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,

- 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.