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

Summary:

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

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

Evaluation research is often controversial

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

It is possible to identify good and bad evaluation research

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

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

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

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

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

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

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

Criteria of validity in evaluation research

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

1 The sampling technique used in a study

2 Sample size

3 Measurement reliability, for all variables included in a study

4 The presence of systematic errors in data

5 Choice of technique of analysis

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

7 Publication bias

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

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

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

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

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

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

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

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

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

2 The direction of causality should be clear.

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

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

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

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

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

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

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

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

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

The criteria of validity have been applied in seven journal papers

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

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

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

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

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

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

Meta-analyses can be widely applied in transport research

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

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

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

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

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