to describe road safety policy were most likely incomplete.

Holder and Wagenaar scored 0.853, Hagge and Romanowicz scored 0.775 and Wong et al. scored 0.513. This example shows that the quality scoring system is able to discriminate between studies that employ the same design, but differ in terms of important study characteristics related to study quality.

8.3.8 Comparison of findings

Table 20 summarises the findings of the pilot testing of the quality scoring system.

Table 20: Quality scores assigned to studies in pilot testing of quality scoring system

	Quality scores (0 = worst; 1 = best)		
Study design	**Study 1**	**Study 2**	**Study 3**
Experiment	0.835		
Multivariate analysis	0.713		
Before-and-after	0.613	0.533	0.863
Cross-section	0.615	0.608	0.540
Case-control	0.524	0.599	0.604
Multivariate analysis	0.735	0.669	0.724
Time series analysis	0.853	0.775	0.513

As can be seen, the quality scores differ from a low of 0.513 to a high of 0.863. If the quality of research is thought of as a continuum, it may be depicted as in Figure 10.

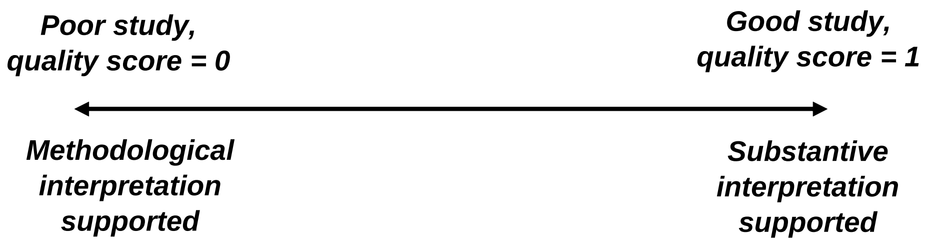
Poor study,
quality score = 0
Good study,
quality score = 1
Methodological interpretation supported
Substantive interpretation supported

Figure 10: Research quality as a continuum

The closer to 1 a study scores for quality, the stronger is the support it gives for concluding that it shows the true effects of a road safety measure. The closer to 0 a study scores, the more likely it is that its findings reflect methodological artefacts only. Can a cutoff be defined? At what score for quality is a study so poor that it should be rejected?

It is tempting to view the various components entering quality scoring as arguments that can be given for and against believing in a study. If the score is less than, say, 0.5, there are stronger arguments against taking it seriously than the arguments in favour of doing so. All the studies scored for quality in this report, scored more than 0.5, but some studies were quite close to that value. The assessment of studies according to quality score could, for example, be as follows:

Studies scoring less than 0.50 (inadequate studies):

Study findings are more likely to reflect methodological weaknesses than the true effect of the road safety measure that has been evaluated.

Studies scoring between 0.50 and 0.599 (weak studies):

The study provides weak evidence for the effects of the road safety measure. Significant methodological shortcomings exist.

Studies scoring between 0.60 and 0.799 (moderately good studies):

Study findings are more likely to show the effects of the safety measure than merely the effects of methodological weaknesses.

Studies scoring between 0.80 and 1.00 (very good studies):

Study findings are clearly more likely to show the true effects of a road safety measure than the effects of methodological factors not adequately addressed by the study.

This sorting is a guideline only. All the studies represented in the examples given here scored more than 0.50 for quality. Is it all conceivable to find a published study scoring less? It certainly is. Gray (1990) reports a simple before-and-after study of driver training in a pharmaceutical company – not controlling for a single confounding factor. Using the quality scoring system presented in this chapter, the study scored 0.131 for quality.

8.4 Summary of main lessons

The main lessons learnt in this chapter can be summarised as follows:

1. It is not possible to develop a formal quality scoring system for road safety evaluation studies that does not have significant elements of arbitrariness. Research designed to minimise the element of arbitrariness by finding a scientific basis for assigning scores and weights to items representing study quality has been unsuccessful. No well-founded scientific basis for quality scoring has been found.
2. A formal quality scoring system for road safety evaluation studies has nevertheless been proposed. The system consists of two parts: a set of standard items that are common to all study designs and a set of items that have been customised to each study design to assess confounding factors that are unique to each design.
3. The quality scoring system scores studies in a bounded range between 0 (poor) and 1 (perfect). The standard items count for half of the total (0.5); control for confounding factors counts for the other half (0.5).
4. The quality scoring system has been tested in a sample of road safety evaluation studies employing different study designs. The system was found to be applicable to all designs and was able to discriminate between good and bad studies. However, a more formal test of validity was not possible.
5. Based on the system, a rough distinction can be made between studies in terms of the strength of the evidence they provide for the effects of road safety measures:
    a. Scores < 0.5: Inadequate studies. The evidence should be rejected on methodological grounds.
    b. Scores 0.50 – 0.59: Weak studies. These studies provide weak evidence and are likely to be influenced by methodological weaknesses.

c. Scores 0.60 – 0.79: Moderately good studies. There are stronger reasons for believing in the studies than for not believing in them, but an influence of methodological weaknesses still cannot be ruled out.

d. Scores > 0.80: Good studies. The evidence is strong and it is unlikely that methodological weaknesses have had a major influence on study findings.

9 The use of quality scores in meta-analysis

9.1 Different approaches to the treatment of study quality in meta-analysis

Six approaches to the treatment of varying study quality in meta-analysis can be distinguished:

1 The "include all studies as reported" approach

According to this approach, no attempt is made to score studies for quality. All relevant studies are included as reported, and their results taken at face value.

This approach dodges the issue and will not be considered further in this report.

2 The "omit bad studies" approach

This approach involves sorting studies into two groups: good and bad. Studies rated as bad are omitted from meta-analysis, which includes just the good studies. To implement this approach, one simply needs to classify studies into those that are good enough to be included in meta-analysis and those that are not good enough. A crude quality scoring system would suffice for this purpose.

This approach will not be considered further in this report, as it is too crude – by disregarding the fact that study quality is a continuous variable – and because determining any cut-off point will be arbitrary.

3 The "sort results by study quality" approach

Studies are classified into a number of groups according to study quality. Meta-analysis is performed in each group, and the results compared between groups. It is then possible to determine whether the results of the analysis are different according to study quality. This approach involves an overall rating of studies by quality (as opposed to item-specific quality scoring; see next section), but this rating can be fairly crude (for example classifying studies as excellent, fair, poor and inadequate).

This approach does not adjust study findings, or estimates of their uncertainty, by study quality. It merely describes the relationship between study quality (according to a crude scale) and study findings. If a relationship is found, the question is what to conclude. Depending on the conclusion drawn, the approach will end up either as a version of approach 1 (include all studies no matter how bad they are) or approach 2 (throw out bad studies). The approach will not be discussed further.

4 The "overall quality score adjustment" approach

An overall quality score is assigned to each study. This overall score could be a function of scores assigned to several components of study quality. The overall quality score, usually normalised to values between 0 and 1, is then used as a variable in a meta regression analysis, which will give estimates of the effect of quality score on study findings. In addition to study quality, meta regression will usually include a number of other factors, for example, publication year, country in which the study was made, and type of institution performing the study.

This approach will be discussed further and illustrated by means of examples.

5 The "item-specific quality score adjustment" approach

Studies are rated for quality on a number of items. The scores assigned to these items are, however, not added up to form an overall quality score for each study. Each item used for quality scoring is retained as a variable in a meta regression analysis, in which study results are statistically adjusted for the effects of other confounding factors as well.

This is the approach advocated by Greenland (1994) and will be discussed further.

6 The "weight studies by quality" approach

This approach involves assigning an overall quality score to each study and using this score as a weight in a meta-analysis, in addition to the ordinary statistical weights. The results of studies can never be more precise than sample size allows for. Accordingly, a formal quality weight would have to take on values between 0 and 1 and always result in a wider uncertainty of summary estimates of effect than sampling variation alone accounts for.

This approach will be discussed further.

9.2 The statistical treatment of quality scores

A meta-analysis is a statistical analysis of set of studies, each of which may contain several estimates of effect, for the purpose of obtaining one or more summary estimates of effect and identify sources of variation in estimates of effect (Elvik 2005A, 2005B). Meta-analysis can be thought of as taking place in three stages: (1) Exploratory analysis, designed to help decide if a meta-analysis makes sense at all; (2) Main analysis, which consists of developing summary estimates of effect and identifying sources of variation in estimates of effect; (3) Sensitivity analysis, which probes how various analytical choices made as part of the main analysis influences its findings.

In analogy to accident modelling, a distinction can be made between various sources of variation in summary estimates of effect developed in meta-analysis. Figure 11 shows the main sources of variation in summary estimates of effect in meta-analysis.

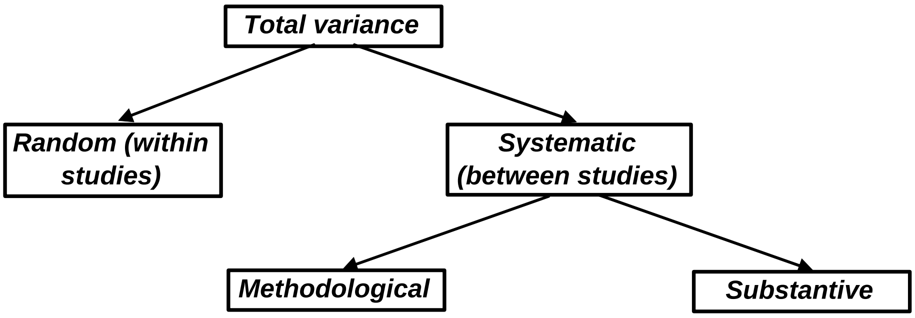

Figure 11: Sources of variation in summary estimates of effect in meta-analysis

Total variance can be decomposed into random variation – or within-studies variation – and systematic variation – or between-studies variation. The latter may in turn be decomposed in methodological variation and substantive variation. Ideally speaking, the statistical treatment of study quality in meta-analysis ought to identify the contribution made to the systematic variation in study findings attributable to methodological and substantive factors.

To illustrate the approaches discussed in section 9.1, a few examples will be used. The first example concerns studies that have evaluated the effects of daytime running lights, summarised by Elvik, Christensen and Fjeld Olsen (2003). The second example refers to studies evaluating the effects of converting junctions to roundabouts, summarised by Elvik (2003B). The third example refers to studies of the relationship between speed and road safety, summarised by Elvik, Christensen and Amundsen (2004).

9.3 Example 1: Daytime running lights

A systematic review of studies that have evaluated the effects of daytime running lights was performed (Elvik, Christensen and Fjeld Olsen 2003). To summarise the findings of these studies, meta-analysis was performed. Meta-regression was used to identify factors that were associated with systematic variation in study findings.

Studies that evaluated the intrinsic effects of daytime running light on cars will be used as an example. By intrinsic effects are meant the effects on the accident rate of each car of using daytime running lights, compared to not using it. There were 13 studies reporting on the intrinsic effects of daytime running lights. Some of these studies reported multiple estimates of effect, referring to different types of accident or different levels of accident severity. For the purpose of the analyses presented here, a single summary estimate of effect for each study will be applied.

If a fixed-effects model (i.e. a model assuming that there is no systematic between-studies variation in effects) of analysis is applied, the summary estimate of effect (accident modification factor) is 0.936 (95 % CI: 0.919; 0.954), corresponding to an accident reduction of slightly more than 6 %. The sum of statistical weights was 10,827.89, meaning that within-study variance was merely 0.00009 (1/10,827.89). A single study contributed to more than 60 % of the total statistical weight assigned to the studies. The (between-studies) variance component was estimated to 0.00511. Total variance thus becomes 0.00009 + 0.00511 = 0.00520; of which 1.8 % is random (within-studies) and 98.2 % is systematic (between-studies). Hence, observed variation in study findings is almost exclusively attributable to between-study variation.

Each study was assigned a quality score ranging from 0 to 1. Details of how this score was defined are given in the original report and will not be repeated here (Elvik, Christensen and Fjeld Olsen 2003). The quality scoring scale was not identical to the one developed in this report. For the thirteen studies that had evaluated the intrinsic effect of daytime running lights on cars, quality score ranged from 0.10 to 0.78, with a simple mean score of 0.47. There was, in other words, a substantial variation in study quality.

Figure 12 shows the simple bivariate relationship between study quality score and estimate of effect. It is seen that, unlike very many other cases, there appears to be a positive relationship: the better the study, the larger its estimate of effects of daytime running lights.

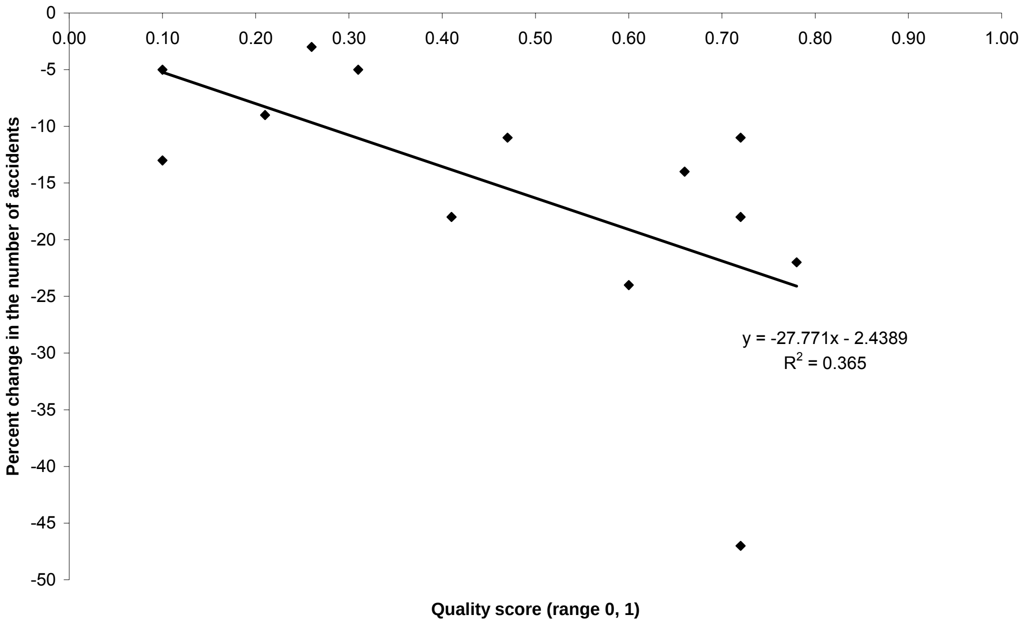

Figure 12. Relationship between study quality and effects attributed to daytime running lights on cars

The relationship is, however, quite noisy. Moreover, in Figure 12, each data point has been assigned an identical statistical weight, which is not correct. In fact, there is a negative relationship between the statistical weight of a study and its quality score: the bigger the study, the lower the quality score.

Two approaches to the treatment of quality in meta-analysis will be illustrated. These are approach 6 (from section 9.1): adjust study weights by quality scores – as proposed in the quality scoring system for the Highway Safety Manual in the United States (Hauer 2007) and approach 4 (from section 9.1): use an overall quality score as a continuous variable in meta-regression. Results obtained by these two approaches will be compared.

The fixed-effects statistical weights assigned to each study were adjusted by multiplying them by quality score:

Quality adjusted weight = Fixed-effects weight · Quality score

This reduced the total statistical weights assigned to the studies from 10,827.59 to 2,894.01. Meta-analysis was applied using the quality-adjusted statistical weights and a summary estimate of effect derived. The quality-adjusted summary estimate of effect was 0.920 (95 % CI: 0.887; 0.954). This does not differ greatly from the fixed-effects summary estimate of effect, but it does adjust the summary estimate of effect in the expected direction, i.e. towards a larger estimate of effect, since the best studies were associated with the greatest estimates of effect.

The between-studies variance was not greatly affected by quality adjusting. The variance component, estimated in order to find the random-effects weights to be assigned to each study, was 0.00511 before adjusting for study quality, 0.00469 after adjusting for study quality. This shows that adjusting for study quality reduces between-study variance by only slightly more than 8 %. If taken at face value, this finding suggests that other sources of between-study variation are more important than varying study quality.

A meta-regression analysis was run, employing three different estimators of effect as dependent variable and a number of potentially explanatory variables including:

1. Study age (0 before 1990; 1 after 1990)
2. Country of origin (0 all other countries; 1 United States)
3. Dummy for fatal accidents (fatal = 1)
4. Dummy for injury accidents (injury = 1)
5. Dummy for property-damage-only accidents (PDO = 1)
6. Dummy for aggregate effects (as opposed to intrinsic; aggregate = 1)
7. Study quality (included as a continuous variable)

Applying the coefficients estimated, the intrinsic effect of daytime running lights for a study published after 1990, not in the United States, applying to injury accidents and having a study quality score of 0.5 (close to the mean of the scores observed) can be estimated to 0.764. If quality score is 0.8 (close to the highest score of the studies represented), the estimated effect is 0.731. If study quality is 0.1 (identical to the lowest scores observed), the summary estimated effect is 0.811. Hence, when adjusting for other variables, the effect of differences in study quality are smaller than suggested by the simple bivariate relationship in Figure 12.

Other variables were found to have a greater influence on study findings than study quality. As an example, if it is assumed that the study was reported in the United States, but had a study quality score of 0.5, the summary estimate of effect becomes 1.080, as opposed to 0.764 if, all else equal, the study was assumed to be made outside the United States. Thus, country of origin changes the summary estimate of effect by nearly 32 percentage points (from 8 % accident increase to 24 % accident reduction), whereas the range in study quality observed between the studies included only generates an 8 percentage point difference in summary estimate of effect (from 27 % accident reduction to 19 % accident reduction).

While meta-regression is, in principle, superior to simply adjusting the weight of each study by means of a quality score it has problems of its own. Variables tend to be co-linear and the estimated coefficients are highly uncertain. Moreover, by combining coefficients, one may generate more or less “hypothetical” estimates of effect that are outside the range of the observations used in fitting the meta-regression model. As an example, one could generate a hypothetical summary estimate of effect for a “perfect” study, i.e. a study scoring 1 for quality. It is not always obvious which combination of coefficients best represent the typical study included in an analysis. These problems are shown in the next example.

9.4 Example 2: Roundabouts

A meta-analysis reported by Elvik (2003B) evaluated the effects of converting junctions to roundabouts. The analysis included 28 studies performed outside the United States. Meta-regression was applied to summarise the findings of these studies.

Study quality was represented by means of a variable describing study design. This variable had the following categories:

1. Before-and-after studies controlling for regression-to-the-mean and long-term trends
2. Before-and-after studies controlling for long-term trends, but not for regression-to-the-mean
3. Before-and-after studies stating traffic volume before and after conversion to roundabout
4. Simple before-and-after studies not controlling for any confounding factors
5. Cross-section studies comparing accident rates in roundabout to accident rates in other types of junctions, controlling for confounding factors by means of stratification.

The first of these designs was regarded as the best. This approach to the treatment of study quality in meta-analysis closely resembles, but is strictly speaking not identical to the item-specific quality adjustment approach presented as approach 5 in section 9.1. With respect to before-and-after studies, it comes very close to being an item-specific approach.

Meta-regression was run and coefficients estimated. Applying these, one may, as an example, estimate the summary effect of converting a three-leg junction previously controlled by a yield sign to a roundabout to:

- 42 % accident reduction if a before-and-after study controlling for regression-to-the-mean and long-term trends was used,
- 56 % accident reduction if a before-and-after study controlling only for long-term trends was used,
- 13 % accident reduction if a before-and-after study providing data on traffic volume was used,
- 42 % accident reduction if a simple before-and-after study was used.

These differences are quite large, but do not systematically indicate that poorer studies are associated with larger estimates of effect. In fact, simple before-and-after studies (the poorest design) produced exactly the same summary estimate of effect as the best controlled before-and-after studies (controlling for regression-to-the-mean and long-term trends).

Five variables were included in the meta-regression analysis (previous type of traffic control, number of legs, size of roundabout, study design, and accident severity). Given the fact that there were 113 estimates of effect in total, using five variables in a meta-regression does not seem excessive. However, the number of combinations of values for the five variables is 400 (2 x 2 x 4 x 5 x 5). This number is well in excess of the number of observations (113) used to fit the multivariate model, indicating that the model could be underdetermined by the data (since not all 400 logically possible combinations of values are found in the data set). Most of the coefficients were indeed not statistically significant, and the coefficient of determination for the final model (not including country variables) was .375 (adjusted R-squared). On top of this, the pattern of covariance between the variables included in the model led to very large confidence intervals for the estimates of effect based on the meta-regression.

The results of meta-regression analysis are, accordingly, not always easy to interpret.

9.5 Example 3: Speed and road accidents

Elvik, Christensen and Amundsen (2004) reviewed 98 studies containing a total of 460 estimates of the relationship between speed and road accidents. Several analyses were performed, including several meta-regression analyses. Study quality was represented by a variable assessing the potential presence of bias in a study due to:

1. Regression-to-the-mean
2. Long-term trends
3. Changes or differences in traffic volume
4. Important risk factors influencing accident occurrence

For each of these factors, an assessment of the likelihood of bias attributable to lack of control for the factor was made. If bias was judged to be likely, a score of “yes” was entered; if bias was judged not to be likely, a score of “no” was entered. The best studies were those that scored no for all four items.

This way of representing study quality permits two approaches to be used to the analysis of how differences in study quality influence summary estimates of effect in meta-analyses. One approach is to treat each of the four factors as one component of study quality and use approach 5 (from section 9.1): an item-specific analysis of components of study quality. The other approach is to summarise the four factors into a count of potential sources of bias, ranging from 0 to 4. This approach corresponds to approach 4 in section 9.1, adjusting for overall study quality. It should be noted that none of the studies reviewed by Elvik, Christensen and Amundsen scored 4 with respect to potential sources of bias. Studies that were judged to be afflicted by all the four potential sources of bias were omitted. Hence, scores ranged from 0 to 3.

Two meta-regression analyses have been selected for illustration. These analyses show how the number of sources of bias influence summary estimates of power for fatal accidents and injury accidents. The results are presented in Table 21.

Table 21: Influence of sources of bias on summary estimates of power in the Power Model of the relationship between speed and road safety. Based on Elvik, Christensen and Amundsen 2004

Number of sources of bias	Power for fatal accidents	Power for injury accidents
0	3.65	2.61
1	4.84	2.64
2	7.75	2.58
3	3.39	3.31

A tendency is seen, albeit somewhat noisy, for the estimates of power to increase as studies are afflicted by more sources of bias.

9.6 The treatment of study quality in meta-analysis

Study quality should be considered explicitly in any meta-analysis. There is no disagreement about this, but different approaches can be taken with respect to how best to integrate an assessment of study quality in meta-analysis. It is important that the approach taken clearly identifies the relationship between study quality and study findings. For a recent, and very striking, illustration, see Erke (2008).

All the three approaches discussed in this chapter are reasonable. While meta-regression relying on components of study quality or an overall score for study quality is by far the most common, adjusting study weights according to study quality is not an approach that should be ruled out. It will be applied in the forthcoming Highway Safety Manual in the United States, and it does make sense from a statistical point of view (Christensen 2003). Adjusting study weights according to quality will, however, not necessarily account for all between-study variation in a set of studies. On the other hand, there may not always be a clear relationship between study quality and study findings. It could be that the findings of poor studies are more variable than the findings of good studies. This will then add an unwanted source of variation to summary estimates in meta-analysis, which can be greatly reduced by assigning weights that adjust for study quality. The poor studies that contribute to artefactual variation in study findings will then be down-weighted and contribute less to the summary estimate of effect.

10 Discussion and conclusions

10.1 Discussion

Road safety evaluation research is a complex field of study. The complexity of the topic is a reflection of the complexity of reality. Road accidents are an extraordinary complex phenomenon, which is influenced by literally hundreds of variables whose contributions are only partly known and in many cases likely to be inaccurately assessed in empirical studies. While it is virtually impossible to perform experimentally designed studies to assess the effects of factors contributing to accidents, randomised controlled trials can sometimes be applied to evaluate the effects of road safety measures. In practice, however, randomised controlled trials are rarely applied and nearly all studies evaluating the effects of road safety measures are observational studies, employing a wide range of study designs and a wide range of approaches to the statistical analysis of the data collected. It is perhaps not surprising that both the quality of these studies and their findings vary considerably.

Some of the complexities facing road safety evaluation research include:

- *Accident reporting* is incomplete and biased; unknown and unrecognised changes in reporting have the potential for seriously biasing evaluation studies – yet adjusting for this bias is next to impossible as no accident record known to be complete exists.

- *Randomness* contributes importantly to variation in the count of accidents when total numbers are low; less so in larger accident samples. Most road safety evaluation studies are based on small accident samples.

- *Systematic variation* in the number of accidents is produced by very many factors, literally hundreds. It is impossible to name and enumerate all these factors; let alone fully and accurately control for their potentially confounding effects in evaluation studies.

- The *effects on accidents of exposure and risk factors* are likely to be non-linear, interactive, partly unobservable and – in general – difficult to model statistically.

Given these complexities, it is hardly surprising that many road safety evaluation studies have been found to be methodologically inadequate. However, it is not merely the inherent complexity of the phenomenon studied that contributes to the sometimes deplorable quality of road safety evaluation studies. Road safety evaluation research is very often done as contract research; sponsoring agencies tend to want the research to be performed as quickly and cheaply as possible. Sometimes research is performed as in-house studies by agencies or individuals who have got a vested interest in the findings. These agencies or individuals may be less inclined to question studies that conclude that the measure was a success than similar studies concluding less favourably.

The great amount and variety of road safety evaluation studies, and the quite detailed information given by some of these studies, can generate a misleading impression that the effects of very many road safety measures are well-known. Care must be taken to steer clear of exaggeration in assessing the current level of knowledge regarding the effects of road safety measures. It would be equally wrong to say that nothing is known as to say that everything we would like to know, is known. Numerical estimates of the effects of a road safety measure can be deceptively precise and detailed; sometimes these estimates unravel almost completely when subjected to critical examination. A case in point is the largely unfounded belief in the effectiveness of treating road accident black spots. The research invoked to bolster this belief collapses almost completely when examined critically, as shown in the example given in Chapter 2.

Given this state of affairs, there ought to be a great interest in developing a practical, easy-to-use, yet reliable and valid instrument for systematically assessing the quality of road safety evaluation studies. Awareness of the problem is widespread, and some efforts have been made to develop instruments for assessing the quality of road safety evaluation studies – notably the effort made as part of the development of the Highway Safety Manual in the United States. However, very few studies have been published whose main purpose was to develop and test a method for assessing the quality of road safety evaluation studies. To be sure, papers can be found that contain some assessment of study quality. But in many cases, this assessment is not the main objective of the paper; rather it is an auxiliary activity needed to accomplish the main purpose of a study.

Moreover, the few examples that can be found of formal instruments for assessing the quality of road safety evaluation studies are all rather crude. Assessment sometimes consists only of classifying studies by design; at best a rough ordinal quality scale containing a few discretionary levels (good, intermediate, bad) is used.

The aim of this study was to develop a more well-founded instrument for assessing the quality of road safety evaluation studies. This task turned out to be considerably more difficult than envisaged at the start of the study. A cogent criticism of previous attempts at developing formal systems for assessing study quality is that any numerical instrument designed score studies for quality is fundamentally arbitrary. In developing such an instrument, choices have to be made with respect to which items to include, how to score each item and how to aggregate scores into an overall score. Unless good reasons can be given for each of these choices, the resulting instrument is arbitrary in the sense that different choices – leading to a different assessment of study quality – could equally well have been made and would have been equally well justified.

Great emphasis has therefore been placed in this study on trying to find ways of minimising the element of arbitrariness in study quality assessment. In an attempt to develop a basis for justifying the choices that must be made in developing a quality scoring instrument, various approaches have been taken, including:

- Studying a sample of 35 quality scoring systems published in research literature,

- Asking a sample of globally leading road safety researchers about their understanding of the concept of study quality and how to assess it,
- Developing a preliminary instrument and conducting a pilot test of it using five researchers scoring five studies for quality independently of each other,
- Developing a typology of study designs employed in road safety evaluation studies and a list of threats to validity relevant for each study design,
- Conducting a survey of methodological studies that have tried to determine how much various confounding factors can influence the results of road safety evaluation studies.

Regrettably, it must be concluded that this research effort was largely unsuccessful and did not produce a very firm basis for justifying the choices that must be made in developing a quality scoring system for road safety evaluation studies.

Given this fact, there are two options about how to proceed. One of them is to give up developing a formal quality scoring system, as there does not seem to exist any way of developing such a system that would not involve a fairly large element of arbitrariness. The other option is to propose a formal quality scoring system, despite the fact that such a system would be somewhat arbitrary and despite the fact that other researchers may legitimately disagree with the system and propose their own, different system. The second option was taken, and a formal quality scoring system proposed. It has to be stressed that the quality scoring system proposed in this report is a first version only. Whether the system makes sense or not, and how well it functions, can only be determined by using and discussing the system. By doing so, the system may undergo modifications and refinements. Perhaps it will be continually revised and developed and never become permanent in any specific form. Were that to be the case, it would, however, only reflect the fact that prevailing notions about study quality tend to evolve over time.

10.2 Conclusions

The main conclusions that can be drawn on the basis of the research presented in this report can be summarised as follows:

1. There exists a large body of road safety evaluation studies. It can be shown that the results of many of these studies are related to aspects of study quality, in particular how well a study has controlled for confounding factors.
2. Formal instruments designed to assess study quality have been developed in a number of fields. These instruments differ greatly with respect to the items included and the weights assigned to these. The reliability and validity of available quality scoring instruments is poorly known and very few of the instruments contain items that are useful for road safety evaluation studies.

3. There is no consensus among leading road safety researchers about the meaning of the concept of study quality or about how best to measure study quality numerically, if such measurement is at all possible.
4. A pilot instrument for assessing the quality of road safety evaluation studies was developed in 2000 and tested by a small sample of researchers coding a few studies for quality. While the instrument was found to have acceptable reliability, no way of testing its validity was found.
5. A broad range of study designs are used in road safety evaluation studies. These designs differ with respect to which potentially confounding factors are most relevant for them. A typology of study designs and lists of important potentially confounding factors was developed.
6. A survey of methodological research was conducted for the purpose of assessing the relative importance of various potentially confounding factors in distorting the results of road safety evaluation studies. The survey was inconclusive. A potentially confounding factor will sometimes introduce great bias if left uncontrolled, in other instances it may not bias a study at all. No regularities or easily interpretable patterns in the findings of methodological studies could be discerned.
7. It was therefore concluded that any formal system for scoring road safety evaluation studies numerically for study quality will contain a large element of arbitrariness. Despite this, a first version of a quality scoring system was proposed and illustrations were given of how to apply the system to road safety evaluation studies.
8. Various approaches can be taken to dealing with study quality in meta-analysis. The three most promising approaches are: (a) To use item-specific scores as variables in meta-regression analysis, (b) To use an overall quality score as a variable in meta-regression analysis, and (c) To adjust study weights by quality. All these approaches are defensible and no strong reasons can be given for preferring one approach to the others.

11 References

Amundsen, A. H.; Elvik, R. (2004). Effects on road safety of new urban arterial roads. *Accident Analysis and Prevention*, **26**, 115-123.

Andrew, E. (1984). Method for assessment of the reporting standard of clinical trials with Roentgen contrast media. *Acta Radiologica Diagnosis*, **25**, 55-58.

Austin, R. D.; Carson, J. L. (2002). An alternative accident prediction model for highway-rail interfaces. *Accident Analysis and Prevention*, **34**, 31-42.

Balk, E. M.; Bonis, P. A. L.; Moskowitz, H. A.; Schmid, C. H.; Ioannidis, J. P. A.; Wang, C.; Lau, J. (2002). Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *Journal of the American Medical Association*, **287**, 2973-2982.

Bangert-Drowns, R. L.; Wells-Parker, E.; Chevillard, I. (1997). Assessing the methodological quality of research in narrative reviews and meta-analyses. Chapter 12 (405-429) in: Bryant, K. J.; Windle, M.; West, S. G. (Eds): The science of prevention: Methodological advances from alcohol and substance abuse research. Washington DC, *American Psychological Association*.

Barley, Z. A. (1988). *Assessment of quality of studies for inclusion in meta-analyses*. Ph D dissertation. Salt Lake City, University of Colorado.

Basile, A. J. (1962). Effect of pavement edge markings on traffic accidents in Kansas. *Highway Research Board Bulletin*, **308**, 80-86.

Beckerman, H.; de Bie, R. A.; Bouter, L. M.; De Cuyper, H. J.; Oostendorp, R. A. B. (1992). The efficacy of laser therapy for musculoskeletal and skin disorders: a criteria-based meta-analysis of randomized clinical trials. *Physical Therapy,* **72**, 483-491.

Bédard, A.; Bravo, G. (1998). Combining studies using effect sizes and quality scores: application to bone loss in postmenopausal women. *Journal of Clinical Epidemiology*, **51**, 801-807.

Berlin, J. A.; Rennie, D. Measuring the quality of trials. The quality of quality scales. Editorial in *Journal of the American Medical Association*, **282**, 1083-1085.

Bjørnskau, T.; Elvik, R. (1992). Can road traffic law enforcement permanently reduce the number of accidents? *Accident Analysis and Prevention*, **24**, 507-520.

Briss, P. A.; Zaza, S.; Pappaioanou, M.; Fielding, J.; Wright-De Agüero, L.; Truman, B. I.; Hopkins, D. P.; Dolan Mullen, P.; Thompson, R. S.; Woolf, S. H.; Carande-Kulis, V. G.; Anderson, L.; Hinman, A. R.; McQueen, D. V.; Teutsch, S. M.; Harris, J. R. (2000). Developing an evidence-based Guide to Community Preventive Services – methods. *American Journal of Preventive Medicine*, **18**, 35-43.

Brüde, U.; Larsson, J.; Thulin, H. (1980). *Trafikolyckors samband med linjeföring – för olika belagd bredd, hastighetsgräns, årstid, ljusförhållanden och region*. VTI-meddelande 235. Statens väg- och trafikinstitut, Linköping.

Chalmers, I.; Adams, M.; Dickersin, K.; Hetherington, J.; Tarnow-Mordi, W.; Meinert, C.; Tonascia, S.; Chalmers, T. C. (1990). A cohort study of summary reports of controlled trials. *Journal of the American Medical Association*, **263**, 1401-1405.

Chalmers, T. C.; Smith, H.; Blackburn, B.; Silverman, B.; Schroeder, B.; Reitman, D.; Ambroz, A. (1981). A method for assessing the quality of a randomized control trial. *Controlled Clinical Trials*, **2**, 31-49.

Christensen, P. (2003). *Topics in meta-analysis. A literature survey*. Report 692. Oslo, Institute of Transport Economics.

Christensen, P.; Elvik, R. (2007). Effects on accidents of periodic motor vehicle inspection in Norway. *Accident Analysis and Prevention*, **39**, 47-52.

Clark, H. D.; Wells, G. A.; Huët, C.; McAlister, F. A.; Salmi, L. R.; Fergusson, D.; Laupacis, A. (1999). Assessing the quality of randomized trials: reliability of the Jadad scale. *Controlled Clinical Trials*, **20**, 448-452.

Cook, T. D.; Campbell, D. T. (1979). *Quasi-experimentation. Design and analysis issues for field settings*. Chicago, ILL, Rand-McNally.

Crombie, I. K. (1996). *The pocket guide to critical appraisal*. London, BMJ Publishing Group.

Cummings, P.; Rivara, F. P.; Thompson, D. C.; Thompson, R. S. (2006). Misconceptions regarding case-control studies of bicycle helmets and head injury. *Accident Analysis and Prevention*, **38**, 636-643.

Cunningham, P.; Baker, C. C.; Clancy, T. V. (1997). A comparison of the association of helicopter and ground ambulance transport with the outcome of injury in trauma patients transported from the scene. *The Journal of Trauma: Injury, Infection, and Critical Care*, **43**, 940-946.

Curnow; W. J. (2003). The efficacy of bicycle helmets against brain injury. *Accident Analysis and Prevention*, **35**, 287-292.

Curnow, W. J. (2005). The Cochrane Collaboration and bicycle helmets. *Accident Analysis and Prevention*, **37**, 569-573.

Curnow, W. J. (2006). Bicycle helmets: lack of efficacy against brain injury. *Accident Analysis and Prevention*, **38**, 833-834.

Curnow, W. J. (2007). Bicycle helmets and brain injury. *Accident Analysis and Prevention*, **39**, 433-436.

Deeks, J. J.; Dinnes, J.; D'Amico, R.; Sowden, A. J.; Skarovitch, C.; Song, F.; Petticrew, M.; Altman, D. G. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment*, **7**, 27 (supplement).

Detsky, A. S.; Naylor, C. D.; O'Rourke, K.; McGeer, A. J.; L'Abbé, K. A. (1992). Incorporating variations in the quality of individual randomized trials into meta-analysis. *Journal of Clinical Epidemiology*, **45**, 255-265.

Downs, S. H.; Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health*, **52**, 377-384.

Egan, M.; Pettigrew, M.; Ogilvie, D.; Hamilton, V. (2003). New roads and human health: a systematic review. *American Journal of Public Health*, **93**, 1463-1471.

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

Elvik, R. (1998). 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.

Elvik, R. (1999). *Assessing the validity of evaluation research by means of meta-analysis. Case illustrations from road safety research*. Report 430. Institute of Transport Economics, Oslo

Elvik, R. (2001A). *Quality scoring of road safety evaluation studies*. Unpublished manuscript. Institute of Transport Economics, Oslo.

Elvik, R. (2001B). *A test of the validity and reliability of a quality scoring system for road safety evaluation studies*. Unpublished manuscript. Institute of Transport Economics, Oslo.

Elvik, R. (2001C). *Quantified road safety targets. An assessment of evaluation methodology*. Report 539. Institute of Transport Economics, Oslo.

Elvik, R. (2002A). The importance of confounding in observational before-and-after studies of road safety measures. *Accident Analysis and Prevention*, **34**, 631-635.

Elvik, R. (2002B). The effect on accidents of technical inspections of heavy vehicles in Norway. *Accident Analysis and Prevention*, **34**, 753-762.

Elvik, R. (2002C). *Measuring study quality: mission impossible?* Paper presented at session 539, Transportation Research Board Annual Meeting 2002. Washington DC. Available from the author on request.

Elvik, R. (2003A). Assessing the validity of road safety evaluation studies by analysing causal chains. *Accident Analysis and Prevention*, **35**, 741-748.

Elvik, R. (2003B). Effects on Road Safety of Converting Intersections to Roundabouts. Review of evidence from Non-U.S. Studies. *Transportation Research Record*, **1847**, 1-9

Elvik, R. (2004). To what extent can theory account for the findings of road safety evaluation studies? *Accident Analysis and Prevention*, **36**, 841-849.

Elvik, R. (2005A). Introductory guide to systematic reviews and meta-analysis. *Transportation Research Record*, **1908**, 230-235.

Elvik, R. (2005B). Can we trust the results of meta-analyses? A systematic approach to sensitivity analysis in meta-analyses. *Transportation Research Record*, **1908**, 221-229.

Elvik, R. (2007A). *State-of-the-art approaches to road accident black spot management and safety analysis of road networks*. Report 883. Institute of Transport Economics, Oslo.

Elvik, R. (2007B). Operational criteria of causality for observational road safety evaluation studies. *Transportation Research Record*, **2019**, 74-81.

Elvik, R. (2008A). *An exploratory analysis of models for evaluating the combined effects of road safety measures*. Unpublished manuscript. Institute of Transport Economics, Oslo.

Elvik, R. (2008B). The predictive accuracy of empirical Bayes estimates of road safety. *Accident Analysis and Prevention*, **40**, 1964-1969.

Elvik, R. (2008C). *The non-linearity of risk and the promotion of environmentally sustainable transport*. Unpublished manuscript. Institute of Transport Economics, Oslo.

Elvik, R.; Christensen, P.; Amundsen, A. (2004). *Speed and road accidents. An evaluation of the Power Model*. Report 740. Institute of Transport Economics, Oslo.

Elvik, R.; Christensen, P.; Fjeld Olsen, S. (2003). *Daytime running lights. A systematic review of effects on road safety*. Report 688. Institute of Transport Economics, Oslo.

Elvik, R.; Mysen, A. B. (1999). Incomplete accident reporting: meta-analysis of studies made in thirteen countries. *Transportation Research Record*, **1665**, 133-140.

Elvik, R.; Vaa, T. (2004). *The handbook of road safety measures*. Oxford, Elsevier Science.

Elwood, J. M. (1998). *Critical appraisal of epidemiological studies and clinical trials. Second edition*. Oxford University Press, Oxford.

Emerson, J. D.; Burdick, E.; Hoaglin, D. C.; Mosteller, F.; Chalmers, T. C. (1990). An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Controlled Clinical Trials*, **11**, 339-352.

Erke, A. (2008). Red light for red-light cameras? A meta-analysis of the effects of red-light cameras on crashes. *Accident Analysis and Prevention*, **40**, forthcoming (published electronically).

Evans, L. (1986A). Double pair comparison – a new method to determine how occupant characteristics affect fatality risk in traffic crashes. *Accident Analysis and Prevention*, **18**, 217-227.

Evans, L. (1986B). The effectiveness of safety belts in preventing fatalities. *Accident Analysis and Prevention*, **18**, 229-241.

Evans, L.; Frick, M. (1988). Helmet effectiveness in preventing motorcycle driver and passenger fatalities. *Accident Analysis and Prevention*, **20**, 447-458.

Evans, M.; Pollock, A. V. (1985). A score system for evaluating random control clinical trials of prophylaxis of abdominal surgical wound infection. *British Journal of Surgery*, **72**, 256-260.

Farrington, D. P. (2003). Methodological quality standards for evaluation research. *Annals of the American Academy of Political and Social Science*, **587**, 49-68.

Fosser, S. (1992). An experimental evaluation of the effects of periodic motor vehicle inspection on accident rates. *Accident Analysis and Prevention*, **24**, 599-612.

Friedenreich, C. M.; Brant, R. F.; Riboli, E. (1994). Influence of methodologic factors in a pooled analysis of 13 case-control studies of colorectal cancer and dietary fiber. *Epidemiology*, **5**, 66-79.

Gibbs, L. E. (1989). Quality of study rating form: an instrument for synthesizing evaluation studies. *Journal of Social Work Education*, **25**, 55-67.

Goodman, S. N.; Berlin, J.; Fletcher, S. W.; Fletcher, R. H. (1994). Manuscript quality before and after peer review and editing at Annals of Internal Medicine. *Annals of Internal Medicine*, **121**, 11-21.

Gray, I. (1990). An attempt to reduce accidents in a company car fleet by driver training and encouragement of low risk driving habits. *Journal of Traffic Medicine*, **18**, 139-141.

Greenland, S. (1994). Invited commentary: a critical look at some popular meta-analytic methods. *American Journal of Epidemiology*, **140**, 290-296, 300-302.

Greer, N.; Mosser, G.; Logan, G.; Wagstrom Halaas, G. (2000). A practical approach to evidence grading. *The Joint Commission Journal of Quality Improvement*, **26**, 700-712.

Grendstad, G. et al. (2003). *Fra riksveg til gate – erfaringer fra 16 miljøgater*. Rapport UTB 2003/06. Statens vegvesen, Vegdirektoratet, Utbyggingsavdelingen, Oslo.

Griffith, M. S. (1999). Safety evaluation of rolled-in continuous shoulder rumble strips installed on freeways. *Transportation Research Record*, **1665**, 28-34.

Gøtzsche, P. C. (1989). Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal anti-inflammatory drugs in rheumatoid arthritis. *Controlled Clinical Trials*, **10**, 31-56.

Hagel, B. E.; Pless, I. B. (2006). A critical examination of arguments against bicycle helmet use and legislation. *Accident Analysis and Prevention*, **38**, 277-278.

Hagge, R. A.; Romanowicz, P. A. (1996). Evaluation of California's commercial driver license program. *Accident Analysis and Prevention*, **28**, 547-559.

Harrington D. M. (1972). The young driver follow-up study: an evaluation of the role of human factors in the first four years of driving. *Accident Analysis and Prevention*, **4**, 191-240.

Harvey, A. C.; Durbin, J. (1986). The Effects of Seat Belt Legislation on British Road Casualties: A Case Study in Structural Time Series Modelling. *Journal of the Royal Statistical Society, Series A*, **149**, 187-227.

Hauer, E. (1991). Comparison groups in road safety studies: an analysis. *Accident Analysis and Prevention*, **23**, 609-622.

Hauer, E. (1995). On exposure and accident rate. *Traffic Engineering and Control*, **36**, 134-138.

Hauer, E. (1997). *Observational before-after studies in road safety*. Oxford, Pergamon Press (Elsevier Science).

Hauer, E. (2004). Statistical road safety modelling. *Transportation Research Record*, **1897**, 81-87.

Hauer, E. (2005A). *Cause and effect in observational cross-section studies on road safety*. Draft report. Highway Safety Information System. U.S. Department of Transportation, Federal Highway Administration, Turner-Fairbank Highway Research Center, McLean, VA.

Hauer, E. (2005B). Fishing for safety information in murky waters. *Journal of Transportation Engineering*, **131**, 340-344.

Hauer, E. (2007). *Introduction and fundamentals*. Draft of Chapter 1 of Highway Safety Manual. Unpublished manuscript. Transportation Research Board, Washington D. C.

Hauer, E.; Ng, J. C. N.; Papaioannou, P. (1991). Prediction in road safety studies: an empirical inquiry. *Accident Analysis and Prevention*, **23**, 595-607.

Hauer, E.; Persaud, B. N. (1983). A common bias in before and after comparisons and its elimination. *Transportation Research Record*, **905**, 164-174.

Hauer, E.; Persaud, B. N. (1987). How to estimate the safety of rail-highway grade crossings and the safety effects of warning devices. *Transportation Research Record*, **1114**, 131-140.

Hirst, W. M.; Mountain, L. J.; Maher, M. H. (2004). Sources of error in road safety scheme evaluation: a quantified comparison of current methods. *Accident Analysis and Prevention*, **36**, 705-715.

Holder, H. D.; Wagenaar, A. C. (1994). Mandated server training and reduced alcohol-involved traffic crashes: a time series analysis of the Oregon experience. *Accident Analysis and Prevention*, **26**, 89-97.

Huwiler-Müntener, K.; Jüni, P.; Junker, C.; Egger, M. (2002). Quality of reporting of randomized trials as a measure of methodologic quality. *Journal of the American Medical Association*, **287**, 2801-2804.

Imperiale, T. F.; McCullough, A. J. (1990). Do corticosteroids reduce mortality from alcoholic hepatitis? A meta-analysis of the randomized trials. *Annals of Internal Medicine*, **113**, 299-307.

Ioannidis, J. P. A. (2005A). Why most published research findings are false. *Public Library of Science: Medicine*, **2**, 696-701 (published electronically at: www.plosmedicine.org).

Ioannidis, J. P. A. (2005B). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, **294**, 218-228.

Jadad, A.; Moore, R. A.; Carroll, D.; Jenkinson, C.; Reynolds, D. J. M.; Gavaghan, D. J.; McQuay, H. J. (1996). Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Controlled Clinical Trials*, **17**, 1-12.

Jonsson, T. (2005). *Predictive models for accidents on urban links*. Doctoral dissertation. Bulletin 226. Lund Institute of Technology, Department of Technology and Society, Traffic Engineering, Lund.

Jüni, P.; Witschi, A.; Bloch, R.; Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association*, **282**, 1054-1060.

Jørgensen, N. O.; Rabani. Z. (1969). *Cykelstiers betydning for færdselssikkerheden*. Rapport 1. Rådet for trafiksikkerhedsforskning, København.

Kallberg, V-P. (1993). Reflector posts – signs of danger? *Transportation Research Record*, **1403**, 57-66.

Khan, S.; Shanmugam, R.; Hoeschen, B. (1999). Injury, fatal, and property damage accident models for highway corridors. *Transportation Research Record*, **1665**, 84-92.

Kim, D-G.; Washington, S. (2006). The significance of endogeneity problems in crash models: an examination of left-turn lanes in intersection crash models. *Accident Analysis and Prevention*, **38**, 1094-1100.

Kleijnen, J.; Knipschild, P.; ter Riet, G. (1991). Clinical trials of homeopathy. *British Medical Journal*, **302**, 316-323.

Koes, B. W.; Assendelft, W. J. J.; van der Heiden, G. J. M. G.; Bouter, L. M.; Knipschild, P. G. (1991). Spinal manipulation and mobilisation for back and neck pain: a blinded review. *British Medical Journal*, **303**, 1298-1303.

Leden, L.; Hämäläinen, O.; Manninen, E. (1998). The effect of resurfacing on friction, speeds and safety on main roads in Finland. *Accident Analysis and Prevention*, **30**, 75-85.

Levine, D. W.; Golob, T. F.; Recker, W. W. (1988). Accident migration associated with lane-addition projects on urban freeways. *Traffic Engineering and Control*, **29**, 624-629.

Levine, J. (1991). Trial assessment procedure scale (TAPS). In: Spilker, B. (Ed): *Guide to clinical trials*, 780-792. New York, NY , Raven Press.

Linde, K.; Jonas, W. B.; Melchart, D.; Willich, S. (2001). The methodological quality of randomized controlled trials of homeopathy, herbal medicines and acupuncture. *International Journal of Epidemiology*, **30**, 526-531.

Linde, K.; Scolz, M.; Ramirez, G.; Clausius, N.; Melchart, D.; Jonas, W. B. (1999). Impact of study quality on outcome in placebo-controlled trials of homeopathy. *Journal of Clinical Epidemiology*, **52**, 631-636.

Lohr, K. N.; Carey, T. S. (1999). Assessing “Best Evidence”: issues in grading the quality of studies for systematic reviews. *The Joint Commission Journal of Quality Improvement*, **25**, 470-479.

Lord, D.; Washington, S. P; Ivan, J. N. (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis and Prevention*, **37**, 35-46.

Lund, T. (Red) (2002). *Innføring i forskningsmetodologi*. Unipub, Oslo.

Lösel, F.; Köferl, P. (1989). Evaluation research on correctional treatment in West Germany: a meta-analysis. In: Wegener, H.; Lösel, F.; Haisch, J. (Eds) *Criminal behaviour and the justice system: Psychological perspectives*, 334-355. Springer Verlag, New York.

Maimaris, C.; Summer, C. L.; Browning, C.; Palmer, C. R. (1994). Injury patterns in cyclists attending an accident and emergency department: a comparison of helmet wearers and non-wearers. *British Medical Journal*, **308**, 1537-1540.

Margetts, B. M.; Thompson, R. L.; Key, T.; Duffy, S.; Nelson, M.; Bingham, S.; Wiseman, M. (1995). Development of a scoring system to judge the scientific quality of information from case-control and cohort studies of nutrition and disease. *Nutrition and Cancer*, **24**, 231-239.

Matthews, L. R.; Barnes, J. W. (1988). Relation between road environment and curve accidents. *Proceedings of the 14th ARRB Conference*, **Part 4**, 105-120.

Moher, D.; Cook, D. J.; Eastwood, S.; Olkin, I.; Rennie, D.; Stroup, D. F. (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *The Lancet*, **354**, 1896-1900.

Moher, D.; Jadad, A.; Nichol, G.; Penman, M.; Tugwell, P.; Walsh, S. (1995). Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Controlled Clinical Trials*, **16**, 62-73.

Moher, D.; Jones, A.; Cook, D. J.; Jadad, A. R.; Tugwell, P.; Klassen, T. P. (1998). Does quality of reports of nonrandomised trials affect estimates of intervention efficacy reported in meta-analyses? *The Lancet*, **352**, 609-613.

Mountain, L. J.; Hirst, W. M.; Maher, M. J. (2005). Are speed enforcement cameras more effective than other speed management measures? The impact of speed management schemes on 30 mph roads. *Accident Analysis and Prevention*, **37**, 742-754.

Mountain, L. J.; Jarrett, D. F.; Fawaz, B. (1995). The safety effects of highway engineering schemes. *Proceedings of the Institution of Civil Engineers Transport*, **111**, 298-309.

Muskaug, R. (1985). *Risiko på norske riksveger*. Rapport. Transportøkonomisk institutt, Oslo.

Nordtyp-projektgruppen. (1980). *Trafikulykker på vejstrækninger. En sammenstilling af ulykkesfrekvenser for nordiske typesektioner*. Rapport. Vejdirektoratet, København.

Nurmohamed, M. T.; Rosendaal, F. R.; Büller, H. R.; Dekker, E.; Hommes, D. W.; Vandenbroucke, J. P.; Briët, E. (1992). Low-molecular-weight heparin versus standard heparin in general and orthopaedic surgery: a meta-analysis. *The Lancet*, **340**, 152-156.

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

Odberg, T. A. (1996). *Erfaringer med etablering av rundkjøringer i Vestfold. Ulykker, atferd og geometri*. Hovedoppgave høsten 1996. Norges teknisk-naturvitenskapelige universitet, Institutt for samferdselsteknikk, Trondheim.

Ogden, K. W. (1997). The effects of paved shoulders on accidents on rural highways. *Accident Analysis and Prevention*, **29**, 353-362.

Onghena, P.; Van Houdenhove, B. (1992). Antidepressant-induced analgesia in chronic non-malignant pain: a meta-analysis of 39 placebo-controlled studies. *Pain*, **49**, 205-219.

Oslo veivesen. (1995). *Ulykkesanalyse. Rundkjøringer i Oslo*. Trafikksikkerhetskontoret, Oslo veivesen, Oslo.

Oxman, A. D.; Guyatt, G. H. (1991). Validation of an index of the quality of review articles. *Journal of Clinical Epidemiology*, **44**, 1271-1278.

Park, Y-J.; Saccomanno, F. F. (2005). *Evaluating factors affecting safety at highway-railway grade crossings*. TRB 2005 Annual Meeting CD-Rom.

Peltola, H. (2000). Seasonally changing speed limits. Effects on speeds and accidents. *Transportation Research Record*, **1734**, 46-51.

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

Persaud, B.; Lyon, C. (2007). Empirical Bayes before-after studies: Lessons learned from two decades of experience and future directions. *Accident Analysis and Prevention*, **39**, 546-555.

Petticrew, M. (2003). Why certain systematic reviews reach uncertain conclusions. *British Medical Journal*, **326**, 756-758.

Phillips, K. B.; McCutchen, J. A. (1991). Economic Regulation vs Safety Regulation of the Trucking Industry - Which More Effectively Promotes Safety? *Transportation Quarterly*, **45**, 323-340.

Poynard, T. (1988). Evaluation de la qualité méthodologique des essais thérapeutiques randomisés. *La Presse Médicale*, **17**, 315-318.

Quddus, M. A. (2008). Time series count data models: An empirical application to traffic accidents. *Accident Analysis and Prevention*, **40**, 1732-1741.

Ragnøy, A. (2004). *Endring av fartsgrenser. Effekt på kjørefart og ulykker*. Rapport 729. Transportøkonomisk institutt, Oslo.

Ragnøy, A.; Christensen, P.; Elvik, R. (2002). *Skadegradstetthet. Et nytt mål på hvor farlig en vegstrekning er*. Rapport 618. Transportøkonomisk institutt, Oslo.

Ragnøy, A.; Muskaug, R. (2003). *Fartsgrenseendringer virker. Reduserte fartsgrenser reduserer kjørefarten*. Arbeidsdokument av 10.11.2003. Transportøkonomisk institutt, Oslo.

Reisch, J. S.; Tyson, J. E.; Mize, S. G. (1989). Aid to the evaluation of therapeutic studies. *Pediatrics*, **84**, 815-827.

Riet, G. ter; Kleijnen, J.; Knipschild, P. (1990). Acupuncture and chronic pain: a criteria-based meta-analysis. *Journal of Clinical Epidemiology*, **43**, 1191-1199.

Rock, S. M. (1995). Impact of the 65 mph speed limit on accidents, deaths, and injuries in Illinois. *Accident Analysis and Prevention*, **27**, 207-214.

Rosenthal, R. (1991A). *Meta-analytic procedures for social research*. Applied social research methods series volume 6. Newbury Park, Ca, Sage Publications.

Rosenthal, R. (1991B). Quality-weighting of studies in meta-analytic research. *Psychotherapy Research*, **1**, 25-28.

Sakshaug, K. (1998). *Effekt av overhøyde i kurver: Beskrivelse av datamaterialet*. Notat av 2.11.1998. SINTEF, Bygg og miljøteknikk, Trondheim.

Saunders, L. D.; Soomro, G. M.; Buckingham, J.; Jamtvedt, G.; Raina, P. (2003). Assessing the methodological quality of nonrandomized intervention studies. *Western Journal of Nursing Research*, **25**, 223-237.

Schrøder Hansen, K.; Engesæter, L. B.; Viste, A. (2003). Protective effect of different types of bicycle helmets. *Traffic Injury Prevention*, **4**, 285-290.

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

Shannon, H. S.; Robson, L. S.; Guastello, S. J. (1999). Methodological criteria for evaluating occupational safety intervention research. *Safety Science*, **31**, 161-179.

Sindhu, F.; Carpenter, L.; Seers, K. (1997). Development of a tool to rate the quality of randomized controlled trials using a Delphi technique. *Journal of Advanced Nursing*, **25**, 1262-1268.

Slim, K.; Nini, E.; Forestier, D.; Kwiatkowski, F.; Panis, Y.; Chipponi, J. (2003). Methodological index for non-randomized studies (MINORS): development and validation of a new instrument. *Australia and New Zealand Journal of Surgery*, **73**, 712-716.

Smith, K.; Cook, D.; Guyatt, G. H.; Madhavan, J.; Oxman, A. D. (1992). Respiratory muscle training in chronic airflow limitation: a meta-analysis. *American Review of Respiratory Diseases*, **145**, 533-539.

Spitzer, W. O.; Lawrence, V.; Dales, R.; Hill, G.; Archer, M. C.; Clark, P.; Abenhaim, L.; Hardy, J.; Sampalis, J.; Pinfold, S. P.; Morgan, P. P. (1990).

Links between passive smoking and disease: a best-evidence synthesis. *Clinical and Investigative Medicine*, **13**, 17-42.

Stewart, D. (1988). Pedestrian guardrails and accidents. *Traffic Engineering and Control*, **29**, 450-455.

Thompson, D. C.; Rivara, F. P.; Thompson, R. S. (1996). Effectiveness of bicycle safety helmets in preventing head injuries. A case-control study. *Journal of the American Medical Association*, **276**, 1968-1973.

Tritchler, D. (1999). Modelling study quality in meta-analysis. *Statistics in Medicine*, **18**, 2135-2145.

Ulmer, R. G.; Preusser, D. F.; Ferguson, S. A.; Williams, A. F. (1999). Teenage crash reduction associated with delayed licensure in Louisiana. *Journal of Safety Research*, **30**, 31-38.

Verhagen, A. P.; de Vet, H. C. W.; de Bie, R. A.; Boers, M.; van den Brandt, P. A. (2001). The art of quality assessment of RCTs included in systematic reviews. *Journal of Clinical Epidemiology*, **54**, 651-654.

Verhagen, A. P.; de Vet, H. C. W.; Vermeer, F.; Widdershoven, J. W. M. G.; de Bie, R. A.; Kessels, A. G. H.; Boers, M.; van den Brandt, P. A. (2002). The influence of methodologic quality on the conclusion of a landmark meta-analysis on thrombolytic therapy. *International Journal of Technology Assessment in Health Care*, **18**, 11-23.

Vaa, T. (1995). *Salting og trafikksikkerhet. Del 2: Sammenligning av ulykkesfrekvens på saltet og usaltet vegnett. Saltingens effekt på kjørefart.* Rapport MITRA 03/95. Statens vegvesen, Vegdirektoratet, Oslo.

Wentz, R.; Roberts, I.; Bunn, F.; Edwards, P.; Kwan, I.; Lefebvre, C. (2001). Identifying controlled evaluation studies of road safety interventions. Searching for needles in a haystack. *Journal of Safety Research*, **32**, 267-276.

Wong, S. C.; Leung, B. S. Y.; Loo, B. P. Y.; Hung, W. T.; Lo, H. K. (2004). A qualitative assessment methodology for road safety policy strategies. *Accident Analysis and Prevention*, **36**, 281-293.

Zaza, S.; Wright-De Agüero, L. K.; Briss, P. A.; Truman, B. I.; Hopkins, D. P.; Hennessy, M. H.; Sosin, D. M.; Anderson, L.; Carande-Kulis, V. G.; Teutsch, S. M.; Pappaioanou, M. (2000). Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services. *American Journal of Preventive Medicine*, **18**, 44-74.

Appendix 1:Test of pilot quality scoring system

Study	Item	RE	PC	AF	AHA	TBJ		C1 (RE)	C2 (PC)	C3 (AF)	C4 (AHA)	C5 (TBJ)	Items	Chance
Rock 95	Sampling	1	1	1	2	1	C1 (RE)	19	17	18	18	19	0.60	0.33
	Data level	2	1	1	2	1	C2 (PC)	0.8					0.40	0.50
	Severity	2	2	2	2	2	C3 (AF)	0.7	0.7				1.00	0.50
	Uncertainty	2	2	2	2	2	C4 (AHA)	0.8	0.7	0.5			1.00	0.33
	Direction	2	2	2	2	2	C5 (TBJ)	0.8	0.8	0.9	0.6		1.00	0.50
	Confounding	4	4	3	4	4	Raters			Items			0.60	0.20
	Mechanism	2	2	2	2	2	Mean	0.73		Mean	0.73		1.00	0.50
	Dose-response	NA	NA	1	NA	1	Std	0.12		Std	0.30		0.40	0.50
	Specificity	2	1	2	NA	2							0.30	0.50
	Theory	2	2	2	2	2							1.00	0.50
Ogden 97	Sampling	1	2	2	2	1		C1 (RE)	C2 (PC)	C3 (AF)	C4 (AHA)	C5 (TBJ)	0.30	0.33
	Data level	2	2	2	2	2	C1 (RE)	13	16	16	20	17	1.00	0.50
	Severity	2	2	1	2	2	C2 (PC)	0.7					0.70	0.50
	Uncertainty	3	3	2	2	2	C3 (AF)	0.5	0.6				0.40	0.33
	Direction	2	2	2	2	2	C4 (AHA)	0.3	0.6	0.7			1.00	0.50
	Confounding	2	3	3	3	3	C5 (TBJ)	0.5	0.6	0.7	0.8		0.60	0.20
	Mechanism	1	2	2	2	2	Raters			Items			0.60	0.50
	Dose-response	NA	NA	NA	1	1	Mean	0.60		Mean	0.60		0.40	0.50
	Specificity	NA	NA	NA	2	NA	Std	0.14		Std	0.24		0.60	0.50
	Theory	NA	NA	2	2	2							0.40	0.50

Study	Item	RE	PC	AF	AHA	TBJ		C1 (RE)	C2 (PC)	C3 (AF)	C4 (AHA)	C5 (TBJ)	Items	Chance
Leden 98	Sampling	1	2	2	2	1		C1 (RE)	C2 (PC)	C3 (AF)	C4 (AHA)	C5 (TBJ)	0.40	0.33
	Data level	2	1	1	2	2	C1 (RE)	15	15	16	22	18	0.40	0.50
	Severity	2	2	1	2	2	C2 (PC)	0.8					0.60	0.50
	Uncertainty	3	3	3	3	2	C3 (AF)	0.6	0.8				0.60	0.33
	Direction	2	2	2	2	2	C4 (AHA)	0.6	0.6	0.6			1.00	0.50
	Confounding	3	3	3	3	3	C5 (TBJ)	0.7	0.5	0.5	0.7		1.00	0.20
	Mechanism	2	2	2	2	2	Raters			Items			1.00	0.50
	Dose-response	NA	NA	NA	2	2	Mean	0.64		Mean	0.64		0.40	0.50
	Specificity	NA	NA	NA	2	NA	Std	0.11		Std	0.26		0.60	0.50
	Theory	NA	NA	2	2	2							0.40	0.50
Ulmer 99	Sampling	1	2	1	2	1		C1 (RE)	C2 (PC)	C3 (AF)	C4 (AHA)	C5 (TBJ)	0.40	0.33
	Data level	1	1	1	2	1	C1 (RE)	15	14	17	18	18	0.60	0.50
	Severity	2	2	2	2	2	C2 (PC)	0.8					1.00	0.50
	Uncertainty	2	2	2	2	2	C3 (AF)	0.9	0.7				1.00	0.33
	Direction	2	2	2	2	2	C4 (AHA)	0.6	0.6	0.7			1.00	0.50
	Confounding	3	3	3	2	3	C5 (TBJ)	0.8	0.6	0.9	0.6		0.60	0.20
	Mechanism	2	2	2	2	2	Raters			Items			1.00	0.50
	Dose-response	NA	NA	NA	NA	1	Mean	0.72		Mean	0.72		0.60	0.50
	Specificity	2	NA	2	2	2	Std	0.12		Std	0.25		0.60	0.50
	Theory	NA	NA	2	2	2							0.40	0.50

Study	Item	RE	PC	AF	AHA	TBJ		C1 (RE)	C2 (PC)	C3 (AF)	C4 (AHA)	C5 (TBJ)	Items	Chance
Griffith 99	Sampling	1	2	1	2	1		C1 (RE)	C2 (PC)	C3 (AF)	C4 (AHA)	C5 (TBJ)	0.40	0.33
	Data level	1	2	2	2	2	C1 (RE)	15	18	18	17	19	0.60	0.50
	Severity	2	2	2	1	2	C2 (PC)	0.7					0.60	0.50
	Uncertainty	3	3	3	3	3	C3 (AF)	0.6	0.7				1.00	0.33
	Direction	2	2	2	2	2	C4 (AHA)	0.4	0.7	0.6			1.00	0.50
	Confounding	3	3	2	3	3	C5 (TBJ)	0.7	0.8	0.9	0.7		0.60	0.20
	Mechanism	1	2	2	2	2	Raters			Items			0.60	0.50
	Dose-response	NA	NA	NA	NA	NA	Mean	0.68		Mean	0.68		1.00	0.50
	Specificity	2	2	2	NA	2	Std	0.13		Std	0.23		0.60	0.50
	Theory	NA	NA	2	2	2							0.40	0.50
All	Sampling							C1 (RE)	C2 (PC)	C3 (AF)	C4 (AHA)	C5 (TBJ)	0.42	0.33
	Data level						C1 (RE)						0.60	0.50
	Severity						C2 (PC)	0.76					0.78	0.50
	Uncertainty						C3 (AF)	0.66	0.70				0.80	0.33
	Direction						C4 (AHA)	0.54	0.64	0.62			1.00	0.50
	Confounding						C5 (TBJ)	0.70	0.66	0.78	0.68		0.68	0.20
	Mechanism						Raters			Items			0.84	0.50
	Dose-response						Mean	0.67		Mean	0.67		0.56	0.50
	Specificity						Std	0.07		Std	0.18		0.54	0.50
	Theory												0.52	0.50

Appendix 2: Scoring of 18 studies by means of quality scale

Design	Item	Values	Score	Relative score	Item weight	Total weight
All	Sampling	Population	4	1.00	0.03	0.03
		Random	3	0.75	0.03	0.03
		Non-random	2	0.50	0.03	0.03
		Convenience	1	0.25	0.03	0.03
	Severity	FaSeSliPdo	4	1.00	0.05	0.05
		FaInjPdo	3	0.75	0.05	0.05
		FatSerSli	3	0.75	0.05	0.05
		InjPdo	2	0.50	0.05	0.05
		Inj	1	0.25	0.05	0.05
		Mixing	0	0.00	0.05	0.05
	Association	Detectable	1	1.00	0.06	0.06
		Not detectable	0	0.00	0.06	0.06
	Strength	Comp	1	1.00	0.03	0.03
		Non-comp	0	0.00	0.03	0.03
	Consistency	Comp	1	1.00	0.03	0.03
		Non-comp	0	0.00	0.03	0.03
	Direction	Correct	1	1.00	0.10	0.10
		Indeterminate	0	0.00	0.10	0.10
	Mechanism	Full empirical	3	1.00	0.06	0.06
		Part empirical	2	0.67	0.06	0.06
		Mentioned	1	0.33	0.06	0.06
		Not ident	0	0.00	0.06	0.06
	Theory	Yes	1	1.00	0.03	0.03
		No	0	0.00	0.03	0.03
	Dose-response	Possible	1	1.00	0.08	0.08
		Not possible	0	0.00	0.08	0.08
	Specificity	Possible	1	1.00	0.03	0.03
		Not possible	0	0.00	0.03	0.03

Design	Item	Values	Score	Relative score	Item weight	Total weight
Experiments	Equivalence	Proved	4	1.00	0.40	0.50
		Presumed	3	0.67	0.40	0.50
		Adjusted	2	0.33	0.40	0.50
		Not adjusted	1	0.00	0.40	0.50
	Diffusion	No	3	1.00	0.20	0.50
		Adjusted	2	0.50	0.20	0.50
		Yes	1	0.00	0.20	0.50
	Attrition	No	3	1.00	0.10	0.50
		Adjusted	2	0.50	0.10	0.50
		Yes	1	0.00	0.10	0.50
	Unintended	No	4	1.00	0.30	0.50
		Adjusted	3	0.67	0.30	0.50
		Suspected	2	0.33	0.30	0.50
		Yes	1	0.00	0.30	0.50
Before-after	RTM	EB-method	3	1.00	0.40	0.50
		Other method	2	0.50	0.40	0.50
		No control	1	0.00	0.40	0.50
	Trend	Control	2	1.00	0.30	0.50
		No control	1	0.00	0.30	0.50
	Volume	Control	2	1.00	0.10	0.50
		No control	1	0.00	0.10	0.50
	Co-incident events	None known	2	1.00	0.05	0.50
		Known	1	0.00	0.05	0.50
	Other measures	Control	2	1.00	0.10	0.50
		No control	1	0.00	0.10	0.50
	Migration	Control	3	1.00	0.05	0.50
		Not likely	2	0.50	0.05	0.50
		Possible	1	0.00	0.05	0.50
Cross-section	Self-selection	Not present	4	1.00	0.20	0.50
		Adjusted	3	0.67	0.20	0.50
		Suspected	2	0.33	0.20	0.50
		Yes	1	0.00	0.20	0.50
	Endogeneity	Not present	4	1.00	0.20	0.50
		Adjusted	3	0.67	0.20	0.50
		Suspected	2	0.33	0.20	0.50
		Yes	1	0.00	0.20	0.50
	Volume	Non-linear	3	1.00	0.20	0.50
		Rates	2	0.50	0.20	0.50
		No control	1	0.00	0.20	0.50
	Composition	Control	2	1.00	0.20	0.50
		No control	1	0.00	0.20	0.50
	Risk factors	Multiple	3	1.00	0.20	0.50
		Few	2	0.50	0.20	0.50
		None	1	0.00	0.20	0.50

Design	Item	Values	Score	Relative score	Item weight	Total weight
Case-control	Equivalence	Proved	4	1.00	0.60	0.50
		Statistical	3	0.67	0.60	0.50
		Stratification	2	0.33	0.60	0.50
		Few or no	1	0.00	0.60	0.50
	Prognostic	Proved	4	1.00	0.20	0.50
		Statistical	3	0.67	0.20	0.50
		Stratification	2	0.33	0.20	0.50
		Few or no	1	0.00	0.20	0.50
	Treatment	Specified	2	1.00	0.20	0.50
		Not speci	1	0.00	0.20	0.50
Multivariate	Endogeneity	Not present	4	1.00	0.40	0.50
		Adjusted	3	0.67	0.40	0.50
		Suspected	2	0.33	0.40	0.50
		Yes	1	0.00	0.40	0.50
	Functional form	Explicit	3	1.00	0.10	0.50
		Default	2	0.50	0.10	0.50
		Implausible	1	0.00	0.10	0.50
	Collinearity	No	3	1.00	0.10	0.50
		Undecided	2	0.50	0.10	0.50
		Yes	1	0.00	0.10	0.50
	Omitted var	No	3	1.00	0.10	0.50
		Undecided	2	0.50	0.10	0.50
		Yes	1	0.00	0.10	0.50
	Residuals	Correct	2	1.00	0.05	0.50
		Dubious	1	0.00	0.05	0.50
	Model form	Plausible	2	1.00	0.10	0.50
		Implausible	1	0.00	0.10	0.50
	Dependent	Appropriate	2	1.00	0.15	0.50
		Inapprop	1	0.00	0.15	0.50
Time series	Independent	Adjusted	2	1.00	0.45	0.50
		Not adjusted	1	0.00	0.45	0.50
	Co-incident	No	2	1.00	0.45	0.50
		Possibly	1	0.00	0.45	0.50
	Residuals	Correct	2	1.00	0.10	0.50
		Dubious	1	0.00	0.10	0.50

Design	Item	Values	Score	Fosser 1992 score
All	Sampling	Population	4	0.030
		Random	3	
		Non-random	2	
		Convenience	1	
	Severity	FaSeSliPdo	4	
		FaInjPdo	3	
		FatSerSli	3	
		InjPdo	2	0.025
		Inj	1	
		Mixing	0	
	Association	Detectable	1	
		Not detectable	0	0.000
	Strength	Comp	1	0.030
		Non-comp	0	
	Consistency	Comp	1	0.030
		Non-comp	0	
	Direction	Correct	1	0.100
		Indeterminate	0	
	Mechanism	Full empirical	3	
		Part empirical	2	0.040
		Mentioned	1	
		Not ident	0	
	Theory	Yes	1	
		No	0	0.000
	Dose-response	Possible	1	0.080
		Not possible	0	
	Specificity	Possible	1	
		Not possible	0	0.000
Experiments	Equivalence	Proved	4	0.200
		Presumed	3	
		Adjusted	2	
		Not adjusted	1	
	Diffusion	No	3	0.100
		Adjusted	2	
		Yes	1	
	Attrition	No	3	0.050
		Adjusted	2	
		Yes	1	
	Unintended	No	4	0.150
		Adjusted	3	
		Suspected	2	
		Yes	1	

Design	Item	Values	Score	Christensen, Elvik 2007 score
All	Sampling	Population	4	
		Random	3	
		Non-random	2	
		Convenience	1	0.008
	Severity	FaSeSliPdo	4	
		FaInjPdo	3	
		FatSerSli	3	
		InjPdo	2	
		Inj	1	
		Mixing	0	0.000
	Association	Detectable	1	
		Not detectable	0	0.000
	Strength	Comp	1	0.030
		Non-comp	0	
	Consistency	Comp	1	0.030
		Non-comp	0	
	Direction	Correct	1	0.100
		Indeterminate	0	
	Mechanism	Full empirical	3	
		Part empirical	2	0.040
		Mentioned	1	
		Not ident	0	
	Theory	Yes	1	
		No	0	0.000
	Dose-response	Possible	1	0.080
		Not possible	0	
	Specificity	Possible	1	
		Not possible	0	0.000
Multivariate	Endogeneity	Not present	4	0.200
		Adjusted	3	
		Suspected	2	
		Yes	1	
	Functional form	Explicit	3	0.050
		Default	2	
		Implausible	1	
	Collinearity	No	3	
		Undecided	2	0.025
		Yes	1	
	Omitted var	No	3	
		Undecided	2	
		Yes	1	0.000
	Residuals	Correct	2	0.025
		Dubious	1	
	Model form	Plausible	2	0.050
		Implausible	1	
	Dependent	Appropriate	2	0.075
		Inappropiate	1	

Design	Item	Values	Score	Odberg 1996	Oslo vei 1995	Nygaard 1988
All	Sampling	Population	4			
		Random	3			
		Non-random	2			
		Convenience	1	0.008	0.008	0.008
	Severity	FaSeSliPdo	4			
		FaInjPdo	3			
		FatSerSli	3	0.038	0.038	0.038
		InjPdo	2			
		Inj	1			
		Mixing	0			
	Association	Detectable	1	0.060	0.060	0.060
		Not detectable	0			
	Strength	Comp	1	0.030	0.030	0.030
		Non-comp	0			
	Consistency	Comp	1	0.030	0.030	0.030
		Non-comp	0			
	Direction	Correct	1	0.100	0.100	0.100
		Indeterminate	0			
	Mechanism	Full empirical	3			
		Part empirical	2	0.040		
		Mentioned	1			
		Not ident	0		0.000	0.000
	Theory	Yes	1	0.030	0.030	0.030
		No	0			
	Dose-response	Possible	1	0.080		0.080
		Not possible	0		0.000	
	Specificity	Possible	1			
		Not possible	0	0.000	0.000	0.000
Before-after	RTM	EB-method	3	0.200		
		Other method	2			
		No control	1		0.000	0.000
	Trend	Control	2	0.150	0.150	0.150
		No control	1			
	Volume	Control	2	0.050		
		No control	1		0.000	0.000
	Co-incident events	None known	2	0.025	0.025	0.025
		Known	1			
	Other measures	Control	2	0.050	0.050	0.050
		No control	1			
	Migration	Control	3			
		Not likely	2	0.013	0.013	0.013
		Possible	1			

Design	Item	Values	Score	Brüde, 1980	Matthews, 1988	Sakshaug, 1998
All	Sampling	Population	4			
		Random	3			
		Non-random	2	0.015		
		Convenience	1		0.008	0.008
	Severity	FaSeSliPdo	4			
		FaInjPdo	3			
		FatSerSli	3			
		InjPdo	2			
		Inj	1			0.013
		Mixing	0	0.000	0.000	
	Association	Detectable	1	0.060	0.060	0.060
		Not detectable	0			
	Strength	Comp	1	0.030	0.030	
		Non-comp	0			0.000
	Consistency	Comp	1	0.030	0.030	0.030
		Non-comp	0			
	Direction	Correct	1	0.100	0.100	0.100
		Indeterminate	0			
	Mechanism	Full empirical	3			
		Part empirical	2			
		Mentioned	1			
		Not ident	0	0.000	0.000	0.000
	Theory	Yes	1			
		No	0	0.000	0.000	0.000
	Dose-response	Possible	1	0.080	0.080	0.080
		Not possible	0			
	Specificity	Possible	1			
		Not possible	0	0.000	0.000	0.000
Cross-section	Self-selection	Not present	4	0.100	0.100	0.100
		Adjusted	3			
		Suspected	2			
		Yes	1			
	Endogeneity	Not present	4	0.100	0.100	0.100
		Adjusted	3			
		Suspected	2			
		Yes	1			
	Volume	Non-linear	3			
		Rates	2	0.050	0.050	0.050
		No control	1			
	Composition	Control	2			
		No control	1	0.000	0.000	0.000
	Risk factors	Multiple	3			
		Few	2	0.050	0.050	
		None	1			0.000

Design	Item	Values	Score	Maimaris, 1994	Thompson, 1996	Schrøder, 2003
All	Sampling	Population	4			
		Random	3			
		Non-random	2			
		Convenience	1	0.008	0.008	0.008
	Severity	FaSeSliPdo	4			
		FaInjPdo	3			
		FatSerSli	3		0.038	
		InjPdo	2			
		Inj	1	0.013		0.013
		Mixing	0			
	Association	Detectable	1	0.060	0.060	0.060
		Not detectable	0			
	Strength	Comp	1	0.030	0.030	0.030
		Non-comp	0			
	Consistency	Comp	1	0.030	0.030	0.030
		Non-comp	0			
	Direction	Correct	1	0.100	0.100	0.100
		Indeterminate	0			
	Mechanism	Full empirical	3			
		Part empirical	2			
		Mentioned	1	0.020		
		Not ident	0		0.000	0.000
	Theory	Yes	1			
		No	0	0.000	0.000	0.000
	Dose-response	Possible	1			
		Not possible	0	0.000	0.000	0.000
	Specificity	Possible	1	0.030		0.030
		Not possible	0		0.000	
Case-control	Equivalence	Proved	4			
		Statistical	3	0.201	0.201	0.201
		Stratification	2			
		Few or no	1			
	Prognostic	Proved	4			
		Statistical	3			
		Stratification	2	0.033	0.033	0.033
		Few or no	1			
	Treatment	Specified	2		0.100	0.100
		Not specified	1	0.000		

Design	Item	Values	Score	Hauer, 1987	Austin, 2002	Park, 2005
All	Sampling	Population	4			
		Random	3			
		Non-random	2	0.015	0.015	0.015
		Convenience	1			
	Severity	FaSeSliPdo	4			
		FaInjPdo	3			
		FatSerSli	3			
		InjPdo	2			
		Inj	1			
		Mixing	0	0.000	0.000	0.000
	Association	Detectable	1	0.060	0.060	0.060
		Not detectable	0			
	Strength	Comp	1		0.030	0.030
		Non-comp	0	0.000		
	Consistency	Comp	1	0.030		0.030
		Non-comp	0		0.000	
	Direction	Correct	1	0.100	0.100	0.100
		Indeterminate	0			
	Mechanism	Full empirical	3			
		Part empirical	2			
		Mentioned	1			
		Not ident	0	0.000	0.000	0.000
	Theory	Yes	1			
		No	0	0.000	0.000	0.000
	Dose-response	Possible	1	0.080	0.080	0.080
		Not possible	0			
	Specificity	Possible	1			
		Not possible	0	0.000	0.000	0.000
Multivariate	Endogeneity	Not present	4	0.200		
		Adjusted	3		0.134	0.134
		Suspected	2			
		Yes	1			
	Functional form	Explicit	3			
		Default	2	0.025	0.025	0.025
		Implausible	1			
	Collinearity	No	3	0.050		0.050
		Undecided	2		0.025	
		Yes	1			
	Omitted var	No	3		0.050	0.050
		Undecided	2	0.025		
		Yes	1			
	Residuals	Correct	2	0.025	0.025	0.025
		Dubious	1			
	Model form	Plausible	2	0.050	0.050	0.050
		Implausible	1			
	Dependent	Appropriate	2	0.075	0.075	0.075
		Inappropriate	1			

Design	Item	Values	Score	Holder, 1994	Hagge, 1996	Wong, 2004
All	Sampling	Population	4	0.030	0.030	0.030
		Random	3			
		Non-random	2			
		Convenience	1			
	Severity	FaSeSliPdo	4			
		FaInjPdo	3			
		FatSerSli	3			0.038
		InjPdo	2		0.025	
		Inj	1	0.013		
		Mixing	0			
	Association	Detectable	1	0.060	0.060	0.060
		Not detectable	0			
	Strength	Comp	1	0.030	0.030	0.030
		Non-comp	0			
	Consistency	Comp	1		0.030	0.030
		Non-comp	0	0.000		
	Direction	Correct	1	0.100	0.100	0.100
		Indeterminate	0			
	Mechanism	Full empirical	3			
		Part empirical	2	0.040		
		Mentioned	1			
		Not ident	0		0.000	0.000
	Theory	Yes	1			
		No	0	0.000	0.000	0.000
	Dose-response	Possible	1	0.080		
		Not possible	0		0.000	0.000
	Specificity	Possible	1			
		Not possible	0	0.000	0.000	0.000
Time series	Independent	Adjusted	2	0.225	0.225	0.225
		Not adjusted	1			
	Co-incident	No	2	0.225	0.225	
		Possibly	1			0.000
	Residuals	Correct	2	0.050	0.050	
		Dubious	1			0.000

Sist utgitte TØI publikasjoner under program: Analyser av miljø- og trafikksikkerhetstiltak

Trafikksikkerhetsindikatoren for alkohol i Safetynet - Datakvalitet i utvalgte land og sammenligning med andre alkoholindikatorer	985/2008
Realisering av nullvisjonen: Hvordan forebygge ulykker og skader blant eldre fotgjengere	972/2008
En komparativ analyse av ulike typer normative premisser for transportsikkerhetspolitikken	964/2008
Nyttekostnadsanalyse av skadeforebyggende tiltak	933/2007
Nyt etappemål for trafiksikerhed i Sverige	930/2007
Utpekning og analyse av ulykkesbelastede steder og sikkerhetsanalyser av vegsystemer - Beste metoder og implementering	919/2007
Beste metoder for utpeking og analyse av ulykkesbelastede steder og sikkerhetsanalyser av vegsystemer	898/2007
Utsiktene til å bedre trafikksikkerheten i Norge	897/2007
Realisering av nullvisjonen: Forebygging av fotgjengerulykker og redusering av ulykkenes alvorlighet	889/2007
Utpekning og analyse av ulykkesbelastede steder og sikkerhetsanalyse av vegsystemer	883/2007
Nullvisjonen - i teori og praksis	873/2007
Effektkatalog for trafikksikkerhetstiltak	851/2006
Trafikksikkerhetsinspeksjoner: effekter og retningslinjer for god praksis	850/2006
Vegdekkets tilstand og trafikksikkerhet. Betydningen av spordybde, ujevnhet og endringer i tverrfall for ulykkesrisikoen	840/2006
Trafikkstøy i boliger. Virkninger av fasadeisoleringstiltak etter grenseverdiforskriften	836/2006

Visiting and postal address:

Institute of Transport Economics
Gaustadalléen 21
NO 0349 Oslo

Telephone: +47 22 57 38 00
Telefax: +47 22 60 92 00
E-mail: toi@toi.no

www.toi.no

Institute of Transport Economics
Norwegian Centre for Transport Research

- carries out research for the benefit of society and industry
- has a research staff of around 70 highly qualified staff working in various areas of transport research
- co-operates with a number of organisations, research institutes and universities in Norway and in other countries
- carries out high quality research and assessment projects within topics such as traffic safety, public transport, the environment, travel behaviour, tourism, planning, decision-making processes, transport economics, and freight
- publishes research findings in the Centre's own report series, on the Internet, in the periodical "Samferdsel", as well as in national and international scientific journals and publications
- participates in the Oslo Centre for Interdisciplinary Environmental and Social Research (CIENS) located near the University of Oslo