SYSTEMATIC REVIEW AND QUANTITATIVE META-ANALYSIS: METHODOLOGICAL FOUNDATIONS AND PRACTICAL APPLICATIONS IN THE DOMAIN OF TRANSPORT POLICY

Dissertation zur Erlangung des Doktorgrades (Dr. rer. soc.)

des Fachbereichs Gesellschaftswissenschaften

der Justus-Liebig-Universität Giessen

Vorgelegt von

Guido Möser

aus Gießen

2006

Supervisor: Prof. Dr. Peter Schmidt

SYSTEMATIC REVIEW AND QUANTITATIVE META-ANALYSIS: METHODOLOGICAL FOUNDATIONS AND PRACTICAL APPLICATIONS IN THE DOMAIN OF TRANSPORT POLICY

Content

1) Introduction

In this dissertation I will try to examine if quantitative methods in synthesising single study findings are helpful in the domain of transport policy. An introductory example deals with the problem:

Suppose there is a community called *A-City,* which is confronted by a lot of traffic. Cars are especially causing air and noise pollution, street damage, traffic jams etc. *A-City*-government is interested in reducing the number of people travelling by car. *A-City* is already running public busses, trains and a school travel system for pupils. The main goal of A-City government is to reduce people travelling by car. Due to a limited budget, *A-City* government has to work cost effectively. In this example, A-City government is right at the starting point of investigating this problem.

Scientists, politicians, policy makers and transportation specialists have wide-ranging information needs and limited budgets. For that reason, they need reliable information on the effectiveness of a large number of different kinds of interventions in the domain of transportation. Moreover, many programmes on national and European policies focus on the problem of an appropriately handling transport policies. For example, *"the European Commission's objective for the next ten years is to refocus Europe's transport policy on the demands and needs of its citizens* (European Commission, 2001)." Measures recommended by the European Commission are for example *improving road safety or preventing congestion.*

A-*City* held a conference, inviting experts and staff from other local governments who had already tried to reduce single occupied vehicle usage (cars and motorcycles). At this conference different kinds of travel interventions were discussed. After the conference, A-City's Government was not satisfied because some of the invited participants reported strong difficulties when implementing different kinds of travel interventions. The implementation of flawed strategies led to massive expenditures and the main goal of reducing car use was not attained. Therefore, A-City's government consulted different sources in order to find out more about the effectiveness of different kinds of policy measures to reduce travelling by car. The main sources, travel policy journals, were reviewed. Since the number of articles reporting different results for the same kind of intervention in reducing car use was very high, A-City's government was dissatisfied again.

Taking into consideration the economic purposes and the limited budget of *A-City Government,* it is necessary for them to implement the most effective car reducing intervention programmes. As a first step, an investigation of other programmes and a literature review will be helpful. If the city government does not obtain satisfying results, a professional researcher should be hired.

A-City government hired a scientific researcher. The researcher collected all literature about travel interventions to reduce car use published so far. Finally, after reading the collected literature carefully, he summarised the results in a report. This report showed no concrete statistics on effectiveness but his own scientific interpretation of the studies.

It seems, that the scientific researcher used techniques of the so called 'narrative reviews' in the example above. In a narrative review, a researcher collects all the relevant information on a special subject to summarise the information for other readers interested in the field. The researcher has to choose which information is relevant and interesting for other readers. This selection process may result in difficulties: "*Narrative reviews do not reveal how the decisions were made about relevance of studies and the validity of the included* *studies.*" (Colins & Fauser, 2004, p. 103). Thus in reporting and summarising the results of synthesising the single study results, "*[d]ifferent reviewers were reaching different conclusions from the same research base and the findings reported often had more to do with the speciality of the reviewer than with the underlying evidence.*" (Colins & Fauser, 2004, p. 104).

The next problem one may assume from the example above is that although narrative reviews are an old and established way to synthesise single study findings, they fail to report concrete statistical parameters on the effectiveness of the results of all studies together. How strong is the overall direction and magnitude of interventions in reducing speed limits to car use? Do interventions lead to significant changes (a reduction of five to ten percent)? Is there an effect at all?

Fortunately, a member of A-City government met an old friend who was a researcher in an institute. This institute was interested in investigating the effectiveness of different kinds of travel intervention programmes. They discussed the problems of A-City government exchanging information on the effectiveness of different travel intervention programmes. The researcher informed his old friend from the A-City government about newer approaches to report statistical parameters on the effectiveness of different kinds of travel intervention programmes. The statistical parameters had been synthesised out of single study results. These approaches used quantitative statistical techniques to solve the problem of summarising single study findings.

In the example above, the researcher talked about *quantitative* methods to summarise research findings from single studies. These methods are the so called s*ystematic review* and m*eta-analysis.* Both methods offer ways to synthesise the results from a large body of empirical studies in a systematic and objective way.

Systematic review and meta-analysis could help the A-City government to obtain a systematic overview of different kinds of travel interventions through synthesising the results from all studies at hand.

Systematic reviews are a further development of the narrative reviews. A systematic review is a better way of summarising research evidence than the narrative review: "*Systematic reviews are like scientific investigations in themselves, using pre-planned methods and an assembly of original studies that meet their criteria as 'subjects'. They synthesise the results of an assembly of primary investigations using strategies that limit bias and random error*." (Cochrane Muscoloskeletal Group, 2006). In other words, a systematic review is any type of review using strategies to avoid bias and including a material and methods section. An example of a strategy used to avoid bias is to identify all relevant studies published since 1985. In the material section of a systematic review all used studies are mentioned. In the method section, methods to summarise the results are described.

Given that there are enough statistical parameters in the collected studies, a meta-analysis could be conducted. Using the approach of meta-analysis, special statistical methods for combining single study findings have been developed. A meta-analysis, sometimes called a quantitative systematic review, is *"…the statistical analysis of a large collection of analysis results from individual studies* *for the purpose of integrating the findings."* (Smith & Glass, 1977). Glass introduced the term "meta-analysis" in one of his studies.

The main difference between systematic review and meta-analysis is that a systematic review does not need a quantitative synthesises of the results, while a meta-analysis is a quantitative analysis using specially developed statistical techniques. One crucial aspect of a meta-analysis is that the outcomes must be reported in a quantitative way. A systematic review may or may not include formal meta-analyses. Sometimes, there is a distinction between qualitative systematic reviews which do not include a meta-analysis and quantitative systematic reviews which do include a meta-analysis.

To consider the size of each study used in a meta-analysis, special weighting approaches have been developed. This is the biggest difference between meta-analysis and using an arithmetic mean. The study size influences the importance of every study, large studies are weighted higher and small studies are weighted lower.

The researcher told his friend that today, most travel intervention programmes have been established in the United Kingdom.

Reviewing the literature one finds that many studies in the United Kingdom report a lot of empirical data. However, the authors frequently used narrative techniques and simple statistics like frequencies and arithmetic means to analyse the data. Hence the subjective way of obtaining results through narrative reviews leads to questionable conclusions. This is a typical problem of narrative reviews.

In one case-study, researchers from Steer Davies Gleave (2003) were very disappointed. In their report they examined the effectiveness of introducing new yellow school buses in five different school areas in the United Kingdom. The main goal of this intervention was to motivate parents not to use the car to bring their children to school. They reported no or sometimes negative effects on public transport use. That is why the researchers from Steer Davies Gleave (2003) were not satisfied.

However the A-City government member was still not pleased. First of all, he thought that it would be interesting to look at travel intervention programmes to reduce car use which had already been carried out in the United Kingdom. He then wondered if it would not make a difference to plan the interventions in the United Kingdom or in Germany. Last but not least, he considered the question if there was a difference in the results whether the intervention was carried out in a rural area or in a large city?

Another important aspect of systematic review and meta-analysis is the influence of possible moderator variables. These include among other variables; country (United Kingdom, Germany, Italy etc.) or area (rural, suburban or urban). Moderator variables show methodological differences among studies, samples and other relevant factors: "*Moderator variables are the keys to explaining differences across studies in the outcomes observed. Their associations with effect sizes provide important clues to why some studies yield large effects while another yield small ones.*" (Lipsey & Wilson, 1999).

Suppose a meta-analysis arrives at a summary effect by combining 21 single studies from different countries with the results of about r ~ 0.2. For example, eight studies were conducted in Australia and 13 in the United Kingdom. Therefore it is crucial to keep in mind the influence of individual factors of each country. To demonstrate further why it is necessary to divide by moderators (country), the summary result combining the eight study results from Australia leads to r ~ 0.1 and in the United Kingdom to r ~ 0.3. Obviously it makes a difference if the travel intervention programme is conducted in the United Kingdom or in Australia. Additionally there are a lot of other moderators influencing the study outcomes.

The A-City government member was still not satisfied. He asked his friend if the A-City government members would be able to understand the calculations and the results of the meta-analysis. The researcher answered that everyone must be able to recalculate the results. Results are normally reported using a simple statistical parameter, like the correlation coefficient r.

In systematic reviews and meta-analysis general, frameworks have been developed - see Higgins and Green (2005) or Cooper and Hedges (1994). These frameworks could help to conduct a synthesis of different study results in a more objective way. Objective because often times the results are reported using a five-stage model of research synthesis, problem formulation, data collection, data evaluation, data analysis and interpretation and public presentation of the results. Working in such an objective way made it possible for nearly everyone to recalculate the results.

The A-City government member asked the researcher, how the results are summarised and how the results could be reported.

Target of a meta-analysis is to calculate the *population effect size*; the summary effect of all studies. This population effect size is easy to understand, even for those unfamiliar with statistics. An effect size produces a statistical standardisation of the study findings.

The population effect size could be a correlation, like Pearson's Correlation Coefficient r. Perhaps the result of the meta-analysis investigating the relationship between the intervention of a speed limit and reduction of car use, is $r \sim 0.1$, summarising 21 studies. The population effect size ($r \sim 0.1$) can be interpreted as a positive (direction) and small effect (magnitude).

In another example, a meta-analysis of the effectiveness in reducing parking lots results in a summary correlation coefficient of about $r \sim 0.15$. This effect is also positive (direction). This means a reduction of parking lots results in reduced car use. The effect is higher than in the first example (magnitude). Comparing both kinds of travel interventions, A-City government should reduce the number of parking lots.

The A-City government member told the researcher, that the public commuting system and infrastructure are old and because of that less frequented by customers. He asked if it wouldn't be the easiest way to invest in new buses or infrastructure. The researcher answered that it is necessary to decide between hard and soft policy measures.

Many studies have been conducted so far, measuring the success of different kinds of travel interventions using hard and soft policy measures. On the one hand, hard policy measures refer to infrastructural, organisational or managing parameters. On the other hand, soft polidy measures try to influence individual decision making by persuasion. That is by changing people's perceptions and motivations. Soft-policy measures in the domain of transport policy are for example: work travel plans, personalised travel planning or school travel planning. In the field of hard policy measures, the *A-City* government could reduce the number of inner city parking lots or invest in new hardware like new public and school buses.

Returning to the example reported by Steer Davies Gleave (2003). The authors described the introduction of new hardware (yellow school buses). Nevertheless, as reported above, the results were disappointing. In their report they described no or sometimes negative effects on parents' habits concerning the transportation of children to their schools. For details and a meta-analytical analysis of the Steer Davies Gleave (2003) results, see Appendix I. This meta-analysis showed a significant negative effect of the introduction of new yellow school buses.

Over the last years there has been a growing interest of transport policy makers in behaviour oriented 'soft' policy measures to reduce private car use. Typical interventions are among others work travel plans, travel awareness programmes, individual marketing, car sharing or school travel plans.

School travel plans are another widely used intervention type. School travel plans "*aim[…] to encourage more families to use environmentally friendly transport options to get to and from schools*". (Energy Efficiency and Conservation Authority, 2006). Alternative measures are to advertise walking to school, to promote the use of the (yellow) school bus or cycle training. The main goal of these soft policy-measures is to foster the use of alternative modes of transportation.

Let us now discuss a second example to show another transport policy problem. In the second example a company is interested in finding a transport policy to reduce single occupied cars used by their staff to commute to work.

A company, let us call it *B-Company*, was interested in reducing the proportion of staff commuting to work by car. Firstly, the management of the company discussed the problem. Secondly, they asked other companies how they managed car reduction. The management of *B-Company* received many different answers. Some companies were experimenting with parking lot restriction and tolling. Others tested the introduction of incentives for using public transport modes, like busses or trains, giving free tickets to their staff.

There are many kinds of intervention to reach the goal of reducing single occupied car use by company staff. For instance, incentives for car-sharing, reducing or tolling the company's parking lots, travel work plans and so forth. The management intends to conduct an intervention while trying to keep the costs as low as possible. Therefore, tolling car lots seems interesting, because extra profits are possible for the company. Obviously this may lead to a loss of motivation among members of the company staff.

B-Company hired a researcher to solve the car use reduction problem. They felt this was necessary because of the many opportunities of travel intervention programmes. The researcher offered the *B-Company* management to conduct a systematic review and a meta-analysis. After hearing about the big advantages of conducting a meta-analysis and systematic review the B-Company management decided to opt for these methods. A meta-analysis and systematic review was run to investigate the best way to solve the car reduction problem.

The *B-Company* management chose the method of work travel plans to reduce car use. Work travel plans offer the best effects on the reduction of staff commuting by car. See for instance chapter three: *Are 'Soft' Policy Measures Effective in Reducing Peoples' Car Use? A Meta-Analytical Review of Research Evidence* and chapter four: *Are Work Travel Plans Effective? – Systematic Review and Meta Analysis in the Transport Policy Domain.* Chapters three and four offer systematic reviews and meta-analyses of studies of soft policy interventions in the transport policy domain. The analyses summarise the direction and magnitude of single effects in order to calculate an overall effect (population effect size) on the effectiveness of different soft policy interventions.

Furthermore the *B-Company* management wanted to know more about the influence of social and psychological factors influencing car use of staff. They wanted to know, for example, if there is an influence of problem awareness of car use and pro-environmental behaviour (i.e. a reduction in car use).

The appropriate model to research determinants of pro-environmental behaviour could be a theoretical one with structural equations modelling based on meta-analytically derived correlation matrices. Structural equation modeling offers an opportunity to test causal relationships and theoretical models. Meta-analysis allowed us to summarise single study findings. For details see chapter five: *Twenty years after Hines, Hungerford, and Tomera: A new meta-analysis of determinants of pro-environmental behaviour*.

At last the single occupied vehicles usage reducing programme (work travel plan) was a success. Meta-analysis and systematic review provided sufficient and reliable data to accomplish this goal.

After looking at the examples which tried to simplify the subject matter, I will now give a brief overview of the different articles. The articles presented in this dissertation are following the historical development of meta-analysis and systematic review.

In the first article, chapter two: '*Meta-analysis: An alternative to narrative review for synthesising social science research?'* methodological foundations and suitability of meta-analysis will be discussed. Theoretical foundations will be shown and practical applications will be critically observed.

In the second article, chapter three: '*Are 'Soft' Policy Measures Effective in Reducing Peoples' Car Use? A Meta-Analytical Review of Research Evidence'* simple meta-analysis techniques will be used to investigate reported univariate single study results. However, in the meta-analysis some of the results showed heterogeneity.

Chapter three will give the results of a meta-analysis synthesising effect sizes obtained in 141 studies. These evaluated the car reduction effect of three identified subgroups in the dataset of soft travel measures: *work travel plans*, *school travel plans* and *personalised travel plans*.

Across all 141 studies, a significant mean effect size of 0.15 (fixed effects model) is estimated. The respective mean effect sizes for the three separate intervention types are: 0.24 for work travel plans, 0.08 for school travel plans, and 0.10 for personalised travel plans. However, the causal inferences one can draw from these results are limited by the weak, quasi-experimental designs used in most of the included intervention studies. Whereas the results of personalised travel plans indicate a homogeneous effect size distribution, this is not the case with the two other methods. The effect size distributions of the other two intervention types are heterogeneous. Results of more detailed moderator analyses are also reported.

In the third article (fourth chapter), *'Are Work Travel Plans Effective? – Systematic Review and Meta Analysis in the Transport Policy Domain'*, a simple univariate meta-analysis like in article two was used. Furthermore, to explain heterogeneity, we discussed the fixed and the random effects models and searched intensively for possible moderators to explain further variance.

In the *fourth chapter*, data on the effectiveness of work travel plans by Cairns, Davies, Newson and Swiderska (2002) and Cairns, Sloman, Newson, Anable, Kirkbride & Goodwin (2004) will be used to demonstrate the application of this strategy in the transportation policy domain. The meta-analytical results differ considerably from the narrative conclusions Cairns et al. (2002 & 2004) draw from the data. They assess the impact of organisational as well as site characteristics as negligible and address parking as the most import success factor of work travel plans. In contrast, the meta-analytical results provide evidence that site and organisational factors as well as characteristics of the monitoring process are strong predictors of the variability in car reduction reported in work travel plan evaluations, whereas parking is of less importance.

Finally, in article four (chapter five), '*Twenty years after Hines, Hungerford, and Tomera: A new meta-analysis of determinants of pro-environmental behavior*', state of the art meta-analytical techniques will be used. This is the combination of meta-analysis and structural equations modeling (MASEM) to look for variance and causality in a more sophisticated way. The combination of meta-analysis and structural equations modeling will be used to investigate research on pro-environmental behaviour since 1986. The first goal of chapter five is to assemble a body of newer studies for an independent replication of the Hines et al. (1986/1987) meta-analytical results. The second goal is to perform a meta-analytical test of a theoretical model of causal determinants of pro-environmental behaviour: use of theoretical models for modelling the interplay of knowledge, behavioural constraints/opportunities as well as personal values and motives in influencing the decision to behave pro-environmentally, see Bamberg and Schmidt (2003) for an overview.

On the whole, chapter's two to five will give an overview of the contemporary state in the domain of travel policy. Possible applications of meta-analysis and systematic review to synthesise single study results to obtain higher evidence levels will be discussed.

Acknowledgements

I would like to very much thank Sebastian Bamberg and Peter Schmidt who, during the work, provided me with useful and helpful assistance. Their advice, many comments and precious time helped me to finalise many parts of my work.

Without the support of my wife Nataliya, my brother Thilo and my parents, this work would most likely have not progressed and matured. Their inspiration was and continues to be a crucial part of my work.

I would like to thank all the people who helped me in discussing the different parts of this work and for their dedication and interest, especially Gero Schwenk, Arlo Ibisch and Klaus Ibisch.

Finally, thank you to Victoria Southworth, who helped me a lot with proof reading my English during the work.

Eingereicht bei der Zeitschrift *European Societies*

2) Meta-analysis: An alternative to narrative reviews for synthesising social science research?

Guido Möser and Peter Schmidt

1 Introduction

The term "meta-analysis" was coined after a paper by Glass in the 1970s [1][1]. Glass defined meta-analysis as *"…the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings."*[2]. Other definitions of meta-analysis describe it as a *collection of statistical techniques for combining studies or as a summary and statistical analysis of the results of several studies testing the same relationship*, see [2].

The important role of research synthesis for contemporary (social) sciences described Hunter & Schmidt (1990) as follows: *Scientists have known for centuries that a single study will not resolve a major issue. Indeed, a small sample study will*

[1] Of course, since early 1900, quantitative methods for combining studies have been available, see [68].

[2] In 1976, Gene V. Glass and Mary Lee Smith meta-analytically investigated if there was any effect of psychotherapy. The results of summarising 375 psychotherapy studies showed that psychotherapy was effective [69].

not even resolve a minor issue. Thus, the foundation of science is the cumulation of knowledge from the results of many studies [3].

Today meta-analysis is accepted as a method of summarising the results of empirical studies within the psychological, health and educational sciences. But, in contemporary political and social sciences meta-analyses are rarely found. Conducting a literature research in the following journals from 2000 up to today showed 7 articles directly related to meta-analysis: American Sociological Review (0), American Journal of Sociology (1) [4]; Annual Review of Sociology (0); European Sociological Review (1) [5]; British Journal of Sociology (0); Sociological Methods and Research (1) [6]; Quantity and Quality (3) [7], [8], [9]; European Societies (0), Sociological Methodology (1) [10], Kölner Zeitschrift für Soziologie (0).

The strength of meta-analytic procedures is to impose transparency and structure on the process of summarizing research findings. Each step will be exactly documented. Specification of criteria must be reported, search strategies, too. It is necessary to report information about the data analysis. Because of this meta-analysis represents, in a more differentiated and sophisticated way, study findings compared with narrative review now used as standard [11]. Meta-Analysis is more systematic and objective compared with the narrative review, it helps to typify a large volume of research literature [12]. The narrative review is another commonly used approach to synthesise and review results. The narrative review relies on a researcher's ability to "*digest the array of findings across studies and arrive at a pronouncement regarding the evidence for or against a hypothesis using some unknown and unknowable … mental calculus*."[13]. For example, the *Annual Review of Sociology* provides overview articles as narrative reviews.

In medicine, detailed and specific checklists and handbooks are available to guide researchers through each step of the meta-analysis. Especially that allows other researchers to check the calculations very exactly. Researchers are recommended to specify their decision rules at each step of the meta-analysis. The results of a meta-analysis should be presented in a standardised form [14].

The Cochrane Collaboration provides a handbook, free downloadable on their website, to support the objective and systematic goals in the field of medicine. Of course, it is also a good example for other scientific fields [15].The handbook is organised in seven sections according to the steps of preparing and maintaining a systematic review: (1) Formulating the problem; (2) Locating and selecting studies; (3) Quality assessment of studies; (4) Collecting data; (5) Analysing and presenting results; (6) Interpreting results; (7) Improving and updating reviews.

Systematic Review is a phrase used in the field of medicine for meta-analysis, although not every Systematic Review must be a meta-analysis. In most papers only the so called *quantitative systematic reviews* equals a typical meta-analysis, for example see [16] Anna Lee, Tony Gin (2002) or [17] Cook D J et al. (1997). In Systematic Reviews, meta-analysis is often a type of systematic review that uses statistical methods to combine and summarise the results of several primary studies[3].

Three of the main statistical approaches to Meta-Analysis are from Hedges and Olkin, Rosenthal and Rubin and Hunter and Schmidt. Focus in the following is on techniques from *Hunter and Schmidt* and *Hedges and Olkin*. These are the most frequently used in psychology and social sciences, see [18].

Limitations of Meta-Analysis exist in, that only quantitative empirical research studies are used. Data will be typically found in articles or other publications and

3 The Systematic Review is described in [70].

normally it is impossible to get the complete data sets. Two important restrictions result from this. The findings must be conceptually comparable and deal with the same constructs and relationships. Secondly the findings must be reported in similar statistical forms. And that is one of the biggest problems in political and social sciences.

Next, we present how to conduct a meta-analysis. For this purpose we describe Cooper's (1994) five stage model of research synthesis. The model encompasses problem formulation, data collection, data evaluation, data analysis and interpretation and public presentation. We present examples within the five steps from published studies to demonstrate applications of meta-analysis. In the last section, advantages and disadvantages will be discussed. An outlook of further developments will be given at the end.

2 Conduction of a meta-analysis

STEPS OF A META-ANALYSIS

Cooper [19] developed a five-stage model of research synthesis. The steps in Cooper's five-stage model of research synthesis are *problem formulation*, *data collection, data evaluation, data analysis and interpretation* and *public presentation,* as shown in Table 1. Especially the two stages of data analysis and interpretation and data collection require a lot of time.

Table 1: Five-stage model of research synthesis (Cooper 1994)

Step	Feature	Content
1	Problem formulation	Defining clearly the question to be asked: Developing a coding form (survey protocol)
2	Data collection	*Searching the literature*: Identifying a research strategy. Sample or population of research reports is gathered
3	Data evaluation	*Coding the Literature*: Studies retrieved coded on critical features: Each research study analyzed by a coder
4	Data analysis and interpretation	Calculating and interpreting the population effect size: Analyzing resulting data using special adaptions of conventional statistical techniques
5	Public presentation of results	Describing the pattern of findings in the selected set of studies

PROBLEM FORMULATION

To specify the problem to be investigated or the question to be answered it will be necessary to start with a statement, which will guide the selection of the research studies [20]. The problem specification process is an iterating process. It starts with the statement and will be modified while information will be found and analysed from relevant studies. Also, it is necessary to identify the form of the research findings to be meta-analysed. Of course, findings must be in form of quantitative data. It is not necessary to look for the same effect size, because there are transformation methods to convert them into each other [21].

The effects of violent video games on aggression have been widely discussed in politics, public and sciences. Sherry conducted a meta-analysis about that topic ("The Effects of Violent Video Games on Aggression – A Meta-Analysis"). The author wanted to find out, if there is any effect on aggression by video-games on children and adolescents, and, if so, how strong the effect is: *Researchers have hypothesized that playing violent-content video games will result in aggression (…) A meta-analysis of existing studies of the effects of violent content video games on aggression will provide important information for consideration of theoretical and methodological issues leading to a systematic program of research in this area* [22].

Reducing inter-group prejudice is one field of interest in social psychology. Pettigrew and Tropp conducted a meta-analysis about reducing inter-group relations and prejudice *through contact* between groups. They defined inter-group contact as "*actual face-to-face interaction between members of clearly distinguishable and defined groups*".

They derived inclusion criteria from that definition for the research process: First, they "*considered only those empirical studies in which intergroup contact acted as a causal, independent variable for intergroup prejudice".* Second, they included "*only research that involved contact between members of discrete, clearly distinguishable groups".* Third, "*to be included, the research had to involve some degree of direct intergroup interaction".* Fourth, "*the prejudice dependent variables had to be collected on individuals, rather than simply as a total aggregate outcome; and comparative data had to be available to evaluate any changes in prejudice"* [23].

DATA COLLECTION

Finding nearly all relevant references to a chosen topic is one of the most important and most difficult tasks in providing a good meta-analysis. This is necessary, because a meta-analysis should give an overview over all studies so far done. It is important to collect all studies, also unpublished ones. Because of that, it is highly recommended to work in a highly systematic and objective way during the data collection process. Copies of collected studies must be obtained to screen and to code for inclusion of the studies in the meta-analysis. Sometimes it is possible to get documents by digital databases; sometimes it is necessary to get printed copies from the library. Hard to get, in which way ever, are unpublished studies[4] [24]. Different sources are shown in table 2. One of the most successful sources are journal articles from relevant journals to the researched topic[5], [6]. Also books related to the interesting topic should be investigated.

Table 2: Possible References

A	Review Articles
B	References in Studies
C	Computerized Bibliographic Databases
	"Keyword Searches": Sociological Abstracts, GESIS SozioGuide (Gesellschaft Sozialwissenschaftler Infrastruktureinrichtungen e.V.), SOSIG (Social Science Information Gateway), (…)
D	Bibliographic Reference Volumes
E	Relevant Journals
F	Conference Programs and Proceedings
G	Authors and Experts
H	Government Agencies

Another problem is biased studies. If the studies included in the meta-analysis are biased, so are the meta-analysis results. For example, conducting only a

[4] As an example, there is the IZ (*Informationszentrum Sozialwissenschaften*) in Germany, where one could search also for grey literature, see http://www.gesis.org/IZ/ for further details.

[5] One advantage is,that the journals are mostly on the latest scientific state [71]. One can start here a literature reseach. Looking in the literature chapter provides a good source for looking for other cited related articles to the interesting topic.

[6] E.g. a study about environmental behavior could be started in journals like Environment and Behavior, Environmental Education Research, Journal of Environmental Planning and Management.

classical literature research will bring more published and significant studies, non-significant studies will be underrepresented. That is called *publication bias* or *file drawer problem*. Non significant studies will be more often not published in journals than significant results. Therefore statistically significant studies are more likely to be included in the meta-analysis which will result in an overestimate of the treatment effect [25]. To identify unpublished studies or heavily reachable studies, possible sources to look for are the university library, doctoral dissertations (database "Dissertation Abstracts International"), conference papers, journal articles, technical reports, information directly from the author, agencies and institutions publishing or sponsoring the research of interest [26]. There are statistical methods to investigate for publication and for other kinds of biases, e.g. the *funnel plot*.

Sherry conducted a search of literature *from January 1975 to July 2000. Sherry* used the following databases and sources: Education Resources Information Center (ERIC), PsychInfo (database for psychological abstracts), Dissertation Abstracts International, Communication Abstracts and Psychological Abstracts. He used the following keywords: *video game* or *computer game.* He found more than 900 citations. Correspondence was undertaken by Sherry with other researchers in the field to locate unpublished studies. At the end of the research process Sherry found 32 independent studies in which violent video game play was the independent variable and some measure of aggression was the dependent variable. After excluding some studies because unavailable (1 study), three studies were excluded because they did not provide a usable control group for comparison, two studies were excluded because they lacked a usable measure of aggression and one study was excluded because the reported data sets were not interpretable. At the end, Sherry included 25 usable studies in his meta-analysis [27].

Pettigrew and Tropp located over 200 studies which met the inclusion criteria. They used in the psychological field *PsychLit* (database), in the sociological field *Sociological abstracts* (*SocAbs*, formerly *sociofile*) (database) , in the political field *GOV*, in the educational field *Educational*

Resources Information Center (ERIC) and for general research *Current Contents*. Pettigrew and Tropp also wrote letters to researchers related to the field to ask for unpublished papers - unfortunately, they reported no information about any success of this action in their article. They reviewed reference lists from previously located studies [28].

DATA EVALUATION

Studies found during the research process have to be coded on interesting features. Each research study is analysed by a coder. This is a very sensitive step, computerisation is nearly impossible. That is, because the articles found have to be read very critical for all collectable information. Coders should be trained before the research and coding process. If possible, two coders should code every study. Study results found by the two coders should be compared and must be as equal as possible.

One of the biggest problems is to investigate the s*tudy quality.* Combing poor quality studies may conduct biased and potentially misleading results [29]. The best research design is the randomised controlled design (RCT), often called 'Gold Standard', followed by Cohort studies, case-control studies, case series, single case reports, ideas and opinions and at last laboratory experiments. For a general discussion see [30].

Another big problem is, that numerous factors can cause variation in effect sizes, so called moderators. Information about measurement reliability, range restrictions, reporting errors, within-study statistical adjustments, unreported factors, age, sex, etc. must be collected. That is necessary to perform further analyses of the variance, especially if there is no homogeneity [31] between single study findings. There is always potential for differences across studies that may be confounded with treatments used. For example, if studies in a meta-analysis were done in different countries, cultural differences may be confounded with treatment differences. These confounding variables should be coded as possible moderators, to investigate later their influences on effect sizes.

At least a dataset using a spreadsheet program, a database or a statistical program, like SPSS or Stata, should be used. Further documentation is necessary [32].

DATA ANALYSIS AND INTERPRETATION

Two steps have to be carried out in the data analysis stage of a meta-analysis. First, it is necessary to *integrate* the single findings and second, the variance has to be examined.

Meta-analysis represents each study's findings in the form of effect sizes. For every relevant study found during the research process an effect size estimate must be calculated. The underlying Concept is a Concept of "standardisation". The *effect size statistic* produces a statistical standardization of the study findings. In that way, numerical values will be interpretable in a consistent way across all the variables and measures involved.

The *integration* of the single findings could be done through different methods. Some reviews categorise findings as significantly positive (favouring the treatment group), significantly negative, or non-significant. The category with the most entries is considered the best representation of research in this area. Rosenthal gives an introduction about the integration of significance levels of single studies [33]. The also so called *Vote-counting* confounds treatment effect and sample size because statistical significance is a function of both. Given the modest power of typical educational research to detect true effects as statistically significant, conclusions from vote-counting can be very misleading [34], [35].

Typical effect sizes used in meta-analysis are correlation coefficients, *r*, Standardised mean differences, d, use of dichotomous measures (differences in proportion), Odds ratios, *o*, Rate differences or Risk differences [36]. A lot more various classes of Effect Sizes exist. For a wider overview about used effect sizes, see appendix [37]. In practice, which effect size to choose depends on the different research situations. The effect size encodes the selected research findings on a numeric scale. Shown here are the most common in the field of social research.

The *standardised mean difference effect size (d)* is calculated by (1). A study design with both, intervention and control group, is necessary. The persons have to be divided randomly between intervention and control group, otherwise it is no *randomised* control trial. Randomised control trials (RCT) reflect the highest evidence level. The effect size is called *d*, and is calculated by the difference between intervention and control group, divided through the pooled within-groups standard deviation, s_{pooled}, given by formula (2) [38].

(1) $$ES_d = d = \frac{\overline{X}_1 - \overline{X}_2}{s_{pooled}}$$

$\overline{X}_1$ is the mean of the intervention condition;

$\overline{X}_1$ is the mean of comparison condition;

s_{pooled} is the pooled within-groups standard deviation

Whereby s_{pooled} is calculated by (2).

(2) $$s_p = \sqrt{\frac{(n_{G1}-1)s_{G1}^2 + (n_{G2}-1)s_{G2}^2}{(n_{G1}-1)+(n_{G2}-1)}}$$

s_p is the pooled standard deviation

s_{G1} is the standard deviation for Group 1

s_{G2} is the standard deviation for Group 2

n_{G1} is the number of subjects in Group 1

n_{G2} is the number of subjects in Group 2

In case of small samples, Hedges and Olkin recommended correction of *d* using of formula (3).

(3) $$d' = \left[1 - \frac{3}{4N-9}\right]d$$

d is the standardised mean difference effect size;

N is the total sample size

Pettigrew and Tropp used the *standardized mean difference effect size* (*d)* (also called *Cohen's d*) as effect size statistic. Information found reported as correlation coefficients (*r)* were converted into *d*. "*The weighted mean estimate for the contact-prejudice effect size among the 203 studies was a Cohen's d of -.42 (or a mean r of -.20).(…) Though in most empirical contexts this effect size would be considered "small" to "medium" in magnitude".* If there is contact to other groups, prejudice is reduced in a small to medium level [39].

Lösel and Beelmann (2003) conducted a meta-analysis on social skills training as a measure for preventing antisocial behaviour in children and youth. The best estimate mean-effect, they found, were d = 0,38 in post-intervention. "*The meta-analysis of the follow-up measures were d =*

0,28. These effects are in the middle range". The results shows, that if there is a social skills training, antisocial behaviour will be slightly reduced.[40].

The variance for the standardised mean difference for small samples is calculated by formula (4). Sample size for the intervention and comparison groups must be known.

(4) $$v_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d'^2}{2(n_1 + n_2)}$$

n_1 and n_2 are the sample sizes for the intervention and comparison conditions

The odds ratio, *o*, effect size, is calculated by formula (5). It "*represents the effect of an intervention as the odds of a favourable (or unfavourable) outcome for the intervention group relative to the comparison group."* [41]. It represents a ratio of two odds, e.g. number of patients living after a medical intervention versus patients died.

(5) $$ES_{OR} = o = \frac{ad}{bc}$$

a and *c* are the number of successful outcomes in the intervention and comparison conditions;

b and *d* are the number of failures in the intervention and comparison conditions (based on a 2 x 2 contingency table)

As an asymmetric measure, the Odds-Ratio has a complex standard error formula. Negative relationships are indicated by values between 0 and 1. Positive relationships are indicated by values between 1 and infinity. The interpretation of the relationships is not easy using the Odds-Ratio as effect size. One possible solution is to transform the Odds-Ratios into the natural log of the Odds-Ratio, see formula (6). In the natural log of the Odds-Ratio are smaller than 0. If there is no relationship, it is indicated by the value of 0. Positive relationship are indicated by values bigger than 0. Results can be converted back into Odds-Ratios by the inverse natural log function, see for the back-transformations of logged odds ratio formula (7) [42].

(6)	$lor = \log(o)$	Log transformation of the odds ratio
(7)	$o = e^{lor}$	Logged odds ratio (lor) transformed into an odds ratio (o); e is the constant 2.7183

Variance for the natural log of the odds-ratios is calculated by (8).

(8)	$v_{lor} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$	The variance for the logged odds ratio; a, b, c and d are the cell frequencies of a 2 x 2 contingency table

MacKenzie et al (2001) provided a meta-analysis (systematic review) about effects of correctional boot camps on Offending. The authors wanted to find out, if boot camps are a better way on offending. They used the odds ratio to calculate effect sizes. They reported an overall mean odds ratio 1.92. The reported confidence interval was 0.90 to 1.17. That "*indicates an almost equal odds of recidivating between the boot camp and comparison group"* [43]. Boot camps are showing no significant effect on offending.

The *Correlation coefficient effect size* represents the relationship between two variables, see formula (9).

(9) $$ES_r = r = r$$ r is the Pearson product-moment correlation coefficient between the two variables of interest

In practice, Fisher's z transformation (formula (10)) is used, because Hedges and Olkin showed, that the correlation coefficient r depends strongly on the unknown true value of the correlation. So the approximate distribution is not very accurate. The solution is Fisher's Zr transformation. It normalises the distribution of r [44], [45].

(10) $$z = .5\log\left(\frac{1+r}{1-r}\right)$$ Fisher's transformation of the correlation effect size

Variance of Fisher's z is calculated by formula (11), whereby N is the total sample size.

(11) $$v_z = \frac{1}{N-3}$$ The variance for the Fisher's transformed correlation coefficient;

N is the total sample size

Finally results can be converted back into "r" with the inverse Zr transformation, given by formula (12).

(12) $r = \frac{e^{2z} - 1}{e^{2z} + 1}$ Transforms the effect size z from equation 6 back into a correlation;

e is the constant 2.7183

Hackney and Sanders conducted a meta-analysis to the topic "Religiosity and Mental Health. Thirty-four (34) studies were located that tested the effects of religiosity and mental-health. Statistic to compute was the Pearson product-moment correlation, other study findings were converted into *r.* Many of the located studies included multiple measures of religiosity and mental-health. Because of this, the final data set consisted of 264 correlations. The average effect testing the relationship between religiosity and mental health is 0.10, with a CI ranging from 0.10 to 0.11. The relationship is significant on a 0,01% level [46]. The result shows a small and positive effect between religiosity and mental health.

There are many procedures for computing effect sizes values from study reports. In many studies, test statistics to obtain common effect size indices, ES_d, ES_o and ES_r, are reported in different ways, such as t-tests or F-test values. So it is frequently necessary to transform the statistical information available in a report to extract or, at least, to estimate the effect size value. Formulae for direct calculation for the common effect sizes are reported above. For a broad overview, see [47].

For example, as shown in Lipsey and Wilson, p .172, 173, it is possible to calculate ES_d out of a t-value. For example, a study reports a t-value of 1.68, favouring the treatment group with treatment n = 10 and comparison group n = 12. Using formulae (13), ES_d is about 0.72 [48].

(13) $$ES_d = t\sqrt{\frac{n_1 + n_2}{n_1 n_2}} \quad = 1.68\sqrt{\frac{10+12}{(10)(12)}} = 1.68\sqrt{\frac{22}{120}} = 0.72$$

Sherry used only randomised control trials, but transformed *d* througha conversion formula into r, so that the results are better interpretable.

In a meta-analysis, larger studies should carry more "weight" in the analyses than smaller studies. A simple approach is to weight each effect size by its sample size. A better approach is to weight by the inverse variance [49], because studies generally vary in size. The standard error (SE) is a direct index of effect size precision. The standard error is the standard deviation of the sampling distribution (the distribution of values we would get if we drew repeated samples of the same size and estimated the statistic for each). Because a larger standard error corresponds to a less precise effect size value, the actual weights are computed as the inverse of the squared standard error value – the so called *inverse variance weight.* SE is used to create confidence intervals. The smaller the standard error, the more precise the effect size. The inverted variance weight is in the case of fixed effects meta-analysis calculated by (14). For example, an effect size based on 100 subjects is assumed to be a more "precise" estimate of the population effect size than is an ES based on 10 subjects. That is one reason that optimal weights must be based on the standard error of the effect size.

(14) $$w = \frac{1}{v} = \frac{1}{se^2}$$ The inverse variance weight; v is the variance, which calculation depends on the kind of the effect size, see equations 2a, 3 and 4

The major goal of every meta-analysis is the estimation of the *population effect size and associated statistics* (15).

(15) $\overline{ES} = \frac{\sum(ES \cdot w)}{\sum w}$

Weighted mean effect size,
where ES is the effect size index (see equations (1), (2), (3) and w is the inverse variance weight (12)

Standard error of mean effect size is calculated by (16). Standard error is necessary to calculate corresponding confidence intervals of the mean effect size.

(16) $se_{\overline{ES}} = \sqrt{\frac{1}{\sum w}}$

The standard error of the mean effect size

The confidence interval for the mean effect sizes will be calculated by (17) and (18). If confidence involves zero, mean effect size is not significant.

(17) $LowerCI = \overline{ES} - 1.96 se_{\overline{ES}}$

Lower bound of the 95 percent confidence interval

(18) $UpperCI = \overline{ES} + 1.96 se_{\overline{ES}}$

Upper bound of the 95 percent confidence interval

Calculating z-value of standard normal distribution, see (19) is an equivalent method to test significance. The z-value should be higher as 1.96, assuming an error-probability of less than 5% to be sure about a mean effect (population effect) [50].

(19) $z = \dfrac{\overline{ES}}{se_{\overline{ES}}}$ A z test; tests whether $\overline{ES}$ is statistically greater than or less than 0

The mean effect size, effect sizes of each study and corresponding confidence intervals could be plotted in a forrest plot [51], because the forrest plot visualises the relationship between effect and study size in very informative way.

The second step is the investigation of the variance. The investigation of the *heterogeneity* between the different studies is a main task in each review or meta-analysis[7]. A *Homogeneity Analysis* is necessary to check if all studies come from the same population or not. For the quantitative assessment of heterogeneity, several statistical tests are available. It is also recommended to investigate the heterogeneity informally by comparing results from studies with different designs, maybe within different geographical regions. Additionally, graphical methods should be used to visualise heterogeneity, such as plots with single studies grouped or ordered according to special co-variables as type of study, publication time, etc., or funnel plots[8] to indicate publication bias, and radial plots. Formula for homogeneity test Q is given by (20) [52].

---

7 Homogeneity analysis tests whether the assumption that all of the effect sizes are estimating the same population mean is a reasonable assumption. If homogeneity is rejected, the distribution of effect sizes is assumed to be heterogeneous. Single mean ES is not a good descriptor of the distribution. There are real between study differences, that is, studies estimate different population mean effect sizes. Two options: model between study differences or fit a random effects model.

8 As written above, a graphical test for detecting the presence of publication bias is the funnel plot. The plot looks like a funnel, if there is no publication bias. The funnel plot compares the effect size (*x-axis*) against the sample size or the standard error (*y-axis*). *Normally,* studies with a smaller sample size should have larger sampling error, and studies with a larger sample size should have lower sampling error [72].

(20) $$Q = \sum \left(\overline{ES}^2 \cdot w \right) - \frac{\left(\sum \left(\overline{ES} \cdot w \right) \right)^2}{\sum w}$$

Homogeneity test Q;

distributed as a chi-square, degrees of freedom equals the number of effect sizes less 1

Another important aspect of meta-analysis is the relationship between effect sizes and moderator variables, like type of study design of an involved study, percentage of men and women in the sample or type of intervention. Moderator variables represent differences among studies in their methods, samples, and interventions [53]: Moderator variables are the keys to explaining differences across studies in the outcomes observed. Their associations with effect sizes provide important clues to why some studies yield large effects while another yield small ones[9]:

MacKenzie et al (2001) reported several moderator variables to investigate sources of heterogneity in their meta-analysis. For example, they used the following method variables: *Qualitative methodological quality score* (Random assignment, not degraded; High-quality quasi experiment; Standard quasi experiment; Poor-quality quasi experiment), .*Randomly assigned participants to conditions* (yes/no), *Used group-level matching* (yes/no), *Prospective research design* (yes/no), *Used statistical controls in analysis* (yes/no), *Boot camp dropouts in analysis* (yes/no), *No overall attrition apparent* (yes/no), *No differential attrition apparent* (yes/no).

They used the following program characteristics, *Aftercare treatment component* (Juveniles (yes/no) vs. Adults(yes/no)), *Academic education* (Juveniles (yes/no) vs. Adults(yes/no)), *Vocational education* (Juveniles (yes/no) vs. Adults(yes/no)), *Drug treatment* (Juveniles (yes/no)

9 Lipsey, Mark W. (2003), wrote a very interesting article about the influence of Moderators and their interaction between each other [73].

vs. Adults(yes/no)), *Counseling (group and individual)* (Juveniles (yes/no) vs. Adults(yes/no)), *Manual labor* (Juveniles (yes/no) vs. Adults(yes/no)) [54].

Testing a random effects model instead of a fixed effects model belongs to the findings of the homogeneity analysis. There are two possible methods, the *fixed effects analysis* for homogeneous distributions and the *random effects analysis* for heterogeneous distributions. Which method one should choose depends on the population and the results found in the Q-Test. In the *fixed effects model*, it is assumed that the underlying true exposure effect in each study is the same. The overall variation and, therefore, the confidence intervals will reflect only the random variation within each study but not any potential heterogeneity between the studies. If individual data is available, the pooled estimator and its variance can be obtained using regression models by incorporating additional dummy variables. The *random effects model* incorporates variation between the studies. It is assumed that each study has its own (true) exposure effect and that there is a random distribution of these true exposure effects around a central effect. The observed effects from the different studies are used to estimate this distribution. In other words, the random effects model allows for non-homogeneity between the effects of different studies. Fixed effects model inverse variance weights are given by (14). Calculating the random effects variance component and weight is a little more complicated; see details in [55].

A Comparison of Random Effect with Fixed Effect Results shows, that the biggest difference to notice is the significance levels and confidence intervals. Confidence intervals will become larger. Effects that were significant under a fixed effect model may no longer be significant. Random effects models are therefore more conservative.

PUBLIC PRESENTATION OF RESULTS

As written above, it is necessary to publish how the scientist(s) calculated the results in a highly systematic way. Table 3 shows the typical contents of a public presentation of the results. The representation of the results of different meta-analyses normally equals each other in the way of their representations of the results. This allows to assess the quality of the meta-analysis and to recalculate the results.

Table 3: Report of Meta-Analysis Results

a.	Abstract or executive Summary
b.	Background information
c.	Hypotheses tested/ question to be addressed in the review
d.	Methods of Review
e.	Details of studies included in the review
f.	Details of studies excluded in the review
g.	Results of Meta-Analysis
h.	Report analysis of the robustness of the results
i.	Discussion
j.	Implications of the Review
k.	References
l.	Dissemination and further research

First, an *abstract or executive summary as introduction* should be reported, followed by a presentation of b*ackground information*. Very important is the purpose of the review, *the hypotheses tested* and *the question to be addressed. Methods of the review* should be reported, especially including the search strategy, assessments of relevance, validity, data extraction and synthesis. A big problem is normally to report details of the studies included in the review because of the large number of studies and information. At least, the references and sources of the studies included should be published. A must to publish are demographic details of groups, year in which the study took place, interventions and outcomes of each study and study design, also information about quality and validity.

Another problem is excluded studies. In many meta-analyses, a lot of studies found during the research process are excluded because of missing information, less statistical information or poor study quality. So, *details of studies excluded in the review* and *Reasons for exclusion* should be provided.

The r*eporting of results of a meta-analysis* is one of the most important parts. First, report point estimates of each study, standard errors and corresponding confidence intervals. Second, report pooled estimates and corresponding standard errors. Information should be presented about fixed or random effects estimates, corresponding confidence intervals and *p*-values of tests. Third, provide a tabular summary of the relative weight of each study. Fourth, provide the result of test of homogeneity, the Q-value and corresponding *p*-value. Fifth, report results in absolute terms (e.g. absolute relative risk or number needed to treat. This allows possible impacts to be assessed.

As a next step, it is recommended to *report analysis of the robustness of the results*. Where is uncertainty or missing data? To assess the robustness of the results it is necessary to perform a sensitivity analysis.

The chapter *discussion* should include discussions on, firstly, the strength of the causal evidence and second, potential biases in the studies and the review. Other topics to discuss are limitations on inferences and potential *implications* of the results for social or political sciences and future research.

Reporting *References*, three lists should be given. Firstly, studies included in the review, secondly, studies excluded in the review and thirdly, other references cited in the review. [56].

There are different areas in which *graphical methods* could be used during a meta-analysis, to discover patterns and relations among variables in a meta-analysis or to check statistical assumptions on which numerical analyses are based. For example, *Funnel* plots and *normal quantile* plots could be used to investigate if all studies come from a single population and to search for publication bias. *The Distribution of effect size* could be shown *by* Boxplots, Stem-and-leaf plots and QQ-Plots [58]. Assessment of heterogeneity could be analysed by Histograms or Forest and L'Abbe plots.

If one is searching for *Publication bias*, also often called the "file drawer problem", the shape of a funnel plot can suggest whether such an effect exists or not. If publication bias exists, the funnel plot shows up in a different form on the funnel plot see [59], [60]. The funnel plot is the most established form of graphical analysis in the field of meta-analysis. A funnel plot is a two-dimensional graph. On one axis is drawn sample size and on the other the effect size. The name funnel plot implies, that sampling error decreases while sample size increases, so that the figure displayed should look like a funnel. If all studies come from a single population, the plot should look like a funnel with the diameter of the funnel decreasing as sample size increases. As sample size increases, the effect size estimates narrow in on the true population effect size [61].

The *publication bias* results in it being easier to find studies with a positive result. Research with statistically significant, positive, results is often found to be more likely to be submitted, published or published faster than that with non-significant results. But there are a lot more reporting biases [64].

Table 4: Possible Biases

No	Kind of bias	Positive results…
1	Publication bias	It is more likely to be published with positive results
2	Time lag bias	It is more likely to be published rapidly with positive results
3	Language bias	It is more likely to be published in English
4	Multiple publication bias	It is more likely to be published
5	Citation bias	More likely to be cited by others

Lit.: Cochrane Collaboration 2002, module 15, page 1.

As a result, all of these biases make it more likely to find positive studies than those with non-significant results. The influence from these biases could be minimised by extensive searching [65].

3 Summary and outlook

We have argued that meta-analysis is a more adequate method of synthesising and reviewing empirical results than the narrative review, which is predominantly used in political science and sociology for this purpose. By using the techniques of meta-analysis all the different steps of problem formulation, literature search, data collection, data evaluation and summarising the empirical evidence over different studies become much more transparent. Furthermore the synthesis of findings is not ad- hoc but has a statistical foundation.

Besides the publications themselves increasing availability of largely standardised data-generating programs by data archives like the European

Household Panel Study (ECHP), the European Social Survey (ESS), the American General Social Survey (GSS), the German General Social Survey (ALLBUS) the International Social Survey programme (ISSP), the Socio-Economic-Panel (SOEP) and the World Value Study (WVS) allow to conduct meta-analyses much easier than in the past.

One drawback may be that a meta-analysis for a certain topic is more time consuming than a narrative review. Furthermore it requires specialised knowledge about underlying methods and statistics including the possible use of a computer programme.

Combining “Oranges and Apples” that is the validation of comparability of single studies remains an important methodological problem both for narrative reviews and metaanalyses14. Since many study designs are possible, it is necessary to evaluate the comparability of the single studies before conducting a review. Evaluation can be conducted partly from published data if enough detailed information is available in the paper. If individual data is available from the authors or the national data archives like Essex, Cologne or others an analysis of the single studies in one common model is possible. A major reason for different results across studies is that different statistical methods/models have been used. Hence heterogeneity can be significantly reduced in a pooled analysis by using the same model for all studies.

The validity of a meta-analysis is dependent on careful attention to design, conduct analysis, and reporting, as is the validity of any other scientific study. Careful consideration and clear documentation of the research questions, procedures, assumptions and methods, supplemented by sensitive analysis will help to secure high-quality meta-analyses that are persuasive to even the most diehard critics [64].

A new and very promising development is the use of structural equation modelling and multi-level- analyses to meta-analyses, which seems very promising to take into account issues of reliability and validity of the measures used and heterogeneity of studies and samples. A search in data bases like sociological abstracts show only a few articles about this topic, for example Frye,Crissie Marie(2001) [65] and Eddy, Erik R.(2000) [66]. The application of structural equation modelling has the advantage to test additionally all the operationalisations via multi-group analysis. An additional advantage is, that not all relationships specified by a theory need to be included in each primary study. It is possible, that 10 studies report the relationship between constructs A and B; 10 other studies report the relationship between B and C or A and C. The correlations between A,B and C can be estimated by using meta-analysis, although no individual study has included all three constructs [67]. In addition, one can test for mediators and moderators and compute partialised co-effiecients.

Reference

[1] Smith, M. L., & Glass, G. V. (1977). *Meta-analysis of psychotherapy outcome studies*. American Psychologist, 32, 752 – 760.

[2] Gelman, Andrew; Carlin, John B.; Stern, Hal S. & Rubin, Donald B. (1995) *Bayesian Data Analysis*, p. 148, 149.

[3] Hunter, John E., & Schmidt, Frank L. (1990). Methods of Meta-Analysis. Correcting Error and Bias in Research Findings, p. 13.

[4] DiPrete (2002). Life Course Mobility. In: American Journal of Sociology, Vol. 108, p. 267-309.

[5] Verhoeven, Willem-Jan; Jansen, Wim and Dessens, Jos (2005). Income Attainment During Transformation Processes: A Meta-Analysis of the Market Transition Theory. In: European Sociological Review, Vol. 21, p. 201 – 266.

[6] Lensvelt-Mulders, Gerty J.L.M.; Hox, Joop J.; van der Heijden, Peter G.M. & Maas, Cora J.M. (2005). Meta-Analysis of Randomized Response Research: Thirty-Five Years of Validation. In: Sociological Methods and Research, Vol. 33, p. 319 – 348.

[7] Maas, Cora J. M.; Hox, Joop J. & Lensvelt-Mulders, Gerty J.L.M (2004). Longitudinal Meta-analysis. In: Quality and Quantity, Vol. 38, Number 4, p. 381 – 389.

[8] Castro, María & Gaviria, José-luis (2000). Application of Hierarchical Linear Models to Meta-Analysis: Study on the Monte Carlo Simulation on the Functioning of Traditional and Empirical-Bayes Effect Size. In: Quality and Quantity, Vol. 34, Number 1, p. 33 – 50.

[9] Borgers, Natascha; Sikkel, Dirk & Hox, Joop (2004). Response Effects in Surveys on Children and Adolescents: The effect of Number of Response Options, Negative Wording, and Neutral Mid-Point. In: Quality and Quantity, Vol. 38, Number 1, p. 17 – 33.

[10] Chang, LinChiat, and Krosnick, Jon A. (2003). Measuring the frequency of regular behaviours: Comparing the "typical week" to the "past week", Vol. 33, 1, p. 55 – 80.

[11] Miller, Norman, and Pollock, Vicki E. (1994). Meta-analytic synthesis for theory development. In: Cooper, Harris, & Hedges, Larry V. (1994). *The Handbook of Research Synthesis*, Russell Sage Foundation, New York, p. 467.

[12] Lipsey, Mark W. & Wilson, David B. (2001). *Practical Meta-Analysis*. SAGE Publications, p. 1.

[13] Wilson, David B. (2001). *Meta-Analytic Methods for Criminology*. The Annals of the American Academy, 578, November 2001, 71 – 89.

[14] The Cochrane Collaboration, http://www.cochrane.org/, 04.12.2005, 15:52 h. http://www.cochrane.org, 04.12.2005, 15:39 h.

[15] The Cochrane Collaboration, http://www.cochrane.org/, 04.12.2005, 15:52 h. Handbook and other helpful information free downloadable, http://www.cochrane.org/resources/handbook/index.htm, 04.12.2005, 16:02 h.

[16] Anna Lee, Tony Gin (2002): *Applying the Results of Quantitative Systematic Reviews to Clinical Practice*, in: Anesth Analg 2002; 94; 372- 377 or Cook D J et al. 1997, p. 376-380)

[17] Cook DJ, Mulrow CD, Haynes RB. *Systematic Review. Synthesis of Best Evidence for Clinical Decisions*. Ann Intern Med 1997; 126(5): 376-380.

[18] Schulze, Ralf (2004). *Meta-Analysis. A Comparison of Approaches*. Hogrefe & Huber, p 19 – 22 and p. 56, 61, 62.

[19] Cooper, Harris and Hedges, Larry V. (1994). *Research Synthesis as a scientific enterprise*. In: Cooper, Harris, & Hedges, Larry V. (1994). *The Handbook of Research Synthesis*, Russell Sage Foundation, New York, p. 9 – 13.

[20] Halvorsen, Katherine Taylor (1994). *The Reporting Format*. In: Cooper, Harris, & Hedges, Larry V. (1994). *The Handbook of Research Synthesis*, Russell Sage Foundation, New York, p. 427.

[21] Halvorsen, Katherine Taylor (1994). *The Reporting Format*. In: Cooper, Harris, & Hedges, Larry V. (1994). *The Handbook of Research Synthesis*, Russell Sage Foundation, New York, p. 427.

[22] Sherry, John L. (2001). *The Effects of Violent Video Games on Aggression. A Meta-Analysis*. Human Communication Research, Vol. 27, No. 3, July 2001, p. 409 – 431.

[23] Pettigrew, Thomas F., and Tropp, Linda, R. (2000). Does intergroup contact reduce prejudice? Recent meta-analytic findings. In S. Oskamp (ed.), Reducing prejudice and discrimination (pp. 93 – 114). Mahwah, NJ: Erlbaum.

[24] Lipsey, Mark W. & Wilson, David B. (2001). Practical Meta-Analysis. SAGE Publications, p. 24 - 31.

[25] Rosenthal R. *Meta-analytic procedures for social research*, Sage Publications; CA, 1991

[26] Lipsey, Mark W. & Wilson, David B. (2001). *Practical Meta-Analysis*. SAGE Publications, p. 31 - 33.

[27] Sherry, John L. (2001). *The Effects of Violent Video Games on Aggression. A Meta-Analysis*. Human Communication Research, Vol. 27, No. 3, July 2001, p. 409 – 431.

[28] Pettigrew, Thomas F., and Tropp, Linda, R. (2000). Does intergroup contact reduce prejudice? Recent meta-analytic findings. In S. Oskamp (ed.), Reducing prejudice and discrimination (pp. 93 – 114). Mahwah, NJ: Erlbaum.

[29] Lipsey, Mark W. & Wilson, David B. (2001). *Practical Meta-Analysis*. SAGE Publications, p. 73 - 75.

[30] Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston: Houghton-Mifflin.

[31] Lipsey, Mark W. & Wilson, David B. (2001). *Practical Meta-Analysis*. SAGE Publications.

[32] Lipsey, Mark W. & Wilson, David B. (2001). *Practical Meta-Analysis*. SAGE Publications, p. 91f.

[33] Rosenthal R. *Meta-analytic procedures for social research*, Sage Publications; CA, 1991.

[34] Hedges, Larry V. & Olkin, Ingram (1985). *Statistical Methods for Meta-Analysis*, Academic Press, p. 47f.

[35] Glass, Gene V., McGaw Barry, and Smith, Mary Lee (1981). Meta-analysis in social research, Sage Publications, p. 15.

[36] Lipsey, Mark W. & Wilson, David B. (2001). *Practical Meta-Analysis*. SAGE Publications, p. 71.

[37] Lipsey, Mark W. & Wilson, David B. (2001). *Practical Meta-Analysis*. SAGE Publications, p. 172f.

[38] Wilson, David B. (2001). *Meta-Analytic Methods for Criminology*. The Annals of the American Academy, 578, November 2001, 71 – 89.

[39] Pettigrew, Thomas F., and Tropp, Linda, R. (2000). Does intergroup contact reduce prejudice? Recent meta-analytic findings. In S. Oskamp (ed.), Reducing prejudice and discrimination (pp. 93 – 114). Mahwah, NJ: Erlbaum.

[40] Lösel, Friedrich & Beelmann, Andreas (2003). Effects of Child Skills Training in Preventing Antisocial Behavior: A Systematic Review of Randomized Evaluations, in: The Annals of the American Academy, 587, May 2003, p. 84 – 109.

[41] Wilson, David B. (2001). *Meta-Analytic Methods for Criminology*. The Annals of the American Academy, 578, November 2001, 71 – 89.

[42] Wilson, David B. (2001). *Meta-Analytic Methods for Criminology*. The Annals of the American Academy, 578, November 2001, 71 – 89.

[43] McKenzie, Doris Layton; Wilson, David B. & Kider, Suzanne B. (2001). *Effects of correctional boot camps on offending*. The Annals of the American Academy, 578, November 2001, p. 130.

[44] Schulze, Ralf (2004). *Meta-Analysis. A Comparison of Approaches*, Hogrefe & Huber, p. 19f.

[45] Hedges, Larry V. & Olkin, Ingram (1985). *Statistical Methods for Meta-Analysis*, Academic Press, p. 227.

[46] Hackney, Charles H. & Sanders, Glenn S. (2003). *Religiosity and Mental Health: A Meta-Analysis of Recent Studies*. Journal for the Scientific Study of Religion, 42:1, p. 43 – 55.

[47] Lipsey, Mark W. & Wilson, David B. (2001). *Practical Meta-Analysis*. SAGE Publications, p. 174, 200, 201, 202.

[48] Lipsey, Mark W. & Wilson, David B. (2001). *Practical Meta-Analysis*. SAGE Publications, p. 172, 173.

[49] Wilson, David B. (2001). *Meta-Analytic Methods for Criminology*. The Annals of the American Academy, 578, November 2001, 71 – 89.

[50] Wilson, David B. (2001). *Meta-Analytic Methods for Criminology*. The Annals of the American Academy, 578, November 2001, 71 – 89.

[51] Lipsey, Mark W. & Wilson, David B. (2001). *Practical Meta-Analysis*. SAGE Publications.

[52] Lipsey, Mark W. & Wilson, David B. (2001). *Practical Meta-Analysis*. SAGE Publications,.

[53] Veenhoven, Ruut (1996). *Average Level of Satisfaction in 10 European Countries. Explanation of Differences*. In: Saris, W.E.; Veenhoven, R.; Scherpenzeel, A.C. & Bunting, B. (eds). *A comparative study of satisfaction with life in Europe*. Eötvös University Press, pp. 243-253.

[54] McKenzie, Doris Layton; Wilson, David B. & Kider, Suzanne B. (2001). *Effects of correctional boot camps on offending.* The Annals of the American Academy, 578, November 2001, p. 130.

[55] Wilson, David B. (2001). *Meta-Analytic Methods for Criminology*. The Annals of the American Academy, 578, November 2001, 71 – 89.

[56] Taylor Halvorsen, Katherine (1994). The reporting format. In: Cooper, Harris, & Hedges, Larry V. (1994). *The Handbook of Research Synthesis*, Russell Sage Foundation, New York, p. 426f.

[57] Light, Richard J., Singer, Judith D., and Willett, John B. (1994). In: Cooper, Harris, & Hedges, Larry V. (1994). *The Handbook of Research Synthesis*, Russell Sage Foundation, New York, p. 439f.

[58] Wang, Morgan C. and Bushman Brad J., 1998, *Using the Normal Quantile Plot to Explore Meta-Analytic Data Sets*, in Psychological Methods, Vol. 3, No. 1, p. 46-54.

[59] Schwarzer, Guido; Antes, Gerd; and Schumacher, Martin (2003). Statistical Tests for the Detection of Bias in Meta-Analysis. In Schulze, Ralf; Holling, Heinz, and Böhning, Dankmar (2003), Meta-analysis. New Developments and Applications in Medical and Social Sciences, p. 73, 74.

[60] Sutton, Alex J., Abrams, Keith R., Jones, David R.; Sheldon, Trevor, A. and Song, Fujian (2000). Methods for Meta-Analysis in Medical Research, p. 113.

[61] Schwarzer, Guido; Antes, Gerd; and Schumacherr, Martin (2003). Statistical Tests fort he Detection of Bias in Meta-Analysis. In Schulze, Ralf; Holling, Heinz, and Böhning, Dankmar (2003), Meta-analysis. New Developments and Applications in Medical and Social Sciences, p. 73, 74.

[62] Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions 4.2.5 [updated May 2005]. http://www.cochrane.org/resources/handbook/hbook.htm (accessed 31st May 2005), Module 15.

[63] White, Howard D. (1994). Scientific Communication and literature retrieval. In: Cooper, Harris, & Hedges, Larry V. (1994). *The Handbook of Research Synthesis*, Russell Sage Foundation, New York, p. 42.

[64] Matt, Georg, E, and Cook, Thomas D. (1994). Threats to validity of research synthesis. In: Cooper, Harris, & Hedges, Larry V. (1994). *The Handbook of Research Synthesis*, Russell Sage Foundation, New York, p. 504f.

[65] Frye, Crissie Marie (2001). *The Effect of Emotional Stability on Job Satisfaction: A Meta-Analysis*, in: The Humanities and Social Sciences, 2001, 61, 11.

[66] Eddy, Erik R (2000). *The Importance of the Critical Psychological States in the Job Characteristics Model: A Meta-Analytic and Structural Equations Modeling Examination*, in: Current Research in Social Psychology, 2000, 5, 12, 22 May, 170-189.

[67] Viswesvaran, C., & Ones, D. S. (1995). Theory testing: Combining psychometric meta-analysis and structural equations modeling. Personnel Psychology, 48, p. 866.

[68] Lipsey, Mark W. & Wilson, David B. (2001). *Practical Meta-Analysis*. SAGE Publications, p. 10, 11.

[69] Smith, M. L., & Glass, G. V. (1977). *Meta-analysis of psychotherapy outcome studies*. American Psychologist, 32, 752 – 760.

[70] Cook DJ, Mulrow CD, Haynes RB. Systematic Review. Synthesis of Best Evidence for Clinical Decisions. Ann Intern Med 1997; 126(5): 376-380.

[71] Lipsey, Mark W. & Wilson, David B. (2001). *Practical Meta-Analysis*. SAGE Publications, p. 19.

[72] Doucouliagos, Chris (2005). Publication Bias in the Economic Freedom and economic growth literature. In: Journal of Economic Surveys, Vol. 19, No. 3, p. 367.

[73] Lipsey, Mark W. (2003), *Those Confounded Moderators in Meta-Analysis: Good, Bad and Ugly*, in: The Annals of the American Academy, 578, p. 69.

Eingereicht bei der Zeitschrift *Transportation*

3) Are 'Soft' Policy Measures Effective in Reducing Peoples' Car Use? A Meta-Analytical Review of Research Evidence

Guido Möser & Sebastian Bamberg

Over the last decades, car use has increased considerably all over the world. The rise in car use is associated with the problems of congestion and pollution that we are all so familiar with. Some of the pollution-related problems can be tackled by reductions in fuel consumption and 'cleaner' vehicle technology. However, other car use related problems cannot be solved by improvements in motoring technology. These include the threats to individual health (through road traffic casualties), the economy (through congestion and time lost), the environment (in terms of land use, noise and effects on wildlife etc) and our communities (severance and loss of community space). These problems can only be solved, if the total level of car use is reduced or at least a further increase in car use is stopped (e.g. Vlek & Steg, 1996; Steg & Tertoolen, 1999).

Whereas the problem diagnosis is clear, finding effective ways for reducing car use seems to be difficult. In the last decade, for this purpose, most local authorities have tried out 'hard' policy measures such as physical improvements to transport infrastructure or operations, traffic engineering, control of road space and changes in price. However, despite huge financial investments, these 'hard' infrastructural initiatives alone fail to deliver the shifts from car use that were hoped for and expected (e.g., Stopher 2004). Similar experiences have been made with transport pricing strategies. Due to the high political costs associated with pricing, politicians were and are very reluctant to practically apply them (e.g. Schade & Schlag, 2003).

Probably these sobering experiences are the background for transport policy's growing interest in a range of initiatives which are widely described as 'soft' measures. Typical soft measure examples are workplace travel plans, personalised travel planning, public transport marketing, and travel awareness campaigns. A consistent definition has not yet been developed to identify what constitutes a 'soft' measure. The word 'soft' is sometimes used to distinguish these initiatives from the above mentioned 'hard' measures, although soft measures often include such 'hard' elements. For example objective improvements of service quality are an important prerequisite for effective public transport marketing and parking fees and restrictions are main elements of effective workplace travel plans. 'Soft' also refers to another typical feature of these measures: they try to influence individual decision making by persuasion that is by changing peoples perceptions and motivations. For this purpose soft measures systematically apply principles of social marketing that is the application of marketing technologies developed in the commercial sector to the solution of social problems (e.g. Ampt, 2003). Psychological concepts like perceptions, values, attitudes, cultural and social norms or perceived self-efficacy play an important role within this approach.

Although some soft travel measures have a long history, their broader application dates to the mid 1990s. In the meantime they have been tested quite frequently in continental Europe, however, at the moment Australia and Britain are the pioneers in the systematic application of this strategy. Consequently most systematic attempts to review the effectiveness of soft travel measures have been undertaken by research teams in these two countries. For example, the British Department for Transport has commissioned a series of scientific reports (Dodgson, Sandbach, McKinnon, Shurmer, van Dijk, & Lane, 1997; Dodgson, Pacey, & Begg, 2000; Atkins, 1999; Halcrow Group, 2001, 2002; James, 2003; Sloman, 2003; Steer Davies Gleave, 2003; Cairns, Sloman, Newson, Anable, Kirkbride & Goodwin, 2004) reviewing national and international evidence in order to make estimates of the overall effect of a combination of soft measures on traffic levels in British conditions. In a recent review Cairns et al. (2004) develop a 'low intensity' and 'high intensity' impact scenario of future soft travel measures implementations based on a review of studies evaluating the effectiveness of these measures. In the low intensity scenario, they assumed that local authorities would carry on introducing these initiatives, so there would be gradual growth in the number of schemes, but no step-change. In the high intensity scenario, the researchers assumed that there would be much more activity and many more resources than at present. In the low intensity scenario, Cairns et al. (2004) estimate that peak hour urban traffic could be cut by about 5%. Nationally, car traffic could be cut by 2%-3%. In the high intensity scenario they estimate that peak hour urban traffic could be cut by 21 % and peak hour non-urban traffic by 14 %. Nationally, under the high intensity scenario, car traffic cut up to 11 % should be possible.

If these claims are real, then soft travel measure programs would be a very valuable approach with the potential to achieve a significant drop in car use. Consequently the British as well as Australian Department for Transport have decided to integrate soft travel measures as a vital part of their local transport strategy. Both governments have decided to invest substantial financial resources over the next 10 years, in motivating authorities to implement soft travel measure programs at the regional and local level (DfT, 2005; Pramberg, 2004).

However, there are scientists who see the danger that the benefits of soft travel measures are being oversold to policy makers (e.g., Stopher & Bullock, 2003). They criticise that the empirical input used for scenario development is of weak and often questionable nature: Most of the evaluation studies undertaken to assess behavioural effects of soft measures have not been as rigorously executed as is required for a credible estimation of the behavioural change effects of soft measures (Stopher & Bullock, 2003; Stopher, Alsnih, Bullock & Ampt, 2004; O'Fallon & Sullivan, 2003; Richardson, 2003)

One central methodological weakness of most available soft measure evaluation studies is their use of weak quasi-experimental designs, namely simple pre-post test designs. The fact that instead of panel data often cross sectional data is used for comparing pre/post intervention car use further reduces the possibility of drawing strong causal conclusions from these studies.

A second weakness of is that many of these evaluation studies have too low sample sizes for providing enough statistical power necessary for detecting true behavioural change effects. Using a nation-wide New Zealand travel behaviour survey for estimating the variability of travel behaviour, O'Fallon & Sullivan (2003) calculate the sample size necessary for having enough statistical power to detect a true 10 % decrease in car use, which is assumed as typical effect of a soft travel measures (e.g., Brög and John, 2001). The sample size required to detect such a behavioural change is 2252 persons in the case of a cross sectional pre-post test design and 727 persons in the case of a panel one-group pre-post test design. In contrast to this calculation the average sample size of the evaluation studies found in the literature is around 600 persons. Perhaps as a consequence of the often low statistical power quasi all available evaluation studies do not report any statistical significance testing of the found behavioural effects. The combination of weak designs and small sample sizes drastically increases the probability that the reported shifts in car use may reflect no systematic effect of the intervention but only random fluctuation.

Another weakness concerns the external validity of the evaluation results that is their 'generalisability' to the total population. In most evaluation studies, highly selected population segments are studied which differ from the total population not only in their socio-demographic background, their location and car availability but also in their attitudes toward as well as actual use of transportation means. Stopher and Bullock (2003) have tried to estimate the total population effect of soft measures corrected for the potential effects of this sample selectivity. They come to the conclusion that when applied to the total population a car reduction of no more than 3 % is more realistic than the 10 % reported in studies based on selected population segments.

The present study

As discussed in the last section, taken as single studies most of the available soft measure evaluations are of only weak methodological quality. On the other hand, together these studies represent a body of research in which a considerable amount of money has been invested and in which thousands of citizens have been involved. Simply throwing away this data would mean wasting a lot of time, money and engagement.

Another reason for working with this data is the chance they provide to demonstrate empirically the effectiveness of behaviour-oriented ‘soft’ interventions in the transport domain. As in most environmental policy fields, transport politicians first ask engineers for advice, then economists or lawyers. Eventually, when the decision concerning the principal problem solving strategy is made, social scientists / psychologists are asked to design a ‘public awareness campaign’ for convincing the people to do what the experts expected them to do.

As discussed above, in the transport domain there is growing insight, that such a one-sided technology oriented ‘hard’ approach obviously offers no effective strategies for changing preferences and behavioural trends. For environmental psychology, this opens the chance to get more influence on the strategic policy planning in this environmentally important domain. However, a fundamental prerequisite for this is the convincing empirical demonstration that behaviour oriented ‘soft’ travel measures are indeed effective in reducing car use.

Thus the central idea of the present paper is to use meta-analytical techniques for an aggregated analysis of this body of evaluation studies. Meta-analysis is a technique for combining effect size estimates from many primary studies to try to estimate the probably true effect size in the population. In simple terms, meta-analysis focuses on two general issues: central tendency and variability. Central tendency relates to the need to estimate the effect size in the population (i.e., the true effect size) and its significance. As such, effect sizes are calculated for individual studies, converted to a common metric, and then combined to obtain an average effect size. Studies in a meta-analysis are typically weighted by their importance, which is achieved by using the inverse of the variance associated with the effect size (itself a function of the sample size) as a weight. The mean effect size can then be expressed in terms of standard normal deviations (a z score) by dividing by the standard error of the mean. A significant value (i.e., the probability of obtaining a z score of such magnitude by chance) can then be computed, or significance can be inferred from the boundaries of a confidence interval constructed around the mean. The issue of variability relates to the similarity of effect sizes across studies and is generally addressed by testing the homogeneity of the single effect sizes.

Thus in the present context meta-analysis provides the possibility to tackle at least two of the deficits found in the existing soft measure evaluation literature: Meta-analysis allows to the calculation of the significance of the mean effect size estimated across all available primary evaluation studies. This allows at least at the population level to test the probability that the mean effect size found across all available primary evaluation studies represent only random fluctuation. The second advantage of a meta-analysis is that by pooling of the results of multiple primary studies the standard error of the estimated mean effect size can be reduced which increases statistical power. Thus at the aggregate level meta-analysis can provide enough statistical power for detecting 'true' effects even if they are relatively small (e.g. Lipsey & Wilson, 1993).

Method

Data collection

Our search in data-banks, journals and the internet quickly showed that the results of most soft travel measure interventions are not reported in publicly available sources. This unfortunate situation is caused by the fact that local authorities often commission commercial consultant firms (e.g., Socialdata, Germany, and Steer Davies Gleave, UK) with the development, implementation and evaluation of soft travel measures. Due to their commercial interests, these consultant firms are reluctant to provide the public with detailed information about the results of their work. Often only short ‘success’ brochures and presentations are available produced mainly for marketing and public relation purposes. Our attempt to receive more detailed information directly from the consultant firms was less successful: Either they did not react or they sent us the known brochures again.

Because of the difficulties in obtaining direct access to the original evaluation reports, we used another strategy to get the evaluation data necessary for our meta-analysis. As mentioned above, over the last years especially the British and Australian Departments for Transport have commissioned a series of research reviews on the effectiveness of soft travel measures. Obviously consultant firms were more ready to collaborate with these research teams probably because the commissioning governmental institutions are important potential clients. Due to the amount of time, labour and money invested in these successive research reviews they provide a rather comprehensive overview of the literature available on this topic. Thus we carefully checked all available research reviews for the reported evidence on the effectiveness of soft travel measures and used the reference sections for searching additional, publicly available documents. In the context of our meta-analysis the following reviews were particular valuable information sources: Cairns, Davies, Newson, & Swiderska (2002), Cairns, Sloman, Newson, Anable, Kirkbride & Goodwin (2004), Ker (2003), Steer Davies Gleave (2003) and GORS (2005).

Inclusion criteria

A problem one is confronted with when trying to meta-analyse the evidence on soft travel measures concerns the question how to classify these measures. For meta-analytical purposes such a classification system is critical because it enables one to define the population of studies investigating the same entity. The calculation of a common effect size only makes sense if we can assume that a collection of primary studies is dealing with the same entity that is in our case the behavioural effect of a specific soft measure intervention type. In the present paper we solved this problem practically by adopting the classification system of the Cairns et al (2004) review. It uses the following ten categories for classifying the found evidence on soft travel measures: (1) workplace travel plans, (2) school travel plans, (3) personalised travel planning, (4) travel awareness campaigns, (5) public transport information marketing, (6) car clubs, (7) car sharing schemes, (8) teleworking, (9) teleconferencing, and (10) home shopping. Our analysis of the available documents shows that the first five intervention types (workplace and school travel plans, personalised travel planning, travel awareness campaigns, and public transport information marketing) seem to be the most 'mature' intervention types, that is the interventions most often implemented and evaluated in the field. Thus we have decided to concentrate our meta-analysis on these intervention types. The next sections present working-definitions of these intervention types we used for classifying the found primary evaluation studies.

Workplace travel plans primarily aim to address the commuting habits of employees, although many also incorporate measures aimed at travel during the course of work, including business and delivery travel, and also travel by patients, students, shoppers, tourists, or other visitors to the employer's site. A workplace travel plan can be described as a package of measures put in place by an employer to encourage more sustainable travel, usually meaning less car use, particularly less single occupancy car use. Workplace travel plans often vary depending on the site and number of employees, but typically include packages of the following elements: New public bus or rail services linking to the site; dedicated 'work buses' shuttling between the site and the town centre; giving all staff public transport information; offering personalised journey plans to staff; interest-free season ticket loans; special deals to reduce the cost of bus and rail travel for employees; secure cycle parking; changing facilities, showers and lockers; business cycle mileage allowance; a car sharing scheme; preferential car parking for sharers; parking 'cash out' (paying employees a small sum on days they do not drive); car parking restricted to essential users; parking charges; publicity: newsletters, prize draws etc. linked to special car-free days; services on site to reduce need to travel (e.g. cafeteria, cash dispenser, convenience store); encouraging teleworking; and variations on the five-day week e.g. compressed working hours.

School travel plans aim to cut the congestion caused by the school run; reduce traffic danger; and support pupils who are already travelling by more sustainable means. It makes it more attractive for pupils to get to school by walking, cycling, public transport or sometimes car sharing. Typical school travel plan measures might include: special walking or cycling promotion days; walking buses or cycle trains; a programme of pedestrian and cycle training for children, including on-road tuition as well as in the playground; cycle parking; improvements to bus or train services; special school buses, with a seat for every child, on-board escorts, seatbelts, a smoking ban, and drivers trained in supervising children; fare cuts; car sharing schemes for families living in the same neighbourhood; activities as part of the curriculum to sell the benefits of sustainable transport and involve children in developing the plan; physical changes to the streets around the school, such as 20mph limits, traffic calming, pedestrian crossings and cycle lanes; and setting out the travel policy in the school prospectus and/or home-school agreement.

Personalised travel planning, travel awareness campaign, and public transport marketing are targeted marketing techniques, providing travel advice and information to people based on an understanding of their personal trip patterns. Employees, school children or households in a particular area might be contacted to find out which range of services and information and sometimes incentives they would find useful. The items on offer might include: pocket sized public transport timetables for the main routes into town; a timetable specifically for their nearest bus stop; a personalised journey plan for a trip they make on a regular basis; a free one-month public transport trial ticket for people who do not already use public transport; the offer of a visit from someone who can provide personal travel advice; a map of walking and cycle routes in their area; and loan of a bike. Participants are sometimes asked to keep a travel diary and may be given tips and suggestions for how to use their cars less. From our viewpoint there is no sharp dividing line between personalised travel planning, travel awareness campaign, and public transport marketing. As well as focusing on local environmental and health impacts, travel awareness campaigns also aim to improve informed knowledge of the facilities available for walking, cycling and public transport use. Where this information is expressed at a more general level it is usually described as a travel awareness campaign, and where it is aimed at specific local conditions and individual journeys it is closer to personalised travel planning. Similarly, there can be considerable overlap between travel awareness campaigns and public transport marketing. Thus in our meta-analysis we will treat these three interventions as one intervention type.

For our meta-analysed we compiled evaluation results for which the documents at least reported information about car use before and after an intervention as well as information about sample size. In this context it has to be mentioned that the three intervention types differ in how they typically operationalise car use reduction: Most evaluations of work and school travel plans use the proportion of employees / pupils arriving on a specific day by car at their workplace / school as central effectiveness measure. Studies evaluating the effects of personalised travel planning / travel awareness campaign / public transport marketing interventions typically use the proportion of trips conducted by car in relation to the total number of reported daily trips (so-called modal-split) as central effectiveness measure. An annoying aspect we were confronted with is that in many evaluation reports, especially those produced by private consultant firms, the information about study sample sizes is unclear or in part missing. Especially in the case of panel designs often only the before net sample sizes are reported, whereas information about response rates and panel mortality is missing.

In the documents, we have access to, we found a total of 141 evaluation reports fulfilling our inclusion criteria. Of these 141 studies 44 evaluated the behavioural effect of workplace travel plans, 25 the effect of school travel plans, and 72 studies evaluated the effect of personalised travel planning / travel awareness campaign / public transport marketing interventions. The tables presented in the Appendix report for all included studies the source as well as the other compiled information.

Meta-analytical procedure

Defining and calculating the effect size statistic. The key to meta-analysis is defining an effect size (ES) statistic capable of presenting the quantitative findings of a set of studies in a standardised form that permits meaningful numerical comparison and analysis across the studies. As reported above, all intervention studies included in the present meta-analysis use proportions as central effectiveness measure. To prevent a negative sign of the ES's, in the first step we have converted each reported car-use proportion into its corresponding no-car-use proportion (1 – car-use proportion, see Appendix A – C, columns 'no car before / after'). However, using the change in the before/after no-car-use proportions directly as ES statistic would have statistical disadvantages (see e.g., Lipsey & Wilson, 2001). Thus instead of the raw proportions statisticians recommend the use of arcsine-transformed proportions for calculating the ES's. The arcsine method is borrowed from statistical power analysis (Cohen, 1988) and creates a ES for the difference between proportions whose statistical power is independent of the location of a proportion between 0 and 1. Appendix A - C presents the arcsine-transformed before/after no-car-use proportions as well as the ES's resulting from subtracting the transformed before proportion from the transformed after proportion (so-called Cohen's h). In a second step for each ES the standard error term (SE = $1/n_{before} + 1/n_{after}$) was calculated. Its inverse ($w = 1/SE^2$) is used as weight with which each primary study contributes to the calculation of the common mean ES. We used n_{before} and n_{after} as conservative estimation of the sample size. For example, in Appendix A, the Estimated Variance was calculated by: 1/(*Study N Before * No car before/100*) + 1/(*Study N After * No car After/100*). Appendix A – C report for each study the variance of Cohen's h as well as the calculated study weight (w). As discussed above, one problem we accounted when calculating w is that 100 of the 141 studies provide only information about the before sample size, whereas the information on the after sample size is missing. In these cases the before sample size was also used as an estimator of the after sample size.

Fixed versus random-effects models. An important controversy surrounding the use of meta-analysis relates to the assumptions made about the population from which studies within the meta-analysis are taken. There are two ways to conceptualise this process: fixed-effects and random-effects models (e.g., Hedges, 1992; Hedges & Vevea, 1998; Hunter & Schmidt, 2000). In essence, in the fixed-effect conceptualisation, studies in the meta-analysis are assumed to be sampled from populations with a fixed-effect size. In other words, the effect size in the population is assumed to be the same for all studies included in the meta-analysis. This situation is known as the homogenous case. The alternative is to assume that the population effect sizes vary randomly from study to study. As such, a study included in a meta-analysis comes from a population that is likely to have a different effect size than any other study in the same meta-analysis.

Population effect sizes can, therefore, be thought of as being sampled from a universe of possible effects – a ‘superpopulation’ (Becker, 1996; Hedges, 1992). This situation is the heterogeneous case.

In statistical terms the two meta-analytical frameworks differ in the calculation of the weight used in the analysis, which in turn affects the standard errors associated with the mean effect size. Fixed-effects models use only within-study variability in their weights because all other ‘unkowns’ in the model are assumed to be constant (see Hedges, 1992; Hedges & Vevea, 1998). However, random-effects models account for the errors associated with sampling from populations that themselves have been sampled from a super-population. The error term, therefore, contains variability arising from differences between studies in addition to within-study variability (see Hedges & Vevea, 1998). Standard errors in the random-effects model are, therefore, larger than in the fixed case, which makes significance tests of combined effects more conservative.

In essence, the problem is whether fixed or random-effects methods are most appropriate for meta-analysing real-world data. Over the last years, there is growing evidence suggesting that the assumption of fixed population effect sizes is not tenable for virtually all real-word data (e.g. Hunter & Schmidt, 2000). The National Research Council (1992) noted that variable population parameters are more common than fixed, and that in virtually all study domains there are always some substantive moderator variables that will create variability in population parameters. Others have argued that methodological factors, such as measurement reliability, range variation, or dichotomisation of continuous variables, will also produce variation in population parameters (Hunter & Schmidt, 1990; Osburn & Callender, 1992). Despite the growing evidence that the assumption of fixed population effects is frequently untenable with real-world data, many researchers continue to routinely apply fixed-effects meta-analytic methods to their data, probably because of the conceptual and computational simplicity of these methods. In a critical appraisal of this practice, Hunter and Schmidt (2000) analyse the effects of applying fixed-effects methods to data for which population effects vary. Hunter and Schmidt come to the conclusion that in this case a huge inflation of the Type I error rate should be expected. Instead of the nominally assumed α of .05, Hunter and Schmidt predicted Type I error rates of 11% (for study sample sizes of 25) and 28% (for study sample sizes of 100). That is inadequately applying fixed-effects methods to data for which population effects vary drastically increases the danger of concluding that there is a genuine effect when in fact there is no effect in the population. Thus most experts in the field heavily recommended routinely using, besides fixed-effects models, random-effects models for checking the potential impact of heterogeneity on the meta-analytic results.

However, one has to be aware of the consequences associated with the use of a random-effects model as meta-analytical framework: By applying a random-effects model we assume that the effect sizes will vary anyway so that one can ask what the value is in seeking an average effect size or worrying about whether it is significant. For example, assume that for Germany evaluation studies of a specific soft travel measure indicate a significant positive effect, for the UK a zero effect and for the USA a negative effect. A meta-analysis across all these evaluation studies may result in an estimated common mean of zero. Readers of such a meta-analysis might conclude that there is no empirical evidence for the effectiveness of this measure. Of course, this conclusion would be wrong: The evaluation studies indicate that it works in Germany, has no effect in the UK, and has a negative effect in the USA. In this case, the issue of interest is not so much the overall effect of the measure, but its effect at the level of a specific country. Put another way, using random-effects model as meta-analytical framework means to concentrate on another function of a meta-analysis: The question of whether there are systematic reasons why effect sizes vary. These factors are known as moderator effects.

Results

Estimating common effect sizes within the fixed-effects approach. A fixed-effects model was used to calculate a weighted mean ES across all 141 found evaluation studies as well as for the studies evaluating the three intervention types separately. As can be seen from Table 1, the weighted mean ES across all 141 evaluation studies is .12. The z-test value of this point estimate is 26.53, which exceeds the critical value of 1.96 (α-level .05). Correspondingly, the 95% confidence interval around the mean ES ($0.11 < \mu < 0.13$) does not include zero.

Table 1 also presents the weighted mean ES estimated for the three intervention types separately. For the intervention type 'travel planning / awareness campaign / PT marketing' the estimated weighted mean ES is .10, which's z-value is statistically significant. For the intervention type 'work travel plan' the estimated weighted mean ES is .24, which's z-value is also statistically significant. However, for the intervention type 'school travel plan' the estimated weighted mean ES is -.01. The z-value of this mean ES is statistically not significant.

Table 1

Model:	Mean ES	-95% CI	+95% CI	SE	Z	DF	Q	V
All Studies (N = 141)								
Fixed Effects	0,121	0,112	0,130	0,0046	26,53***	140	1517,58***	
Random Effects	0,153	0,119	0,187	0,0174	8,75***			0,0296
Travel Planning / Travel awareness campaign / PT Marketing (N = 72)								
Fixed Effects	0,099	0,085	0,113	0,0071	14,01***	71	122,72***	0,0029
Random Effects	0,105	0,080	0,130	0,0127	8,25***			
Work Travel Plans (N = 44)								
Fixed Effects	0,237	0,221	0,252	0,0077	30,61***	43	758,36***	
Random Effects	0,244	0,174	0,314	0,0355	6,87***			0,0454
School Travel Plans (N = 25)								
Fixed Effects	-0,012	-0,030	0,009	0,0094	-1,23	24	205,55***	
Random Effects	0,079	0,019	0,140	0,0309	2,56**			0,0174

Note: ES = effect size; CI = Confidence Interval; Q = Homogeneity Measure; V = Random Effects Variance Component

Testing the homogeneity of the primary study effect sizes. As mentioned above, the calculation of a fixed common mean effect size is based on the assumption of homogeneous ES that is that all studies reflect the same ‘true’ population effect. One practical way to assess whether population effect sizes are likely to be fixed or variable is to use the Q statistic (Hedges & Olkin, 1985) for testing the homogeneity of the primary study effect sizes. If the value of the Q statistic is non significant then it can be argued that population effect sizes are likely to be homogenous (and hence fixed to some extent). Table 1 presents the respective Q statistics calculated across all 141 studies as well as for the three types of intervention studies separately. In all four cases the Q-value is significant, that is the effect sizes are characterised by a degree of heterogeneity which can not be explained by sampling error alone.

Estimating common effect sizes within the random-effects approach. As discussed above, when there is evidence for heterogeneous ES’s, the random-effects approach is a substantially better way to control the Type I error rate than the fixed-effects approach. For this reason we have recalculated the mean weighted ES’s within the random-effects approach. For this purpose, the additional random effects variance component V was estimated by a non-iterative method of moments (Lipsey & Wilson, 2001) and is presented in Table 1. V was then added to the prior estimated ES variance and the inverse of this term was used as study weights.

As can be seen from Table 1 across all 141 intervention studies the random-effects model results in a slightly higher weighted mean ES of .15. The z-test value of 8.75 again is statistically significant, however, much lower than the respective z-value calculated under the assumption of a fixed-effects model. Calculating random-effects models for the two intervention types ‘travel planning / awareness campaign / PT marketing’ and ‘work travel plans’ separately provide similar results: For both intervention types the estimated random-effects weighted mean ES’s are slightly higher, the respective z-values are still significant, however, much lower compared with the z-values calculated under the fixed-effects model. For the intervention type school travel plan the application of the random-effects model make a more substantive difference: For this intervention type adding V the study variance obviously results in a stronger weighting of primary studies reporting higher ES of school travel plans. This is reflected in a random-effects weighted mean ES of .08, which z-value is also statistically significant.

Exploring heterogeneity. As discussed above, main focus of the random-effects approach is exploring the sources of ES heterogeneity. There are different methodological ways to deal with this question. The easiest way consists in visually inspecting the ES distribution, e.g. via a box-and-whisker plot. Aim of this analysis is the identification of potential outliers that is single ES’s which expressed in standard deviation units lay considerably above or under the mean of the total ES distribution. The logic behind this analysis is that such extreme strong or weak ES are probably caused by random fluctuation and are thus not representative for the ‘true’ population ES of an intervention. Theoretically the application of an outlier analysis can be justified within the fixed-effects approach: Heterogeneity is seen as caused by random ‘noise’ in the data. Deleting this ‘noise’ should result in a homogeneous set of ES representing the same fixed ‘true’ population effect.

Another approach, which is closer to the logic of the random-effects approach, is to use study and effect size descriptors captured in the study coding protocol as potential moderators of the ES distribution. Potential moderators may be differences in the target population, the realised intervention, the target organisation or site, or the evaluation process. Mixed-effects weighted meta-regression (e.g., Hedges, 1992; Lipsey & Wilson, 2001) provides an adequate statistical tool for such moderator analyses. This statistical model assumes that the effects of study and effect size descriptors are systematic but that there is a remaining unmeasured random effect in the effect size distribution in addition to sampling error. That is, variability in the effect size distribution is attributed to systematic (modelled) between study differences, sampling error, and an additional random component.

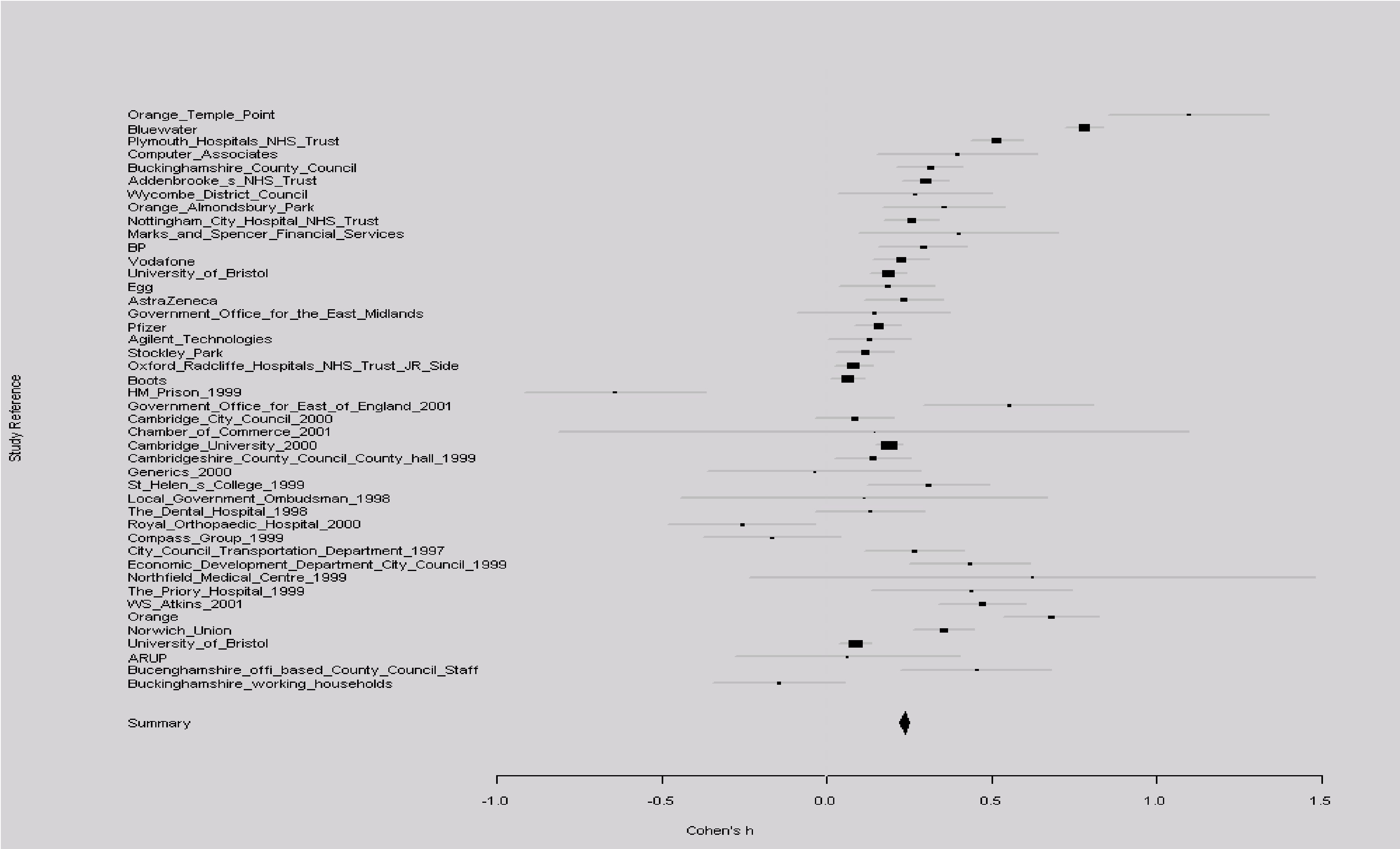
However, in the present paper the possibility of performing systematic moderator analyses is limited by the small set of available study and effect size descriptors (see Appendix). The different documents used for obtaining evaluation data consistently provide only information about the intervention type and the country in which the intervention was conducted. A part of the found study reports also provide information about the year when the intervention was conducted as well as the time period between before and after measurement. Thus in the present study we can use only this limited information for exploring the observed ES heterogeneity.

We started our heterogeneity analyses with the 72 studies evaluating the effect of travel planning / awareness campaign / PT marketing interventions. As first step a box-and-whisker plot of this ES distribution was inspected for potential outliers. Indeed, the box-and-whisker plot indicates two potential out-liners (Bike Busters, ES = 0.74, and Buckinghamshire Country Council, ES = 0.45) which's ES's lay six respectively three SD units over the mean of the ES distribution. Deleting these two outliers from the ES distribution results in a non significant value of the Q-statistic (Q-value = 41,28, df = 69, p = .99). The fixed-effects weighted mean ES for the remaining 42 evaluation studies results is .09 (z-test value = 12.30; 95% CI = 0.07 < μ < 0.10). Figure 1 shows the Forest plot for all 72 studies.

For the 44 studies evaluating the effects of work travel plans, a box-and-whisker plot also indicates two potential outliers (Orange Temple Point, ES = .78, and HM Prison, ES = -.64). Deleting these two potential out-liners from the ES distribution significantly reduces the value of the Q statistic (difference of Q-value = 87,67, df = 2, p < .001), however, the remaining Q-value is still significant. Calculating a random-effects model without the two out-liners results in a weighted mean ES of .24 (z-test value = 7.07; 95% CI = 0.18 < μ < 0.31). Obviously the heterogeneity of the ES reported in the studies evaluating the intervention work travel plan can not be explained by random noise alone. There must be additional systematic heterogeneity sources. Figure 2 shows the Forest-Plot of the 44 studies evaluating the effects of work travel plans.

Figure 1: Forest plot of the 72 studies evaluating the effect of travel planning/awareness campaign/PT marketing interventions

Figure 2: Forest plot of work travel plans (fixed effects model)



Thus in the second step we estimated a mixed-effects meta-regression model for the remaining 42 work travel plan intervention studies with data source, study year and before sample size as potential moderators. Data source is used as potential moderator because descriptive analysis showed that the average ES reported in one data source (Cairns et al., 2002) were slightly higher than the ES reported in the other sources (ES = .26 vs. ES = .19). Study year was used as potential moderator because one can argue that later implemented work travel plans may be more 'mature' intervention which take into account the experiences made in earlier interventions. Trichotomised before sample size (under 860, between 861 and 2100, and above 2100) is used as a proxy variable of organisational size. It can be argued that the greater the organisation the more resources are available for implementing drivers elements of the work travel plan package. As can be seen from Table 2, none of these three potential moderators are statistically significant. Thus they do not appear to add anything to explaining variability across the ES.

For the 25 studies evaluating the effects of school travel plans a box-and-whisker plot indicates an association between a specific data source and the reported ES's. The ES's taken from the Cairns et al. (2004) review all lay considerably over the mean calculated across all 25 studies. Figure 3 shows the Forest plot for the 25 studies

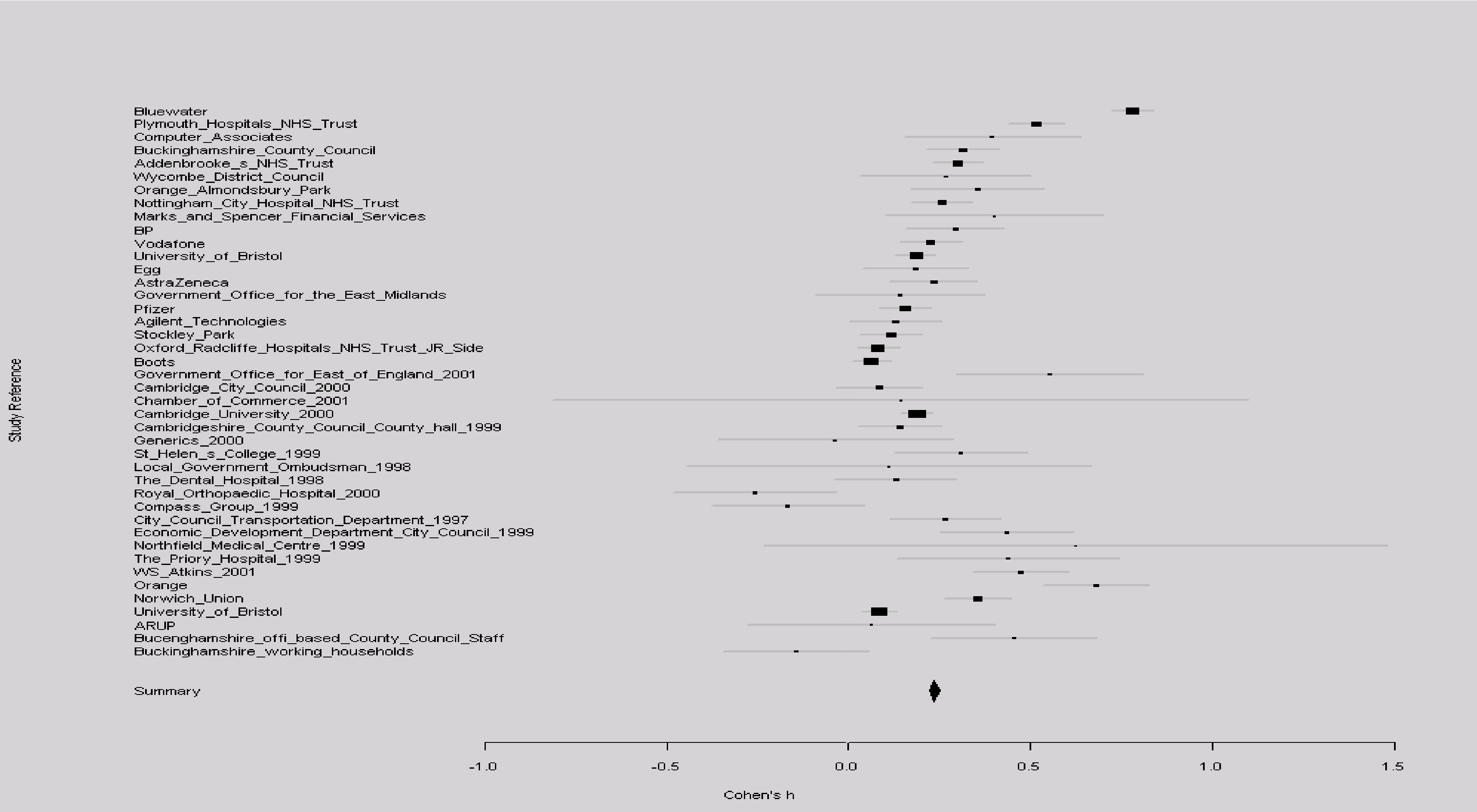
Table 2: Mixed-effects Meta-Regression Model for 42 workplace travel plan intervention studies with data source, study year and before sample size as potential moderators

Homogeneity Analysis	Q		df	P				
Model	0.81		3	.85				
Residual	25,52		34	.85				
Total	26,33		37	.90				
Random Variance Component V	0,051							
Study descriptor	B	SE	95% CI		Z	P	Beta	Model R^2
Constant	,1935	0,1957	-0.19	0.58	0,99			.03
Data source	,0815	0,1291	-0.17	0.33	0,63		0.20	
Size of organisation	-,0057	0,0511	-0.10	0.09	-0.11		-0.02	
Study year	,0007	0,0272	-0.05	0.05	0,03		0.01	

Notes:

B = unstandardized bivariate regression coefficient; CI = confidence interval; Z = value of z statistic; *** $p < .001$; ** $p< .01$; Beta = standardized regression coefficient; R^2 = percentage of heterogeneity explained by the descriptor set

Figure 3: Forest plot school travel plans (fixed effects model)



Whereas for the five ES's reported in the Cairns et al. (2004) review the fixed-effects weighted mean ES is .58 (z-test value = 10.10; 95% CI = 0.47 < μ < 0.69), the respective mean ES for the 20 studies found in the Steer Davies Gleave (2003b) and GORS (2005) reports is -.03 (z-test value = -2.94; 95% CI = -0.05 < μ < -0.01). Thus the ES's reported in the different documents are extremely heterogeneous: Whereas the ES's reported in the Cairns et al. (2004) review indicate a substantive positive car use reduction effect of the intervention school travel plan, the ES's found in the other two reports indicate a significant overall negative effect of this intervention type that is a decrease of the proportion of pupils not using the car after the intervention.

Table 3 presents the results of a mixed-effects meta-regression model with the 25 ES's as dependent and data source (Cairns et al. = 1; Steer Davies Gleave & DfT = 0), school size (number of pupils), and study year as independent variables (potential moderators). The regression results confirm the moderating impact of data source. The regression weight of these variables is significant (ß = .67, z-value = 5.72). Besides data source school size also seems to have a significant, however, negative moderating effect on the ES distribution (ß = -.28, z-value = 2.51): The greater a school, the weaker is the reported school travel plan ES's. The third potential moderator, study year, has no significant impact. Together data source and school size can explain 71% of the observed ES variance. However, the Q-statistic of the remaining residual ES variance is still significant (Q-value = 70.05; df = 22; p <.001). Obviously besides data source and school size there must be additional sources of the observed ES heterogeneity.

Table 3: Mixed-effects Meta-Regression Model for the 25 School Travel Plan Intervention Studies with Data Source, School Size, and Study Year as moderators

Homogeneity Analysis	Q	P	df					
Model	77.06	***	3					
Residual	31,13		21					
Total	108,18	***	24					
Random Variance Component V	0,006							
Study descriptor	B	SE	95% CI		Z	p	Beta	Model R^2
Constant	5,9275	31,4137	-55.64	67.50	0,19			.71
Data source	0.4807	0,0840	0.31	0.65	5,72	***	0.67	
School's pupil number	-0,0001	0,0000	-0.00	0.00	-2.51	*	-0.28	
Study year	-0,0029	0,0157	-0.03	0.03	-0,19		-0.02	

Notes: B = unstandardized bivariate regression coefficient; CI = confidence interval; Z = value of z statistic; *** $p < .001$; ** $p < .01$; * $p < .05$; Beta = standardized regression coefficient; R^2 = percentage of heterogeneity explained by the descriptor set

Discussion and conclusion

To prevent misunderstandings we want to stress that we do not claim that meta-analysis provides an easy way for ‘heeling’ fundamental methodological deficits of the momentarily existing soft policy measure evaluation literature. The most critical of these deficits is the dominance of weak quasi-experimental evaluation designs. Ultimately meta-analytical results can only be as reliable as the primary studies on which they are based. The inability of simple one-group pre-post-test designs to allow strong causal inferences severely limits the policy recommendations one can draw from our results.

Confronted with these limitations one may ask what the benefits of the present meta-analysis are. Would it not be better to wait with such an enterprise until a sufficient body of random control trials are available? Such a position may perhaps be adequate within an academic context however it is not very realistic in a policy making context. The reason for transport politicians’ interest in research syntheses is their pressure to develop effective car use reduction strategies. In such a context it is probably better to base decision making on weak empirical evidence than no evidence. From environmental research decision makers expect research syntheses as comprehensive, reliable and valid as possible. Adequate research syntheses are also important starting points for the preparation of future random control trials.

Before this background we see the main benefit of our meta-analysis in the fact that it provides a more transparent and reliable method of research synthesis than the traditional narrative literature review used by former research teams. These teams have done a good job in systematically searching and assembling existing evaluation evidence, however, the methods they use for synthesis this body of evidence appear to be quite unsystematic and intuitive: For example, in the Cairns et al. (2004) review at the end of each of the 12 chapters, suddenly a short paragraph appears in which the authors provide a numerical summary of the presented evidence. Typically this numerical summary consists of an estimated minimum-maximum percentage range of car use reduction expected from a specific intervention. The authors give no further information how they weight and combine the prior presented evidence information to get these estimates. Furthermore they neither try to test the probability that these estimates may only reflect random fluctuation nor do they analyse the homogeneity of their estimates. To summarise, on an aggregated level, this narrative research synthesis replicates the methodological deficits found on the level of the primary evaluation studies. In contrast to the less transparent and thus difficult to replicate narrative research synthesis the present meta-analysis reports in detail the methods used for research synthesis. Thus everybody who is interested in should be able to replicate our analyses.

But what substantive conclusions can be drawn from our meta-analysis? On the most aggregated level the positive message sounds that it provides empirical support for the claim that soft travel measures are an effective strategy for reducing car use: Across all 141 soft travel measure evaluation reports we found a statistically significant random-effects mean ES of .15. The confidence intervals indicate a 95% probability that the population mean is between .12 and .19. What is the practical meaning of this mean ES? Compared with the convention established by Cohen (1988) it indicates a small effect. However, in the transport policy context, even such a small effect is of considerable significance. As discussed above, the reduction, or even more modest the stop of further car use increase seems to be a goal very difficult to obtain. Most transport experts agree in the conclusion that the expensive 'hard' infra-structural measures alone have failed to reach this goal. Thus empirical evidence that a combination of 'hard' and 'soft' measure may be a more effective strategy has important implications for future transport policy.

To get a better impression what an ES of .15 practically means we have used the weighted original percentage proportions to calculate the average shift in the no-car proportion observed across all 141 studies. Expressed in this more common metric, the estimated mean ES of .15 reflects a average increase of the non-car-use proportion from 35.3 % to 42.1 % or a average 16 % reduction of car use. This is considerably higher than the 10 % car use reduction discussed in the literature as typical benchmark for the effect of soft policy measures. However, we want to stress again that because these results are based on weak evaluation designs, it remains unclear how much of this change should be attributed to the causal effect of soft policy measures.

The second important result of our meta-analysis concerns the heterogeneity of the ES's between as well as within the three intervention types. With a random-effects mean ES of .24 workplace travel plans seem to produce the strongest average car use reduction effect. For the combined intervention type travel planning / awareness campaign / PT marketing the respective mean ES is .11 and for school travel plans .08. However, when judging these ES's one has to remember the different goals of the three intervention types. Whereas workplace as well as school travel plans are targeting specific population segments and trips, the aim of travel planning / awareness campaign / PT marketing is to increase for the total population and all daily trips the proportion of no-car trips. When judging the possible environmental impact of soft travel policy measures, besides effectiveness it also has to be taken into account the potential coverage degree of the measures. Due to its greater coverage on a global level the total amount of car reduction reached by the intervention type travel planning / awareness campaign / PT marketing may be higher than that of school or workplace travel plans. A third factor necessary to take into account when judging the impact of the different intervention types are the implementation costs. In case studies Cairns et al. (2004) found for workplace travel plans average costs of 0.1 – 2.0 pence per reduced car km, for school travel plans this range is 1.4 – 9.9 pence and for travel planning / awareness campaign / PT marketing 0.2 – 4.4 pence.

Before the analyses especially for the intervention type travel planning / awareness campaign / PT marketing we expected to find ES heterogeneity. The relatively high number of evaluations available for this intervention type increases the probability of finding a significant Q-value. However, after deleting two outliers from the ES-distribution, our analyses indicate, for this intervention type a homogeneous ES-distribution. Thus for this intervention type the assumption of a fixed population effect is empirically not rejected. One possible reason for this finding may be that the main share of travel planning / awareness campaign / PT marketing interventions was conducted by two consultant firms: Social Data and Steer Davies Gleave. A high degree of professionalism and standardisation of both firms in the implementation of this intervention type provides a possible positive explanation of this result. An alternative, negative explanation would be that the firms have a commercial interest in reporting publicly only very similar findings.

For the other two intervention types our analyses indicate strong heterogeneity of the ES distributions. Instead of a fixed common population effect, the results support for these two intervention types more the assumption that the random-effects model is adequate. Obviously, differences in characteristics of the implemented interventions, the target groups, organisations and sites create sub-populations with varying mean ES's. This reduces the practical significance of the estimated intervention type specific mean ES's and underlines the importance of systematic moderator analyses. Unfortunately in the present study the little information available about potential moderators limits a more detailed exploration of heterogeneity.

For school travel plans our heterogeneity analysis indicates a moderating effect of the source from which we obtained the analysed primary ES's: Whereas the five ES's found in the Cairns et al. (2004) review are quite substantive, the ES's obtained from two other sources indicate a zero mean effect of school travel plans. Critically reading the three documents provides some explanations for this finding. The Cairns et al. (2004) review reports the results of five selected case studies. Our impression is that these case studies represent 'best practice' school travel plan examples. Besides infra-structural elements, these case study interventions include many of the awareness and behaviour change oriented elements mentioned in the working definition of this intervention type. Furthermore, the interventions were not only developed in participation and with support of school boards and parents, they also were implemented in communities with a longer experience in implementing this intervention type. In contrast to the Cairns et al. (2004) review, the second source (Steer Davies Gleave; 2003b) reports the results of a study evaluating a specific infra-structural element of school travel plans, a new (the ‘yellow’) school bus schema. This infra-structural element was not supported by awareness and behaviour change oriented elements. The authors of the third data source (GORS, 2005) themselves express doubt concerning the validity of the found evaluation results. The report’s aim was to evaluate the effects of travel plans in schools receiving a governmental grant for the implementation of this measure. However, the authors are skeptical whether this was a valid criterion for identifying newly implemented intervention. Some schools seem to receive a grant, which had carried out travel planning work for many years, whereas other schools have not started their travel plan even if they have already received a grant.

To summarise, our results indicate that moderator analysis, that is the theory-driven search for factors causing the variability of soft travel measures will be an important task of future meta-analysis in this field. Identifying substantive moderators also allows to specify regression equations which can be used for estimating subgroup specific ES’s. Compared with the mean ES within a random-effects approach such a procedure provides a more adequate way of predicting the potential impact of an intervention in a specific context. However, one precondition for conducting better future meta-analyses in general and moderator analyses in particluar is a drastic change of the publication practice in this field. At the moment the difficulty or inability to get direct access to original evaluation results is one main obstacle for performing meta-analysis. A second main obstacle is the less professional reporting of evaluation results in many documents. An especially annoying aspect is often the absence of precise and detailed information about research design and sample size. Furthermore, often in the documents also basic descriptive statistics of evaluation results like means and standard deviation are missing. Because most soft travel measures are implemented by local authorities using public money, it would be relatively easy to develop a guideline prescribing the commissioned firms or universities how to report their evaluation results. Such a guideline should also include a list of study and intervention descriptors which should also be reported. Ideally a central governmental institution should collect all local, regional and national intervention studies and should make them publicly available at best via an internet site.

Reference

References marked with an asterisk indicate studies included in the meta-analysis.

Ampt, L. (2003, August). *Voluntary Household Travel Behaviour Change – Theory and Practice.* Paper presented at the 10th International Conference on Travel Behaviour Research, Lucerne.

Armitage, P., & Berry, G. (1987). *Statistical methods in medical research* (2nd ed.). Oxford, England: Blackwell.

Atkins, W. S.(1999). *Assessing the effect of transport white paper policies on national traffic.* (Final Report and Appendices, Department for Environment, Food and Rural Affairs (UK) (DETR)). London.

Bregg, C. B. (1994). Publication bias. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 399-409). New York: Russell Sage Foundation.

Brög, W., & John, G. (2001, August). *Personalised Marketing – the Perth success story.* Proceedings: Marketing Public Transport Conference, 3. August 2001, Auckland, New Zealand. Retrieved December 12, 2005, from www.dpi.wa.gov.au/travelsmart/techpub.html.

Button, K., & Kerr, J. (1996, August). *Synthesising the results of quantitative case studies in transport analysis.* Paper presented at 36th Congress of the European Regional Science Association, 26-30. August, ETH Zürich, Switzerland.

* Cairns, S., Davies, A., Newson, C. & Swiderska, C. (2002). *Making travel plans work: Research report.* ((former) Department for Transport, Local Government and the Regions (DTLR)). London.

* Cairns, S., Sloman, L., Newson, C., Anable, J., Kirkbride, A. & Goodwin, P (2004). *'Smarter Choices - Changing the Way We Travel'.* (Final report of the research project: The influence of soft factor interventions on travel demand. Research report for the Department for Transport). London. Retrieved December 1, 2005, from http://www.dft.gov.uk/stellent/groups/dft_susttravel/documents/page/dft_susttravel_029722.pdf.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Dodgson, J., Pacey, J., & Begg, M. (2000). *Motors and Modems Revisited* (National Economic Research Associates (NERA)). London.

Dodgson, J., Sandbach, J., McKinnon, A., Shurmer, M., van Dijk, T., & Lane, B. (1997). *Motors or Modems* (National Economic Research Associates (NERA)). London.

* GORS (2005). *Travelling to School Initiative: Annexes to the Report on the Findings of the Initial Evaluation.* (Operational Research Unit for Sustainable Travel Initiatives Branch (GORS)).

Halcrow Group. (2002). *Multi-modal studies: soft factors likely to affect travel demand* (update final report, Department for transport). London.

Hedges, L. V. (1992). Meta-analysis. *Journal of Educational Statistics, 17,* 279-296.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3,* 486-504.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative knowledge in psychology. *International Journal of Selection and Assessment, 8,* 275-293.

James, A. (2002). *Review of Halcrow soft factors report* (Unpublished report for South West Transport Activists Roundtable).

* Ker, I. (2003). *Travel Demand Management: Public Transport Business Case,* (Contract Report RC5051 for Dept. of Infrastructure). Victoria, June.

Light, R. J., Singer, J. D., & Willett, J. B. (1994). The visual presentation and interpretation of meta-analysis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 439-453). New York: Russell Sage Foundation.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatments: Confirmation from meta-analysis. *American Psychologist, 48,* 1181-1209.

Lipsey, M. W., & Wilson, D. (2001). *Practical Meta-Analysis.* Thousand Oaks, CA: Sage.

O'Fallon, C., & Sullivan, C. (2003, October). *Personalised Marketing – Improving evaluation.* Paper presented a the 26th Australasian Transport Research Form, Wellington, New Zealand.

Richardson, A. J. (2003). *Temporal variability of car usage as an imput to the design of before & after surveys.* Paper presented at the 82nd Annual Meeting of the Transportation Research Board, Washington, D.C.

Rosenthal, R. (1979). The file-drawer problem and tolerance of null results. *Psychological Bulletin, 86,* 638-641,

Rosenthal, R.(1991). Meta-analytic procedures for social research. *Applied social research methods Series (Vol. 6).* Thousand Oak, CA: Sage.

Schade, J., & Schlag, B. (Eds.) (2003). *Acceptability of transport pricing strategies.* Amsterdam: Elsevier.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton-Mifflin.

Sloman, L. (2003). *Less traffic where people live: how local transport schemes can help cut traffic* (Transport for Quality of Life). Machynlleth.

Steer Davies Gleave (2003). *Weymouth relief road: alternatives to the proposed scheme.* (SDG). London.

* Steer Davies Gleave (2003b). *Evaluation of first Yellow Bus.* (Report for the Department for Transport). UK: London.

Steg, L., & Tertoolen, G. (1999). Sustainable Transport Policy: The Contribution from Behavioural Scientists. *Public Money & Managemant, 1,* 63-69.

Stopher, P. & Bullock, P. (2003, October). *Travel behaviour modification: a critical appraisal.* Paper presented to the 26th Australasian Transportation Research Forum, Wellington New Zealand.

Stopher, P. R. (2004). Reducing Road Congestion: A Reality Check. *Transport Policy*, *11,* 117-131.

Stopher, P., Alsnih, R., Bullock, P., & Ampt, L. (2004). *Evaluating Voluntary Travel Behaviour Interventions* (Working Paper ITS-WP-04-17, Institute of Transport Studies). Sydney: University of Sydney.

UK Department for Transport (2005). *Making Smarter Choices Work* (DfT Publication). London.

Vlek, C., & Steg, L. (1996, March). Societal reasons, conditions and policy strategies for reducing the use of motor vehicles: a behavioural-science perspective and some empirical data in OECD. In OECD, *Towards Sustainable Transportation*, *Proceedings of the International Conference Toward Sustainable Transport (pp. 10-16),* Vancouver.

Appendix A: Primary Studies Evaluating the Effect of the Intervention Type 'Travel Planning / Awareness Campaign / PT Marketing'(N = 72)

Intervention Study	Study Start	Study End	Data Source	Country	Study N	Study N Before	Study N After	No car Before	No car After	Arcsin Before	Arcsin After	Effect size Cohens h	Var. H	Weight W
Armandale	03.2003	03.2004	3	3		247	210	44	49	1,451	1,551	0,100	0,019	52,85
Cambridge	11. 2001	11. 2002	3	3		529	400	39,5	43	1,359	1,430	0,071	0,011	94,34
Frementle	03. 2001	03. 2004	3	3		476	615	48	53	1,531	1,631	0,100	0,007	134,3
Melville	03. 2000	11.2003	3	3		972	634	34	40	1,245	1,369	0,124	0,007	143,49
Subiaco	1998	09.2003	3	3			400	44	51	1,451	1,591	0,140	0,011	94,48
Marangaroo	05. 2002	05. 2003	3	3		300	201	40	42	1,369	1,410	0,041	0,020	49,56
South Perth	1999	10. 2000	3	3			706	40	48	1,369	1,531	0,161	0,006	154,04
Vincent	04. 2000	03.2004	3	3		416	409	44	52	1,451	1,611	0,160	0,010	98,38
Gloucester pilot	10. 2001		3	1	187			51	60	1,591	1,772	0,181	0,019	51,55
Gloucester largescale	07. 2003		3	1	2018			46	55	1,491	1,671	0,180	0,021	46,84
Bristol VIVALDI phase 1	09. 2002		3	1	232			52	57	1,611	1,711	0,100	0,020	50,85
Bristol Bishopston	05. 2003		3	1	5364			56	66	1,691	1,897	0,205	0,018	56,65
Frome pilot			3	1	282			53	59	1,631	1,752	0,121	0,019	52,21
Kingston TfL pilot			3	1	793			52	63	1,611	1,834	0,223	0,019	53,27
Southwark TfL pilot			3	1	257			50	66	1,571	1,897	0,326	0,019	53,20
Enfield TfL pilot			3	1	235			74	37	2,071	1,308	-0,764	0,022	46,13
Stepchange pilot in Scotland			3	1	1754			76	25	2,118	1,047	-1,070	0,028	35,18
East Hampshire 2003		2003	4	1	1000	1115	956	13	15	0,738	0,795	0,058	0,014	72,09
York (12 hour day)	2000	2002	4	1	500			53,8	56,3	1,647	1,697	0,050	0,007	137,55
York (morning peak)	2000	2002	4	1	500			57,4	63	1,719	1,834	0,114	0,007	150,17
Bike Bus'ters	1995	1996	4	1	175			20	55	0,927	1,671	0,744	0,014	73,33
Bishopsworth/ Hartcliffe project I	09.2002	10.2002	5	1	2500			53	56	1,631	1,691	0,060	0,007	136,15
Bishopston Indi. Marketing campaign	04.2003	06.2003	5	1	5364			62	66	1,813	1,897	0,083	0,006	159,84
Buckinghamshire County Council	08.1998	06.2003	5	1	3000			28,7	50,6	1,131	1,583	0,452	0,011	91,56

Appendix A, Continuation: Primary Studies Evaluating the Effect of the Intervention Type 'Travel Planning / Awareness Campaign / PT Marketing'(N = 72)

Intervention Study	Study Start	Study End	Data Source	Country	Study N	No car Before	No car After	Arcsin Before	Arcsin After	Effect size Cohens h	Var. H	Weight W
Quedgeley pilot	2001	2002	6	1	177	55	58	1,671	1,731	0,061	0,007	141,15
Quedgeley large scale 2004			6	1	954	75	80	2,094	2,214	0,120	0,005	193,55
South Perth (Western Australia)			7	3	1000	6	7	0,495	0,536	0,041	0,062	16,15
Brisbane (Queensland)			7	3	1100	6	10	0,495	0,644	0,149	0,053	18,75
Gloucester (UK)			7	1	445	3	5	0,348	0,451	0,103	0,107	9,38
Frome (UK)			7	1	500	5	6	0,451	0,495	0,044	0,073	13,64
Dulwich (South Australia)			7	3	515	3,8	3,6	0,392	0,382	-0,011	0,108	9,24
Christies Beach (South Australia)			7	3	215	2,9	3,7	0,342	0,387	0,045	0,123	8,13
Holland Park (South Australia)			7	3	102	9,3	8,6	0,620	0,595	-0,025	0,045	22,34
Lisbon (Portugal)			7	4	548	18	20	0,876	0,927	0,051	0,021	47,37
Copenhagen A			7	6	585	21	23	0,952	1,000	0,049	0,018	54,89
Helsinki (Finland)			7	7	176	37	42	1,308	1,410	0,102	0,010	98,35
Leipzig (Germany)			7	2	188	14	18	0,767	0,876	0,109	0,025	39,38
Magdeburg (Germany)			7	2	212	14	18	0,767	0,876	0,109	0,025	39,38
Halle (Germany)			7	2	154	19	23	0,902	1,000	0,098	0,019	52,02
Bremen (Germany)			7	2	189	17	20	0,850	0,927	0,077	0,022	45,95
Pinneberg (Germany)			7	2	501	15	19	0,795	0,902	0,107	0,024	41,91
Ludwigshafen (Germany)			7	2	197	9	13	0,609	0,738	0,128	0,038	26,59
Cologne (Germany)			7	2	235	19	21	0,902	0,952	0,050	0,020	49,88
Munich (Germany)			7	2	229	24	27	1,024	1,093	0,069	0,016	63,53
Borken (Germany)			7	2	410	4	6	0,403	0,495	0,092	0,083	12,00
Delft/Den Haag (Netherlands)			7	8	124	4	6	0,403	0,495	0,092	0,083	12,00
Liverpool (UK)			7	1	32	20	20	0,927	0,927	0,000	0,020	50,00
Hampshire (UK)			7	1	162	4	9	0,403	0,609	0,207	0,072	13,85

Appendix A, Continuation: Primary Studies Evaluating the Effect of the Intervention Type 'Travel Planning / Awareness Campaign / PT Marketing'(N = 72)

Intervention Study	Data Source	Country	Study N	No car Before	No car After	Arcsin Before	Arcsin After	Effect size Cohens h	Var. H	Weight W
Bern A (Switz)	7	5	247	31	33	1,181	1,224	0,043	0,013	79,92
Bern B (Switz)	7	5	27	18	23	0,876	1,000	0,124	0,020	50,49
Montpelier (France)	7	9	411	3	5	0,348	0,451	0,103	0,107	9,38
Bologna (Italy)	7	10	681	26	34	1,070	1,245	0,175	0,014	73,67
Parma (Italy)	7	10	721	13	24	0,738	1,024	0,286	0,024	42,16
Reggio Emilia (Italy)	7	10	691	17	17	0,850	0,850	0,000	0,024	42,50
Turin (Italy)	7	10	213	34	47	1,245	1,511	0,266	0,010	98,64
Venice (Italy)	7	10	742	37	45	1,308	1,471	0,163	0,010	101,52
Madrid (Spain)	7	11	382	7	13	0,536	0,738	0,202	0,044	22,75
Porto (Portugal)	7	4	421	22	22	0,976	0,976	0,000	0,018	55,00
Lison (Portugal)	7	4	548	23	25	1,001	1,047	0,047	0,017	59,90
Oslo (Norway)	7	12	1153	29	31	1,137	1,181	0,044	0,013	74,92
Arnhem (Netherlands)	7	8	106	2	3	0,284	0,348	0,064	0,167	6,00
Liverpool (UK)	7	1	33	12	16	0,707	0,823	0,116	0,029	34,29
Luxembourg	7	13	230	38	39	1,328	1,349	0,021	0,010	96,23
Nürnberg	8	2	4940	17	23	0,850	1,000	0,150	0,020	48,88
Wiesbaden	8	2	4632	17	19	0,850	0,902	0,053	0,022	44,86
Hannover-Südstadt	8	2	40990	25	30	1,047	1,159	0,113	0,015	68,18
Baunatal	8	2	6918	7	13	0,536	0,738	0,202	0,044	22,75
Kassel	8	2	13012	20	20	0,927	0,927	0,000	0,020	50,00
Vollmar	8	2	5655	4	8	0,403	0,574	0,171	0,075	13,33
Stuttgart-Freiberg	8	2	5330	21	22	0,952	0,976	0,024	0,019	53,72
Linz	8	14	15141	19	21	0,902	0,952	0,050	0,020	49,88
Salzburg	8	14	5500	16	19	0,823	0,902	0,079	0,023	43,43

Appendix B: Primary Studies Evaluating the Effect of the Intervention ‚Work Travel Plan', (N = 44)

Case Study	Study Start	Study End	Data Source	Country	Study N	No car Before	No car After	Arcsin Before	Arcsin After	ES Cohens h	Var. h	Weight w
Orange (Temple Point)	Oct 01	Oct 01	9	1	400	21	73	0,952	2,049	1,097	0,015	65,23
Bluewater	Mar 99	May 00	9	1	5500	31	69	1,181	1,961	0,780	0,001	1176,45
Plymouth Hospitals NHS Trust	1995	Oct 01	9	1	4193	22	46	0,976	1,491	0,514	0,002	624,02
Computer Associates	Jun 00	Oct 01	9	1	850	11	26	0,676	1,070	0,394	0,015	65,70
Buckinghamshire County Council	Sep 98	Feb 01	9	1	2200	29	44	1,137	1,451	0,313	0,003	384,55
Addenbrooke's NHS Trust	Oct 93	Oct 99	9	1	4977	26	40	1,070	1,369	0,299	0,001	784,25
Wycombe District Council	Mar 98	Mar 99	9	1	502	23	35	1,000	1,266	0,266	0,014	69,67
Orange (Almondsbury Park)	Jul 96	Oct 01	9	1	2000	08	20	0,574	0,927	0,354	0,009	114,29
Nottingham City Hospital NHS Trust	Nov 97	Nov 00	9	1	3500	27	39	1,093	1,349	0,256	0,002	558,41
Marks and Spencer Financial Services	Oct 98	Dec 99	9	1	1100	05	17	0,451	0,850	0,399	0,024	42,50
BP	Feb 98	Mar 01	9	1	2100	16	28	0,823	1,116	0,292	0,005	213,82
Vodafone	Jun 98	Oct 01	9	1	5400	16	25	0,823	1,047	0,224	0,002	526,83
University of Bristol	Nov 98	Nov 01	9	1	4177	56	65	1,691	1,875	0,184	0,001	1256,55
Egg	Sep 99	Jan 01	9	1	880	38	47	1,328	1,511	0,182	0,005	184,90
AstraZeneca	Oct 97	Oct 01	9	1	4200	10	18	0,644	0,876	0,233	0,004	270,00
Government Office for the East Midlands	Dec 97	Dec 99	9	1	245	55	62	1,671	1,813	0,142	0,014	71,41
Pfizer	Apr 98	Apr 01	9	1	5500	25	32	1,047	1,203	0,155	0,001	771,93
Agilent Technologies	Nov 97	Nov 99	9	1	1500	29	35	1,137	1,266	0,129	0,004	237,89
Stockley Park	Oct 97	Oct 99	9	1	7700	12	16	0,707	,8230	0,116	0,002	528,00
Oxford Radcliffe Hospitals NHS Trust	Mar 00	Mar 01	9	1	5170	42	46	1,410	1,491	0,081	0,001	1135,05
Boots	Jun 95	Nov 99	9	1	7500	35	38	1,266	1,328	0,062	0,001	1366,44
HM Prison	1999	2001	10	1	650	36	10	1,287	0,644	0,544	0,020	50,87

Appendix B, Continuation: Primary Studies Evaluating the Effect of the Intervention ‚Work Travel Plan', (N = 44)

Case Study	Study Start	Study End	Data Source	Country	Study N	No car Before	No car After	Arcsin Before	Arcsin After	ES Cohens h	Var. h	Weight w
Government Office for East of England	2001	2002	10	1	290	30,5	57,5	1,170	1,721	1,146	0,008	129,54
Cambridge City Council	2000	2002	10	1	800	65,3	69,2	1,882	1,965	1,273	0,005	218,38
Chamber of Commerce	2001	2002	10	1	18	43,4	50,5	1,438	1,581	1,076	0,007	151,72
Cambridge University	2000	2002	10	1	6250	64,3	73	1,861	2,049	1,319	0,005	222,22
Cambridgeshire County Council	1999	2002	10	1	1100	49	56	1,551	1,691	1,131	0,006	169,87
Generics	2000	2002	10	1	220	34,3	32,5	1,251	1,213	0,888	0,009	108,47
St Helen's College	1999	2002	10	1	800	23	37	1,000	1,308	0,938	0,011	92,19
Local Government Ombudsman	1998	2002	10	1	85	27	32	1,093	1,203	0,883	0,011	95,19
The Dental Hospital	1998	2001	11	1	400	66	72	1,897	2,026	1,306	0,004	223,83
Royal Orthopaedic Hospital	2000	2002	11	1	500	38	26	1,328	1,070	0,810	0,010	100,34
Compass Group	1999	2003	11	1	400	39	31	1,349	1,181	0,871	0,009	112,26
City council Transportation Department	1997	2001	11	1	562	52	65	1,611	1,875	1,225	0,005	187,78
Economic Develop. Depart. City council	1999	2003	11	1	423	50	71	1,571	2,004	1,294	0,005	190,70
Northfield Medical Centre	1999	2001	11	1	50	14	41	,7670	1,390	0,980	0,015	67,84
The Priory Hospital	1998	2001	11	1	300	21	41	,9521	1,390	0,980	0,011	90,27
WS Atkins	2001	2003	11	1	783	47	70	1,511	1,982	1,282	0,005	182,78
Orange			11	1	700	45	73	1,471	2,049	1,319	0,006	180,95
Norwich Union			11	1	1300	63	79	1,834	2,190	1,400	0,004	227,82
University of Bristol			11	1	5000	64	68	1,855	1,939	1,259	0,005	214,30
ARUP			11	1	109	59	62	1,752	1,813	1,193	0,005	196,50
?Buckinghamshire?	1998	2003	12	1	410	28,1	50	1,117	1,571	1,071	0,009	116,93
Buckinghamshire	2000	2003	12	1	850	26,2	20,1	1,075	0,930	,7288	0,014	73,932

Appendix C: Primary Studies Evaluating the Effect of the Intervention ‚School Travel Plan‘, (N = 25)

Case Study	Study Start	Study End	Data Source	Country	Study N Before	Study N After	No car Before	No car After	Arcsin Before	Arcsin After	Effect size Cohens h	Var. h	Weight w
Wrexham secondary		2003	1	1	2615	2463	81	77	2,240	2,141	-0,098	0,004	264,69
Runnymede secondary		2003	1	1	1459	1349	62	62	1,813	1,813	0,000	0,004	266,35
Wrexham primary		2003	1	1	80	47	55	66	1,671	1,897	0,226	0,090	11,07
Hebden Bridge Primary		2003	1	1	179	146	60	64	1,772	1,855	0,082	0,033	30,31
Wrexham post 16		2003	1	1	161	70	79	81	2,190	2,240	0,050	0,105	9,55
Bedfordshire - Lower School	2003	2004	2	1	3679	3640	59,8	57	1,768	1,711	-0,057	0,001	760,43
Bedfordshire - Middle School	2003	2004	2	1	1693	1669	76,1	74,2	2,120	2,076	-0,044	0,005	208,61
Bracknell Forest - Primary School	2003	2004	2	1	828	904	54	54,8	1,651	1,667	0,016	0,005	197,13
Hartlepool – Primary School - March	2003	2004	2	1	837	813	54,2	56	1,655	1,691	0,036	0,005	185,04
Lancashire – Primary Schools	1999	2004	2	1	3153	2911	49,8	42,7	1,567	1,424	-0,143	0,001	812,14
Lancashire – Urban Secondary Schools	1999	2004	2	1	1561	1734	75,9	77,8	2,115	2,160	0,045	0,005	190,26
Leeds – Primary Schools	2001	2004	2	1	2190	2317	50,5	51,9	1,581	1,609	0,028	0,002	549,53
Redcar & Cleveland - Primary Schools	2001	2004	2	1	2341	2173	76,1	70	2,120	1,982	-0,138	0,003	301,09
Redcar & Cleveland - Secondary Schools	2001	2004	2	1	1265	1284	65	70	1,875	1,982	0,107	0,005	205,99
Telford & Wrekin - Primary School	2000	2004	2	1	1712	1854	47,5	49,2	1,521	1,555	0,034	0,002	459,91
Thurrock – Primary Schools	2001	2005	2	1	4532	4331	60,4	54,8	1,780	1,667	-0,113	0,001	936,30
Shropshire – Primary Schools	2002	2004	2	1	2218	2299	72,4	75,2	2,035	2,099	0,064	0,003	295,21
Shropshire – Secondary Schools	2002	2004	2	1	2738	2692	90,4	91,3	2,512	2,543	0,031	0,008	123,85
Stockport – Primary Schools	2002	2004	2	1	1029	991	59,9	62,6	1,770	1,826	0,055	0,005	195,25
John Hampden Infant	2000	2003	5	1	275	275	45	85,4	1,471	2,357	0,887	0,032	31,73
Holmer Green Infant	2000	2003	5	1	180	180	30	74,6	1,159	2,085	0,926	0,030	33,55
Holy Trinity CE	2000	2003	5	1	266	266	41	69,6	1,390	1,974	0,584	0,019	53,37
West Wycombe Combined	2001	2003	5	1	209	209	22	44,1	0,976	1,453	0,476	0,015	68,06
Little Kingshill Combined	2001	2003	5	1	262	262	21,7	39	0,970	1,349	0,380	0,011	89,83
Marlow CE Infant	2000	2003	5	1	186	186	51	50,6	1,591	1,583	-0,008	0,022	45,76

Eingereicht bei der Zeitschrift Journal of Environmentally Psychology

4) Are Work Travel Plans Effective? –Systematic Review and Meta Analysis in the Transport Policy Domain

Sebastian Bamberg & Guido Möser

For transport policy the ongoing changes in central demographical, economic, ecological and political frame conditions in combination with massive financing problems creates a strong pressure to act. This challenge can only be solved if policymakers develop innovative, effective and affordable transport solutions. This task is not easy because transport is a complex and dynamic system, which is influenced by a variety of external factors. Thus policy and practice depend on scientific support to get reliable and valid knowledge about the causes of transport problems as well as the impact and cost/benefit ratio of alternative problem solving strategies. The complexity of transport-related research issues is also reflected by the trend that important research questions are tackled increasingly by international, often competitive research consortiums. As a consequence, the volume of data that need to be considered and evaluated by researchers as well as decision makers is constantly expanding. In many areas it has become simply impossible for the individual to read and critically synthesise the state of current knowledge, let alone keep updating this on a regular basis. Thus systematic research synthesis is a key part of evidence-based transportation research and policy.

The traditional and still most frequently used research synthesis approach is the narrative research review. For this purpose a respected expert in a field is asked to collate existing knowledge. Unfortunately, in the meantime there is strong evidence (e.g. Mulrow, 1987; Teagarden, 1989; Spector & Thompson, 1991; Antman, Lau, Kupelnick, & Chalmers, 1992; Lau , Antman, Jimenez-Silva, Kupelnick, Mosteller, Chalmers, 1992; Chalmers, Frank, Reitman, 1990) that the traditional narrative review provides no scientifically defensible way for sound research synthesis: Without guidance by formal rules, narrative reviews are subjective and therefore prone to bias and error. Different reviewers often disagree about issues as basic as what types of studies it is appropriate to include and how to balance the quantitative evidence they provide. Selective inclusion of studies that support the author's view is common: the frequency of citation of studies is related to their outcome, with studies in line with the prevailing opinion being quoted more frequently than unsupportive studies (e.g. Ranskov, 1992; Götzsche, 1987). Once a set of studies has been assembled, a common way to review the results is to count the number of studies supporting various sides of an issue and to choose the view receiving the most votes. This procedure is unsound as it ignores sample size, effect size, and research design. It is thus hardly surprising that narrative reviews often reach opposite conclusions and miss small, but potentially important, differences (e.g., Chalmers, Frank, & Reitman, 1990; Cooper & Rosenthal, 1980; Mulrow, 1987).

In medicine the disadvantages of the traditional narrative review has stimulated the search for more objective, transparent and valid research synthesis methods. An influential international initiative, the Cochran Collaboration, has devoted its work completely to the development of better review methods as well as the production of methodologically sound reviews. The central difference between the traditional narrative and the systematic review approach promoted by the Cochran Collaboration is that in a systematic review the whole selection and synthesis process of relevant information is guided by explicit rules aiming to minimise biases and random errors (Chalmers & Altman, 1995). Generally, a systematic review is guided by the following basis rules (e.g. Clarke, & Oman, 2000; Cooper & Hedges, 1994; NHS CRD, 2001):

Focusing on a precise review question. A criticism of the narrative review is its often unfocused nature. In contrast, systematic reviews focus on a specific question or questions. Developing the question(s) is an important, but often complex and time consuming part of the review process. However, concentrating on specific questions or problems gives systematic reviews a clarity of purpose and of content that should enhance their usefulness to others.

Using protocols to guide the review process**.** In a review protocol the strategy is a-priori specified which the review will follow to identify, appraise and collate evidence. By specifying a review protocol the reviewer is forced from the beginning to be as explicit as possible about how the review will be carried out. Thus a review protocol is a important tool for promoting transparency, transferability and replicability.

Seeking to identify as much of the relevant research as possible. Systematic reviews take a wide ranging and comprehensive approach to search for relevant research. They use the technology now available to carry out global searches for relevant data. They aim to identify as much of the relevant research as possible, not just the most well known, well promoted and successful. While it is not always possible to locate all the research in a given area, the review explains how studies were identified and obtained, and highlights any known gaps.

Appraising the quality of the research included in the review**.** Using exclusion and inclusion criteria set out in the protocol, the reviewers appraise the methodological quality of the studies identified to decide which studies warrant inclusion in the review. This means that decisions on inclusion are made explicit rather than implicit. The quality of the studies included in the review is also assessed.

Synthesising the research findings in the studies included**.** The findings of included studies are synthesised in different ways. CRD's guidance on systematic review (CRD, 2001, Report 4) discusses the different approaches to synthesis available to the systematic reviewer in more depth. The best known techniques are narrative synthesis and meta-analysis.

Quantitative research synthesise via meta-analysis

The term meta-analysis is reserved for the use of statistical methods to combine the results of multiple studies, generally with the aim to produce a single estimate of a treatment effect. Although the statistical methods involved may at first appear to be complex, their purpose is simple. They are trying to answer four basic questions (Lau, Loanidis & Schmid, 1997): Are the results of the different studies similar? In as far as they are similar, what is the best overall estimate? How precise and robust is this estimate? Finally, can dissimilarities be explained? To answer these questions, the tasks of a meta-analysis consists in (1) evaluating the statistical heterogeneity of the data, (2) estimating a common effect, (3) explaining heterogeneity, (4) assessing the potential for bias, and (5) presenting the results.

Evaluating the statistical heterogeneity of the data. This step is intended to answer the question ‚Are the results of the different studies similar (homogeneous)?'. It is important to answer this question before combining any data. To do this, one must calculate the magnitude of the statistical diversity (heterogeneity) of the study findings included in a review. Statistical diversity can be thought of as attributable to one or both of two causes. First, study results can differ because of random sampling error. Even if the true effect is the same in each study, the results of different studies would be expected to vary randomly around the true common fixed effect. This diversity is called the subject-level variance. Second, each study may have been drawn from a different population depending on the particular participants chosen and the interventions and conditions unique to the study. Therefore, even if each study enrolled a large sample, the treatment effect would be expected to differ. These differences, called study-level variance, describe the between-study variation with regard to an overall mean of the effects of all of the studies that could be undertaken. The test most commonly used to assess study-level heterogeneity is the Q statistic (Hedges & Olkin, 1985). It provides a measure of the sum of the squared differences between the results observed and the results expected in each study under the assumption that each study estimates the same common treatment effect. A large total deviation indicates that a single common treatment effect is unlikely.

Estimating a common effect. The questions that this step tries to answers are: (1) in as far as data are similar, what is their best common point estimate of a treatment effect, and (2) how precise is this estimate? To answer these questions the findings of different studies are combined (pooled) into an overall estimate. For this purpose each study is given a weight reflecting the precision of its results. Because studies with a greater sample size provide more reliable parameter estimations, they should be weighted more heavily than studies with small sample size. Mathematically, this precision can be expressed by the inverse of the variance of the estimate of each study, which is a direct function of study sample size. When the study-level variance is found to be or assumed to be zero, each study is simply weighted by this term. This approach characterises a fixed-effects model. Pooled results are generally reported as a point estimate with its 95% confidence intervals (CIs). Another advantage of pooling different studies is the increase of the statistical power to detect 'true' population effects of a treatment (Lipsey & Wilson, 1993).

However, when theoretical reasoning or the Q statistic indicate that the variability across effect sizes may be greater than expected from sampling error (subject-level variance) alone, it is difficult to defend the assumption of the fixed-effects model that study weighting by a term representing only subject-level sample error is sufficient to account for their differential precision as statistical estimates of population values. In this case a random-effects model provides a more adequate approach for calulating a common point estimate of a treatment effect. The random-effects model assumes that to represent the variation among effect sizes another variance component must be included in the statistical model in addition to subject-level sampling error (Kalaian & Raudenbush, 1996; Overton, 1998; Raudenbush, 1994). Since this additional variance component is assumed to either be, or act like, study-level sampling error, sampling error in the random effects model represents random variability at both the study-level and the subject-level. The random effects model, therefore, involves a different inverse variance weight than the fixed effects model, which must be used for recalculating the weighted mean and confidence interval computations. Generally, the random-effects model produces wider CIs than does the fixed-effects model, and the level of statistical significance may therefore be different depending on the model used. The pooled point estimate should be similar under both models (Lipsey & Wilson, 2001).

Explaining heterogeneity. Rather than assuming that effect size heterogeneity (study-level variance) is due to unobserved random sources, a researcher may believe that between studies variability can be systematically explained by study and effect size descriptors. One option to handle this situation is that a researcher continues to assume a fixed effects model, but adds the assumption that the variability beyond subject-level sampling error is systematic, that is, derived from identificable differences between studies. This added assumption is the basis for further analysing effect size variation in terms of the characteristics of the source studies that generate the effect sizes. Two statistical approaches can be used for modelling the study-level variance: Hedges' (1982) analogue to the analysis of variance and Hedges and Olkin's (1985) modified weighted multiple regression. The former handles categorical independent variables and, as the name implies, is similar to one-way analysis of variance (ANOVA). The latter handles continuous or dichotomous independent variables and can model multiple independent variables in a single analysis.

However, an effect size distribution may remain heterogeneous (significant Q-statistic) even after using study and effect size descriptors for modelling between-study differences. This indicates that the assumption of a fixed-effects model with only systematic variance (the modelled component) and subjective-level sampling error is untenable and a mixed-effects model should be considered. A mixed-effects model assumes that the effects of study-level variables, such as treatment type, are systematic but that there is a remaining unmeasured random effect in the effect size distribution in addition to sampling error. That is, variability in the effect size distribution is attributed to systematic (modelled) between study differences, subject-level sampling error, and an additional random component. Fitting a mixed effects model to the effect size data is similar to the method for fitting random-effects model. Under a mixed effects model the CIs of estimated parameter will be larger than under a fixed-effects model. As a consequence, the regression coefficients that may be significant under fixed-effects assumptions may no longer be so. An important statistical advantage of random- as well as mixed-effects models is that in the case of varying effect sizes they provide more correct estimations of the Type I error rates. Thus, a comparison of the results from fixed- and mixed-effects models is always advisable (Mosteller & Colditz, 1996; Overton, 1998).

Assessing the potential for bias. The assessment of potential bias should always be part of a meta-analysis. One major source of bias for meta-analysis is the failure to find all of the studies performed in a domain. Publication bias is often a problem because studies with negative results are more likely to remain unpublished. Furthermore, some studies may be impossible to retrieve and include in a meta-analysis despite a thorough search of potential databases. Publication bias is difficult to eliminate, but some procedures may be helpful in detecting its presence. Often an inverted funnel plot is used to visually explore the possibility that publication bias is present (e.g., Light & Pillemer, 1984). A funnel plot depicts graphically the relation between sample sizes and effect sizes. The logic behind this procedure is that small studies produce more variable effect size estimates than larger studies. Therefore, the most aberrant values that occur by chance are much farther from the mean ES than the aberrant values for large studies. If selective publication causes the more extreme effect sizes to be selected for publication, regardless of the sample size, then the ES from the small studies will be more extreme than those from the larger studies, leading to a relation between sample size and ES. If no bias is present, this plot should be shaped like a funnel, with the spout pointing up – that is, with a broad spread of points for the highly variable small studies at the bottom and decreasing spread as the sample size increases. However, the mean ES should be the same regardless of sample size. That is, one should be able to draw a vertical line through the mean effect size, and the points should be distributed on either side for all sample sizes. In other words, the funnel should not be skewed.

Presenting the Results. The results of meta-analyses are typically presented in a graphical form (so-called ‚Forrest plot') that shows the point estimates and the CIs of the single studies as well as the calculated weighted mean effect size and its CIs. This presentation aims to convey an impression of the results of the individual studies, to convey the extent of heterogeneity, and to report the pooled estimate.

The present research

The main aim of the present paper is to demonstrate how the techniques systematic review and meta-analysis can be used for synthesising research evidence in the transportation policy domain. Because of this programmatic goal we have decided to use a body of evaluation results which have been synthesised within a traditional narrative review approach thus far. This strategy also allows us to reach the second main goal of our paper: We want to compare the conclusions drawn within a narrative research synthesis approach from these data with the conclusions we draw from our quantitative meta-analysis of the same data.

For this purpose we use a body of evaluation data assembled in a research review commissioned in 2001 by the UK Department for Transport to Transport 2000 Trust in collaboration with Addison & Associates, University College London and Adrian Davis Associates. Review title is ‚Making travel plans work'. Authors are Sally Cairns (ESRC Transport Studies Unit, UCL), Adrian Davies (Adrian Davies Associates), Carey Newson (Transport 2000) and Camilla Swiderska (Transport 2000). The final report was published in 2002 and is electronically available at the homepage of the UK Department of Transport (www.dft.gov.uk., Sustainable Travel Section, 20.12.05). Although Cairns et al. (2002) never mention the term ‚systematic review', their work provides a good example of how to conduct a systematic review in the transportation policy domain. However, the conclusions Cairns et al. draw from the assembled data are based on a mainly narrative synthesis approach.

The Cairns et al. review as an example of a systematic review

Review focus. In the last decade growing evidence has been reported in the literature (e.g. TCRP, 1994; Schreffler and Organizational Coaching, 1996; Shoup, 1997; Ligtermoet, 1998; Touwen, 1999) that the implementation of travel work plans provide an effective strategy to reduce work-related car use, particularly commuting. A work travel plan is defined as a package of measures combining 'carrots' and 'sticks' with the aim of encouraging staff to commute to work in a more sustainable way. But there is still little knowledge about how organisational factors (e.g. type, size, staff characteristics), site factors (e.g. location, accessibility by public transport, scarcity of off-site parking), and travel plan characteristics (the measures introduced to promote reduction in work related car use) influence the success of a work travel plan. The aim of the Cairns et al. (2002) review is to analyse the impact of these factors on the effectiveness of implemented work travel plans that exemplify on many dimensions best practice in encouraging staff to commute to work more sustainably.

Review protocol. The report does not explicitly mention whether the reviewing process was prepared and guided by an a-priori formulated review protocol. However, it can be assumed that the Department of Transport's decision on who to commission with the review was based on a critical appraisal of competing proposals describing in detail the planned review methodology.

Identifying the relevant research. Cairns et al. (2002) have used the following three stage approach for systematically collecting their best practice work travel plan examples. In the first stage a comprehensive survey by Steer Davies Gleave (2001) served as one key resource for identifying potential best practice candidates. Taking all the local authorities in England and Wales into account, including 554 businesses, 45 hospitals, 29 higher education establishments, this survey has collected information whether they had introduced or plan to introduce a work travel plan. In the second stage a sample of best practice case studies were selected from this list of potential candidates which includes organisations from a range of sectors (health facilities, private sector companies, government organisations and local authorities); a range of locations (town centre, suburban and rural), and a range of sizes. The key criterion for including an organisation in the review was the existence of data allowing the effectiveness of the work travel plan to be monitored. During this stage 95 organisations were considered in detail but rejected – primarily due to insufficient monitoring data being available. A total of 38 organisations reporting that they had reasonable monitoring results available were short-listed for consideration. Final selection was based on a desire to ensure that a cross-section of organisations was represented, that organisations had achieved traffic reduction, that innovative good practice was retained in the study, and that the case study experience would be useful for as many other UK organisations as possible. For these reasons airports, organisations based in Central London and schools were excluded because it was felt that their experience might not be generalisable to other organisations or parts of the country. In the appendix of their review, Cairns et al (2002) provide a detailed description of the selection process. Table 1 presents the list of the 21 case studies finally included in the review.

To obtain comprehensive information on potential success factors for these 21 case studies, in the third stage a standardised interview was carried out in each organisation with the persons responsible for co-ordinating the organisation's work plan activities. The standardised interview included questions regarding the following domains: Site characteristics (number of employees, type of organisation, level of parking, type of location and staff profile like age, sex, income, duration of the travel plan), a full list of all travel plan measures, information on costs and savings associated with the travel plan and on sources of funding used to finance it, evidence of the travel plan's impact over time, including information on reduction in car use, and changes in modal split. The questionnaire is also documented in the review appendix.

<u>Appraising research quality.</u> As reported above, the existence of before-after data on staff's car use was a central inclusion criteria. Because the included case studies differ on how they measure this central effectiveness measure, Cairns et al. reported for each case study the information on which the effectiveness measure is based as well as methodological problems encountered when calculating it (Cairns et al., 2002, Table 4.1.1., p. 38-40).

Table 1: Name, Type, Staff Size and Location of the 21 Work Travel Plan Case Studies

Organisation	**Organisation Type**	**Staff size**	**Location**
Addenbrooke's NHS Trust	Health care	4977	Edge of town
Agilent Technologies	Telecommunication	1500	Rural
AstraZeneca	Pharmaceuticals	4200	Rural
Bluewater	Shopping centre	5500	Rural
Boots	Pharmaceuticals	7500	Outer suburbs
BP	Oil company	2100	Outer suburbs
Buckinghamshire County Council	County council	2200	Town centre
Computer Associates	Software	850	Edge of town
Egg	Financial services	880	Edge of town
Government Office for the East Midlands	Government	245	Town centre
Marks and Spencer Financial Services	Financial services	1100	Edge of town
Nottingham City Hospital NHS Trust	Health care	3500	Edge of town
Orange (Almondsbury Park)*	Telecommunication	2000	Edge of town
Orange (Temple Point)*	Telecommunication	400	Town centre
Oxford Radcliffe Hospitals NHS Trust	Health care	5170	Outer suburbs
Pfizer	Pharmaceuticals	5500	Rural
Plymouth Hospitals NHS Trust	Health care	4193	Outer suburbs
Stockley Park	Business park	7700	Outer suburbs
University of Bristol	University	4177	Town centre
Vodafone	Telecommunication	5400	Town centre
Wycombe District Council	District council	502	Town centre

* Two examples of Orange's travel planning work were examined. Hence, 20 organisations were contacted but, de facto, 21 travel plans were examined.

Synthesising the results of the included studies. A first important result is the considerable variability in car use reduction reported across the 21 case studies. Car reduction varies from a –52 %-point shift (Orange Temple Point) to a –3 %-point shift (Boots). The Median calculated across the findings of all 21 case studies is a 15% car reduction, equivalent to an average reduction of 12 commuter cars per 100 staff.

Exploring factors expected to explain the observed variability of work travel plan effectiveness, Cairns et al. (2002) view no evidence that organisation features like size, organisational type, lower paid staff, proportion of women employed, the age of the workforce or site location (e.g. rural vs. town centre) are associated with variation in effectiveness. However, according to these researchers the data underlines the central role of parking as a success factor of work travel plans: On the average the 13 case studies where parking has been addressed (either by restricting the number of staff entitled to park, introducing charges or providing incentive payments not to park) report a 24% reduction of commuting journeys made in a car, compared with an average of 10% reduction for those who had not used the car. Furthermore, the analyses suggest that the proportion of staff who are permitted to park is the key determinant of levels of car use, but parking charges will then have a secondary effect.

Meta-analytical re-analysis of the Cairns et al. (2002) data set

The Cairns et al. (2002) review contains a number of tables presenting the information assembled and analysised by the authors. Our meta-analysis uses these tables as data base.

Defining and calculating the effect size statistic. The key to meta-analysis is defining an effect size (ES) statistic capable of presenting the quantitative findings of a set of studies in a standardised form that permits meaningful numerical comparison and analysis across the studies. Cairns et al. (2002) use the before / after number of commuter cars arriving per 100 staff as a central effectiveness measure. From the change of these proportions they calculate the percentage reduction in commuter journeys made as a car driver. Staff who parked off-site were counted as bringing a car. Staff using Park-and-Ride services for commuting were not counted as bringing a car. In the majority of cases, calculations were based on results from before / after staff travel surveys – where the number of cars arriving per 100 staff was inferred from the percentage of staff arriving as a car driver on a typical day (including both solo drivers and car sharer drivers). In some cases, the indicator was calculated by dividing counts of the number of cars arriving by the number of staff on site on a typical day (plus home workers, where appropriate).

To prevent a negative sign of the ES's, in the first step we converted the car-use proportions reported by Cairns et al. into their corresponding no-car-use proportion (1 – car-use proportion, see Table 2, column 2 and 3). However, using change in the before/after no-car-use proportions directly as an ES statistic would have statistical disadvantages (see e.g., Lipsey & Wilson, 2001). Thus instead of the raw proportions statisticians recommend the use of arcsine-transformed proportions for calculating the ES's. The arcsine method is borrowed from statistical power analysis (Cohen, 1988) and creates an ES for the difference between proportions whose statistical power is independent of the location of a proportion between 0 and 1. Column 5 and 6 of Table 2 present the arcsine-transformed before/after no-car-use proportions as well as the ES's resulting from subtracting the transformed before proportion from the transformed after proportion (so-called Cohen's h, Table 2, column 7). In a second step for each ES the standard error term ($1/n_{NoCARbefore}$ + $1/n_{NoCARafter,}$ Table2, column 10) was calculated, whose inverse ($w = 1/SE^2$, Table 2, column 11) is used as weight with which each primary study contributes to the calculation of the common mean ES. One problem encountered when calculating the variance of Cohen's h was that Cairns et al. (2002, Table 4.1.1, p. 38) provide no complete information about the sample sizes of before/after surveys on which the reported number of commuter cars arriving per 100 staff is based. To solve this missing value problem we have used for each case the total number of staff using no car before and after the intervention for calculating the variance. We view total number of staff using no car before and after the intervention (Table 2, column 7 and 8) as the most reliable estimate of the probable sample size. For example, the estimated variance of the effect size Orange (Temple Point) in Table 2 was calculated by multiplying staff size (see Table 1) with No car per 100 staff before = No car before total, here 400 * 21% = 84 (no car before total). No car per 100 staff after for Orange (Temple Point) was calculated by 400 * 73% = 292 (no car after total). Finaly, the estimated variance was than calculated by 1/84 + 1/292 = 0,015239.

<u>Evaluating the statistical heterogeneity of the data</u>. As mentioned above, an important question to ask is whether the various ES's that should be averaged into a common mean value all estimate the same population ES (Hedges, 1982; Rosenthal & Rubin, 1982). For the present study set, just a short view on the %-point shift of no-car use (Table 2, column 3) reported in the 21 case studies

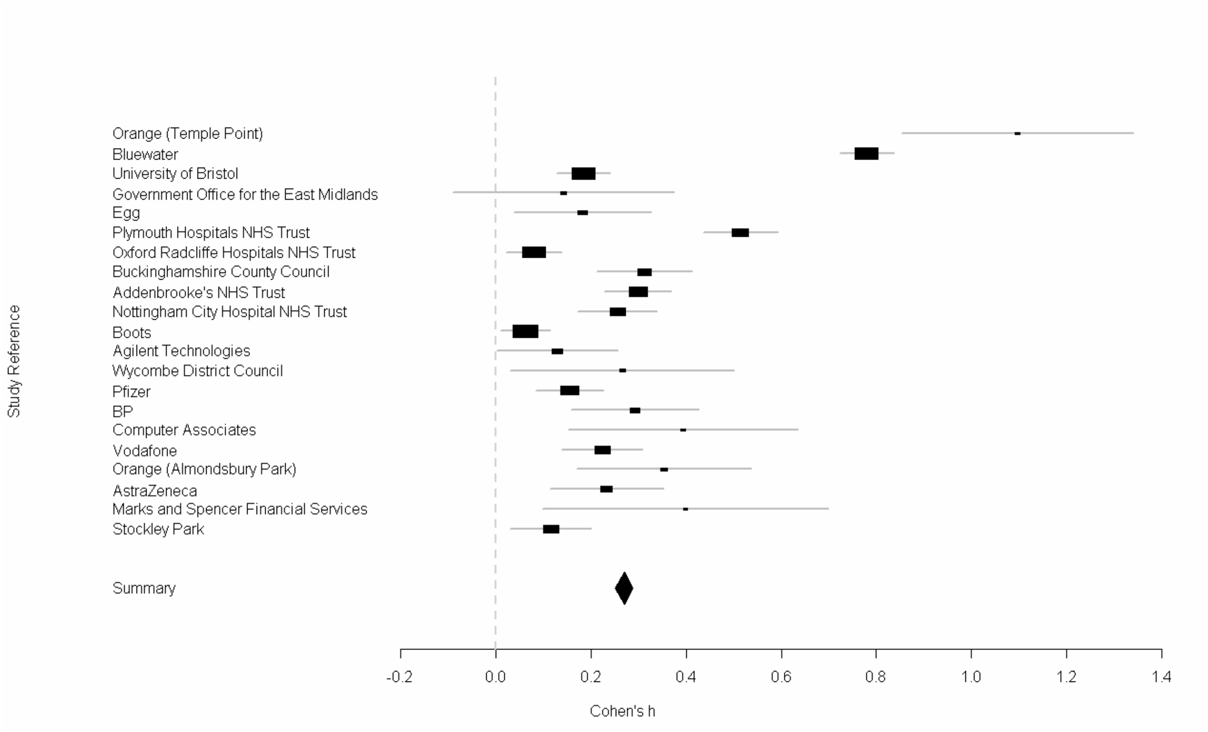
indicates strong heterogeneity of the findings. This was confirmed by the formal Q statistic (Q-value = 531,87; df = 20; $p < .001$). To check the role of potential outliers as heterogeneity source, the ES-distribution was inspected visually by a box-and-whisker plot. For the present study set the box-and-whisker plot indicates two potential out-liners (Orange Temple Point, ES = 1,1; and Bluewater, ES = 0.78). Recalculating the Q-statistic without these two outliners resulted in a much lower, however still significant Q-value (134,8122, df =18, $p < .001$).

Estimating a common effect size. Firstly a fixed-effects model was used for calculating a mean ES. For this purpose, for each of the 21 case studies the calculated Cohen's h was weighted by the inverse of its variance (w). Across all 21 case studies, the calculated fixed-effects mean ES is .27. The z-test value of this point estimate is 27.56, which exceeds the critical value of 1.96 (α-level .05). Correspondingly, the 95% CIs around the mean ES ($0.25 < \mu < 0.29$) do not include zero. Figure 1 shows Forest Plot of the 21 ES's.

Table 2: Before / After Percentage of Staff Not Commuting With Car and the Calculation of the Effect Size (Cohen's h) for the 21 Work Travel Plans reported by Cairns et al. (2002)

Case Study	No car per 100 staff - before	No car per 100 staff - after	%-Point Shift	No car before arcsine	No car after arcsine	Effect size Cohens h	No car before total	No car after total	Variance Cohens h	Weight w
Orange (Temple Point)	21	73	52	0,952	2,049	1,097	84	292	0,015329	65,23
Bluewater	31	69	38	1,181	1,961	0,780	1705	3795	0,000850	1176,45
University of Bristol	56	65	9	1,691	1,875	0,184	2339	2715	0,000796	1256,55
Government Office for the East Midlands	55	62	7	1,671	1,813	0,142	135	152	0,014004	71,41
Egg	38	47	9	1,328	1,511	0,182	334	414	0,005408	184,90
Plymouth Hospitals NHS Trust	22	46	24	0,976	1,491	0,514	922	1929	0,001603	624,02
Oxford Radcliffe Hospitals NHS Trust	42	46	4	1,410	1,491	0,081	2171	2378	0,000881	1135,05
Buckinghamshire County Council	29	44	15	1,137	1,451	0,313	638	968	0,002600	384,55
Addenbrooke's NHS Trust	26	40	14	1,070	1,369	0,299	1294	1991	0,001275	784,25
Nottingham City Hospital NHS Trust	27	39	12	1,093	1,349	0,256	945	1365	0,001791	558,41
Boots	35	38	03	1,266	1,328	0,062	2625	2850	0,000732	1366,44
Agilent Technologies	29	35	06	1,137	1,266	0,129	435	525	0,004204	237,89
Wycombe District Council	23	35	12	1,000	1,266	0,266	115	176	0,014353	69,67
Pfizer	25	32	7	1,047	1,203	0,155	1375	1760	0,001295	771,93
BP	16	28	12	0,823	1,115	0,292	336	588	0,004677	213,82
Computer Associates	11	26	15	0,676	1,070	0,394	94	221	0,015220	65,70
Vodafone	16	25	9	0,823	1,047	0,224	864	1350	0,001898	526,83
Orange (Almondsbury Park)	08	20	12	0,574	0,927	0,354	160	400	0,008750	114,29
AstraZeneca	10	18	8	0,644	0,876	0,233	420	756	0,003704	270,00
Marks and Spencer Financial Services	05	17	12	0,451	0,850	0,399	55	187	0,023529	42,50
Stockley Park	12	16	4	0,707	0,823	0,116	924	1232	0,001894	528,00

Fig. 1: Graphical Representation of the distribution of the 21 ES's from Cairn's et al.



The black rectangle and horizontal lines correspond to the ES of each case study and their 95% CIs. The area of the black rectangles reflects the weight each study finding contributes in the meta-analysis. The diamond (.27) reflects the weighted mean ES across the 21 studies with its 95% CIs.

Because the above inspection of the ES distribution indicates that the ES of two case studies should be treated as outliers, we calculated a fixed-effects model without the two outliers. The resulting mean ES is .20 (z-test value = 19.05; 95% CI = $0.18 < \mu < 0.22$).

As reported above, after deleting the two potential outliers the Q-statistic is still significant. That means that the variability across effect sizes is greater than expected from sampling error (subject-level variance) alone. For this reason, an random-effects model was calculated, which takes study-level heterogeneity as additional variance component into account. In the present analysis a noniterative method of moments (Lipsey & Wilson, 2001) is used for estimating the additional random variance component (v = 0.053) and add to the variance of each study ES. Across the 21 case studies the random-effects mean ES is .30, which is similar to fixed-effects mean ES of .27. The random-effects mean ES is also significant (z-value of 5.69; $p > .001$), however, compared with the fixed-effects model the z-value is much lower. This results demonstrate that under the condition of varying effect sizes the inadequate application of a fixed-effects model inflates Type I error. In inadequate z-values may indicate in this case a significant mean ES which is actually nonsignificant. The random-effects mean ES without the two out-liner case studies is .23 (z-value = 7,48; 95% CI = $0.17 < \mu < 0.29$).

Explaining heterogeneity: Rather than assuming that ES heterogeneity is due to unobserved random sources, the central aim of the Cairns et al (2002) review is to identify characteristics of the organisation, the site, and the implemented work travel plan that might explain the strong variability in the effectiveness of the 21 work travel plans. As discussed above, within the meta-analytical approach meta-regression provides a statistical tool for systematically analysing the impact of these independent study-descriptors on heterogeneity. Although most statistical software programs, such as SPSS or SAS, perform weighted least squares regression, they report inaccurate standard errors and, hence, statistical significance when applied to effect size data (for further details see Lipsey & Wilson, 2001). With some side calculations performed on the output of these programs, however, this problem is easily overcome. In the present analysis a SPSS macro written by Lipsey and Wilson (2001) is used for this purpose.

When practically conducting a meta-regression, one should recall that all of the problems that arise in connection with multiple regression analysis can also arise in meta-regression. Collinearity may degrade the quality of estimates of regression coefficients, wildly influencing their values and increasing their standard errors. The same procedures that are used to safeguard against excessive collinearity in normal regression analysis in primary research are useful in meta-regression. Examination of the correlation matrix of the predictors and the exclusion of some predictors that are too highly intercorrelated with others can often be helpful.

Table 3-5 summarises the distribution of 25 case study descriptors extracted from the Cairns et al. (2002) review. These descriptors can be assigned to four predictor sets: descriptors related to the monitoring process, organisational descriptors, site descriptors, and descriptors of the implemented work travel plans.

Because the combination of a relatively small sample of studies with many potential predictors makes collinearity problems likely, a multi-stage approach was used for conducting the meta-regression. Because in small sample sizes outliers exert a strong bias on the estimated regression coefficients, the two studies identified above as outliers were excluded from meta-regression. Then the bivariate association between each study descriptor and the ES distribution was calculated (Table 6). The later comparison of the bivariate and multivariate results (signs, magnitude of coefficients) provides one possibility to check for collinearity. In the third step the intercorrelation of the predictors was inspected for identifying highly associated predictors. In the fourth step, a multivariate meta-regression was performed for each of the four predictor sets (Table 7) separately. In the last step all study descriptors, significant in the four separate set-specific analyses, were used as predictors in a final fixed- respective mixed-effects multivariate meta-regression model (Table 8).

Table 3: Characteristics of the Monitoring Process and Organisation's Staff for the 21 Travel Plan Case Studies reported by Cairns et al. (2002)

Case study	Data problems	Date of before monitoring	Survey period (month)	Gender bias	Below average income	Age (Young)	Staff within 3-5 miles
Orange (Temple Point)	1	2001	0	0	0	1	13
Bluewater	1	1999	14	1	1	1	13
University of Bristol	0	1998	36	0	0	0	
Government Office for the East Midlands	0	1997	24	0	0	0	13
Egg	1	1999	16	0	1	1	20
Plymouth Hospitals NHS Trust	0	1995	72	1	0	0	13
Oxford Radcliffe Hospitals NHS Trust	0	2000	12	1	0	1	27
Buckinghamshire County Council	0	1998	29	0	0	0	39
Addenbrooke's NHS Trust	0	1993	72	1	0	0	25
Nottingham City Hospital NHS Trust	0	1997	36	1	0	0	30
Boots	1	1995	53	0	1	0	10
Agilent Technologies	0	1997	24	0	0	0	8
Wycombe District Council	0	1998	12	0	0	0	44
Pfizer	0	1998	36	0	0	0	10
BP	0	1998	37	0	0	0	8
Computer Associates	0	2000	16	2	0	0	4
Vodafone	1	1998	40	2	0	1	22
Orange (Almondsbury Park)	0	1996	63	0	0	1	14
AstraZeneca	0	1997	48	0	0	0	0
Marks and Spencer Financial Services	1	1998	14	1	0	1	4
Stockley Park	0	1997	24	0	0	1	13

Notes:
Methodological problems: derived from Table 4.1.1 in Cairns et al. (2002); Survey period: months between before/after monitoring; Gender bias: 0 = no bias, 1 = >70% female, 2 = >70% male; Below average income: 0 = average, 1 = below average; Age: 1= staff mainly < 35 years, Staff within 3-5 miles:the proportion of staff living within reasonable cycling distance.

Table 4: Site Characteristic of the 21 Travel Plan Case Studies reported by Cairns et al. (2002)

Case study	Walking access	Cycle access	No. of am peak services	Off-side Parking	Parking per 100 Staff	<100% Parking
Orange (Temple Point)	3	3	38	2	14	1
Bluewater	2	2	58	1	31	1
University of Bristol	3	3	30	2	26	1
Government Office for the East Midlands	3	1	10	3	18	1
Egg	2	1	28	2	57	0
Plymouth Hospitals NHS Trust	2	3	44	1	32	1
Oxford Radcliffe Hospitals NHS Trust	2	3	60	2	28	1
Buckinghamshire County Council	3	2	40	3	27	1
Addenbrooke’s NHS Trust	2	3	21	2	48	0
Nottingham City Hospital NHS Trust	2	3	4	1	34	0
Boots	3	4	25	1	57	0
Agilent Technologies	2	1	12	3	66	0
Wycombe District Council	3	2	86	3	100	0
Pfizer	1	3	23	1	73	0
BP	2	1	14	3	76	0
Computer Associates	3	3	6	1	97	0
Vodafone	2	1	12	1	72	0
Orange (Almondsbury Park)	1	2	9	2	55	0
AstraZeneca	1	1	14	1	78	0
Marks and Spencer Financial Services	1	2	13	1	84	0
Stockley Park	3	4	24	1	78	0

Notes:
Walking access: 1 = poor, 2 = medium, 3 = good; cycle access: 1 = difficult, 2 = average, 3 = good, 4 = excellent; number of am peak services refers to the number of bus and train services arriving within a quarter of a mile of the site between 8am and 9am (including free company huttle buses and Park and Ride services); <100% park: organisations where less than 100% of staff are entitled to park in the organisation’s own car park; Parking per 100 staff has been calculated using the number of full-time staff, or the number of staff on site during core hours; amount of off-site parking opportunities: 1 = few, 2= some, 3 = sample.

Table 5: Description of the Implemented 21 Travel Plan Cases Studies Reported in the Cairns et al. (2002) Review

Case study	Parking charge	Parking charge (£)	FI	PT (6)	Cycle (7)	Walk (4)	Car share (4)	Costs (£)per 100 staff (gross)
Orange (Temple Point)	0	0	1	3	3	2	2	51
Bluewater	0	0	0	6	5	3	0	36
University of Bristol	1	10,6	0	5	7	4	1	53
Government Office East Midlands	0	0	0	3	2	2	2	42
Egg	1	3,8	0	4	3	2	1	91
Plymouth Hospitals NHS Trust	1	2,5	1	4	4	2	3	36
Oxford Radcliffe Hospitals NHS Trust	1	0,4	0	2	4	2	1	22
Buckinghamshire County Council	0	0	0	2	5	3	2	57
Addenbrooke's NHS Trust	1	1,5	0	1	6	3	1	41
Nottingham City Hospital NHS Trust	1	1,2	0	4	6	2	0	71
Boots	0	0	0	3	7	3	2	43
Agilent Technologies	0	0	0	3	3	0	1	2
Wycombe District Council	0	0	0	0	6	2	1	6
Pfizer	0	0	0	5	5	3	1	50
BP	0	0	0	6	4	3	0	226
Computer Associates	0	0	1	3	6	4	4	325
Vodafone	0	0	1	4	2	0	2	431
Orange (Almondsbury Park)	0	0	0	4	3	0	0	100
AstraZeneca	0	0	0	3	6	0	3	108
Marks and Spencer Financial Services	0	0	0	2	6	0	4	71
Stockley Park	0	0	0	4	6	3	0	39

Notes:

Parking charge: weekly charge for parking on the site for those entitled to do so; FI: a general incentive payment to staff for giving up their parking permit or for using alternative modes; PT (6): introduced measures for promoting bus/rail: having cheap or free services available; providing a dedicated company shuttle bus, increasing the number of services available, improving the quality of services, negotiating discounts on public services and improving off-site infrastructure; Cycle (7): introduced measuresfor promoting cycling: improvements to off-site cycle access; increasing cycle parking; introducing showers, changing and locker facilities; existence of a Bicycle Users Group; events to promote cycling; cycle repairs service and discounts on cycle equipment;Walk (4): introduced measures for promoting walking: good site access or improvements to site access during the monitoring period; improvements in on-site conditions; on-site security and direct marketing of walking (usually on health grounds); Car share (4): introduced measures for promoting car sharing: offering a centrally co-ordinated matching system; events to enable car-sharers to meet such as major lunch events; dedicated parking and significant incentive payments or relief from parking charges.

Table 6: Bivariate Fixed-Effects Meta-Regression with Effect Size as Dependent Variable and the 25 Study Descriptors as Predictors

Monitoring descriptors	B	95% CI	Z	P	Beta	R^2
Data problems	-.1024	-.1509 / -.0539	-4.14	***	-.31	.13
Date of before monitoring	-.3365	-.4868 / -.1861	-4,39	***	-.29	.09
Monitoring (in month)	.0033	.0022 / .0044	5,96	***	.49	.24
Organisation descriptors						
Organisation Type	.1274	.0837 / .1711	5,71	***	.55	.30
Staff size	-.0034	-.0045 / -.0023	-6,01	***	-.16	.03
Female bias	.0883	.0452 / .1314	4,02	***	.57	.33
Male bias	.0475	-.0358 / .1307	1,12		.05	.00
Below average income	-.1467	-.2012 / -.0921	-5,27	***	-.50	.25
Age (Young)	-.0768	-.1226 / -.0311	-3,29	***	-.28	.08
Staff within 3-5 miles	.0017	.0006 / .0039	1,46		.23	.05
Site descriptors						
Location	.0284	.0083 / .0486	2,76	**	.17	.03
Walking access	-.0454	-.0755 / -.0153	-2,96	**	-.28	.08
Cycle access	-.0418	-.0635 / -.0202	-3,79	***	-.25	.06
No. of am peak services	-.0008	-.0020 / .0005	-1,18		-.01	.00
Off-side parking	.0091	.0212 / .0395	0,59		-.05	.00
Parking per 100 Staff	-.0008	-.0018 / .0002	-1,54		-.26	.07
<100% Parking	.0395	-.0027 / .0816	1.83		.24	.06
Travel plan descriptors						
Parking charge	.0667	.0258 / .1075	3,20	**	.43	.19
Parking charge (£)	.0034	-.0024 / .0092	1,15		.05	.00
FI	.2115	.1512 / .2719	6,87	***	.65	.42
PT (6)	.0050	-.0104 / .0203	0,63		.01	.00
Cycle (7)	-.0192	-.0326 / -.0058	-2,82	**	-.29	.08
Walke (4)	-.0161	-.0340 / -.0017	-1,77		-.20	.04
Car share (4)	.0435	.0206 / .0664	3,72	***	.31	.10
Costs per 100 staff (gross)	.0002	.0000 / .0004	1,86		.09	.00

Notes: B = unstandardized bivariate regression coefficient; CI = confidence interval; Z = value of z statistic; *** p < .001; ** p < .01; * p < .05; Beta = standardized bivariate regression coefficient; R^2 = percentage of study-level variance explained by each study descriptor

Table 7: Four Multivariate Fixed-Effects Meta-Regression Models with Effect Size as Dependent Variable and One of the Four Blocks of Study Descriptors as Predictor

Monitoring descriptors	B	95% CI		Z	P	Beta	Model R^2
Data problems	-,1325	-,1817	-,0832	-5,27	***	-.46	.47
Date of before monitoring	---	---	---	---	---	---	
Monitoring (in month)	,0039	,0028	,0050	6,79	***	.59	
Organisation descriptors							
Organisation Type	,1502	,0763	,2241	3,98	***	,54	.63
Staff size	-,0016	-,0030	-,0002	-2,28	*	-,07	
Female bias	,0578	,0027	,1129	2,06	*	,20	
Male bias	---	---	---	---	---	---	
Below average income	-,0940	-,1637	-,0242	-2,64	**	-,35	
Age (Young)	-,0311	-,0902	,0280	-1,03		-,16	
Staff within 3-5 miles	-,0043	-,0073	-,0013	-2,78	**	-,29	
Site descriptors							
Location	,0427	,0105	,0750	2,59	**	,36	.26
Walking access	-,0679	-,1158	-,0201	-2,78	**	-,38	
Cycle access	-,0223	-,0592	,0146	-1,19		-,17	
No. of am peak services	-,0003	-,0019	,0013	-0,33		-,04	
Off-side parking	-,0258	-,0695	-,0179	-1,16		-,14	
Parking per 100 Staff	-,0011	-,0025	-,0003	-1,49		-,18	
<100% Parking	---	---	---	---	---	---	
Travel plan descriptors							
Parking charge	,0495	-,0010	,0999	1,92		,20	.44
Parking charge (£)	---	---	---	---	---	---	
FI	,2623	,1409	,3836	4,24	***	,73	
PT (6)	,0016	-,0169	,0200	0,17		,02	
Cycle (7)	,0043	-,0184	,0270	0,37		,05	
Walke (4)	-,0035	-,0323	,0254	-0,24		-,03	
Carshareing (4)	-,0024	-,0384	,0335	-,13		-,02	
Costs per 100 staff (gross)	-,0003	-,0006	,0001	-1,58		-,21	

Notes: B = unstandardized bivariate regression coefficient; CI = confidence interval; Z = value of z statistic; *** $p < .001$; ** $p < .01$; * $p < .05$; Beta = standardized bivariate regression coefficient; R^2 = percentage of study-level variance explained by each study descriptor set; --- = predictors deleted because of high intercorrelation with other predictors ($r > .70$).

Table 8: Final Fixed-Effects Multivariate Meta-Regression Model

Homogeneity Analysis	Q	df	P				
Model	105,48	5	***				
Residual	29,32	13	**				
Total	134,80	18	***				
Study descriptor	B	95% CIs		Z	P	Beta	Model R^2
Constant	-,1236	-.2053	-,0418	-2,96	**		.78
Difficult/average cycling access	,1126	,0567	,1685	3,94	***	.39	
Female bias	,0929	,0453	,1406	3,82	***	.36	
Organisation Type	,1093	,0646	,1539	4,80	***	.42	
Monitoring (in months)	,0027	,0015	,0039	4,42	***	.41	
Incentive for not parking (FI)	,0951	,0265	,1636	2,72	**	.27	

Notes:
B = unstandardized bivariate regression coefficient; CIs = confidence interval; Z = value of z statistic; *** p < .001; ** p< .01; Beta = standardized bivariate regression coefficient; R^2 = percentage of study-level variance explained by the final descriptor set

Table 9: Final Mixed-Effects Multivariate Meta-Regression Model

Homogeneity Analysis	Q	Df	P				
Model	55,06	5	***				
Residual	18,72	13					
Total	73,78	18	***				
Study descriptor	B	95% CIs		Z	P	Beta	Model R^2
Constant	-,0966	-,2054	,0122	-1,74			.75
Difficult/average cycling access	,1046	,0360	,1732	2,99	**	,40	
Female bias	,0866	,0193	,1539	2,52	*	,34	
Organisation Type	,0990	,0385	,1594	3,20	**	,38	
Monitoring (in months)	,0026	,0010	,0043	3,18	**	,40	
Incentive for not parking (FI)	,0994	,0143	,1845	2,29	*	,29	
Random Variance Component V^a	,00114	SE (V)		,00105			

Notes:
B = unstandardized bivariate regression coefficient; CIs= confidence interval; Z = value of z statistic; Beta = standardized bivariate regression coefficient; R^2 = percentage of total variance explained by the final descriptor set; *** p < .001; ** p < .01; * p< .05; ^{a} = ML-estimator; SE = standard error.

As can be seen from Table 6 the results of the bivariate fixed-effects meta-regression indicate a significant bivariate association with the ES distribution for 15 of the 25 study descriptors. The study descriptors duration of monitoring, organisational type, staff with female bias, below average income, charge for parking, and incentive payment for giving up parking show the strongest bivariate association with the ES distribution.

Results of the four setwise multivariate meta-regressions (Table 7) indicate that the set of the six organisation descriptors explains most ES variance, followed by the set of three monitoring process descriptors, the set of eight work travel plan descriptors, and the set of seven site characteristic descriptors.

Table 8 presents the results of the final fixed-effects multivariate meta-regression model: From the 10 predictors, significant in the four separated setwise regression analyses, only the following five remain significant predictors of the ES variability: cycling access, staff with female bias, organisation type, duration of the monitoring process, and incentive payment for giving up parking. More detailed analyses show that work travel plans implemented on sites with poor or average cycling access have stronger ES's than travel plans implemented on sites with good or excellent cycling access. Therefore, in Table 8 a dichotomous variable was used with poor/average cycling access as 1 and good / excellent cycling access as 0. All predictors have positive signs that is higher values are associated with greater ES's. That means that travel plans implemented on a site with poor/average cycling access, in a public organisation with a proportion of female staff above 70%, with a longer duration of the monitoring process, and financial incentive for giving up parking report the strongest ES's. Together the five descriptors explain 78 % of the study-level heterogeneity.

However, as can be seen from Table 8, even after modelling systematic sources of between-study variability, the residual Q-value is still significant. As discussed above, this indicates that a fixed-effects model which takes into account only subject-level sampling error and systematic variance (the modelled component) is untenable and a mixed-effects model should be considered. In a mixed-effects model variability in the ES distribution is attributed to subject-level sampling error, systematic (modelled) between study differences, and an additional random variance component v. Table 9 presents the meta-regression coefficients estimated under the mixed-model assumption. Adding the additional random variance component to the study weights results in a slight decrease of the estimated regression coefficients. However, despite the wider CIs all the coefficients remain significant. This supports the stability of the specified regression model. Per definition in the mixed-effects model the residual Q statistic is insignificant.

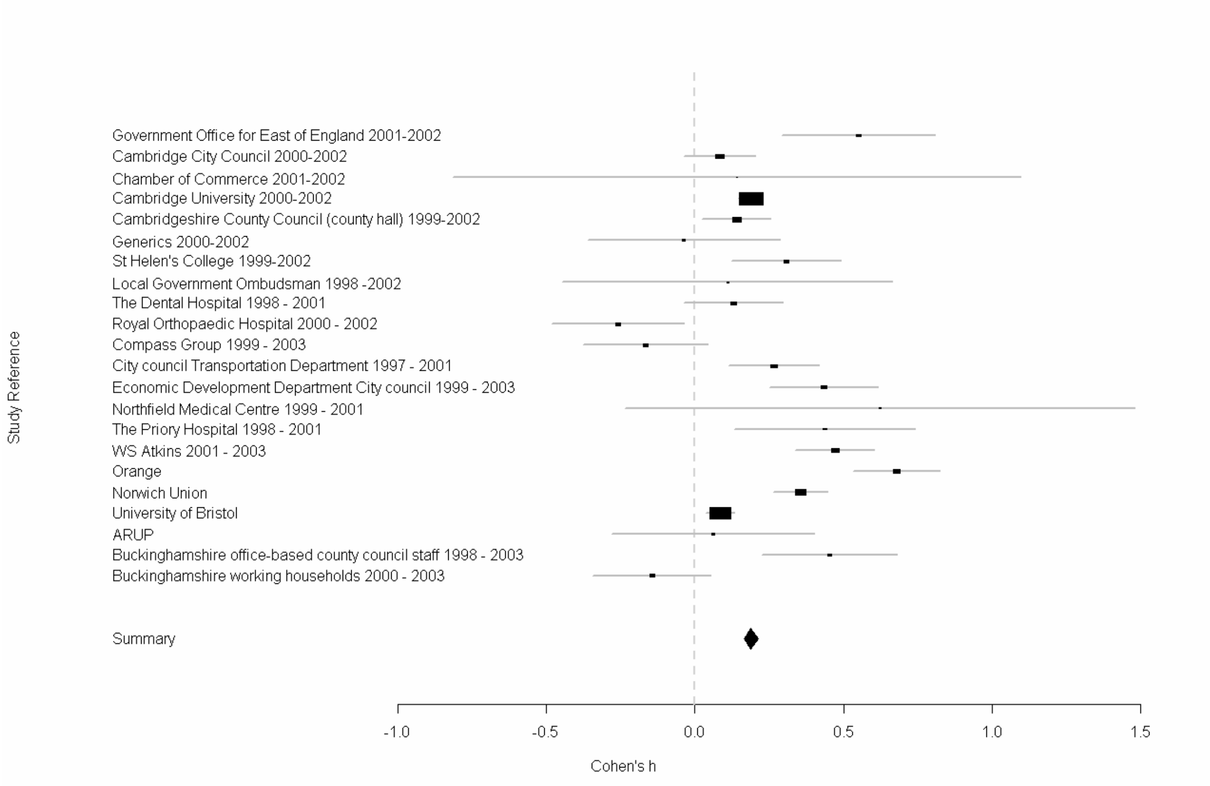
Testing the generalisability of the estimated mean ES. Across the 21 case studies the above analyses indicate for work travel plans a random-effects mean ES of .30 which decreases to .23 after deleting the two out-liners. Because this result is based on a highly pre-selected sample of ‚best-practice' case studies one may ask whether it provides a generalisable estimate for the average effect of ‚normal' travel plans. Fortunately we have found another review by Cairns, Sloman, Newson, Anable, Kirkbride, and Goodwin (2004), which documented additional data monitoring the effectiveness of 'normal' work travel plans and thus allows an empirical test of this question. The review titled 'Smarter Choices: Changing the Way We Travel' is also available via the homepage of the UK Department of Transport (www.dft.gov.uk., Sustainable Travel Section, 20.12.05). Because the scope of the Cairns et al. (2004) review is much broader (it tries to summarise the existing evidence for the effectiveness of a variety of so-called ‚soft transport policy measures') it does not provide the detailed organisation, site and intervention information reported in the Cairns et al. (2002) review. However, it uses the same before/after proportion of commuter cars arriving per 100 staff as the central effectiveness measure. Thus the second data set can be directly compared with the previous one. Table 10 presents the information necessary for calculating the ES statistic (Cohen's h) for these additional 23 studies.

Table 10: Before / After Percentage of Staff Not Commuting With Car and the Calculation of the Effect Size (Cohen's h) for the 44 Work Travel Plans reported by Cairns et al. (2004)

Case Study	No car per 100 staff - before	No car per 100 staff - after	%-Point Shift	No car per Arcsine Before	100 staff Arcsine After	Effect size Cohens h	Variance Cohens h	Weight W	Staff Size (bef.)	Staff Size (after)
Generics	34	33	-1	1,251	1,213	-0,038	0,027	36,71	220	
Arup	59	62	3	1,752	1,813	0,061	0,030	32,95	109	
Compass Group	39	31	-8	1,349	1,181	-0,168	0,011	87,33	400	640
WS Atkins	47	70	23	1,511	1,982	0,472	0,005	216,35	783	750
Orange	40	73	33	1,369	2,049	0,679	0,006	180,88	700	
Government Office for East of England	31	58	27	1,170	1,723	0,551	0,017	57,79	290	
Norwich Union	63	79	16	1,834	2,190	0,356	0,002	455,64	1300	
HM Prison	36	10	-26	1,287	0,644	-0,644	0,020	50,87	650	
Cambridge City Council	65	69	4	1,882	1,965	0,083	0,004	268,77	800	
University of Bristol	64	68	4	1,855	1,939	0,084	0,001	1648,49	5000	
Local Government Ombudsman	27	32	5	1,093	1,203	0,110	0,080	12,45	85	
Cambridgeshire County Council	49	56	7	1,551	1,691	0,140	0,003	287,47	1100	
Chamber of Commerce	43	51	8	1,438	1,581	0,142	0,238	4,20	18	
Cambridge University	64	73	9	1,861	2,049	0,188	0,000	2136,70	6250	
St Helen's College	23	37	14	1,000	1,308	0,307	0,009	113,47	800	
City council Transportation Depart.	52	65	13	1,611	1,875	0,265	0,006	166,22	562	593
Economic Develop.Depart. City council	50	71	21	1,571	2,004	0,433	0,009	114,26	423	350
Buckinghamshire working households	26	20	-6	1,075	0,930	-0,145	0,010	96,68	850	
Buckinghamshire office county	28	50	22	1,117	1,571	0,453	0,014	73,76	410	
Royal Orthopaedic Hospital	38	26	-12	1,328	1,070	-0,258	0,013	77,19	500	500
The Dental Hospital	66	72	6	1,897	2,026	0,130	0,007	137,74	400	400
The Priory Hospital	21	41	20	0,952	1,390	0,438	0,024	41,66	300	300
Northfield Medical Centre	14	41	27	0,767	1,390	0,623	0,192	5,22	50	50

The box-and-whisker plot of this new ES distribution indicates one potential out-liner (HS Prison, ES = -.64). For the remaining 22 studies, the fixed-effects mean ES is .19. (z-value = 14.89; 95% CI = 0.16 < μ < 0.21). Figure 2 shows the Forest plot of this new data set.

Fig. 2: Forrest plot for the additional 22 studies

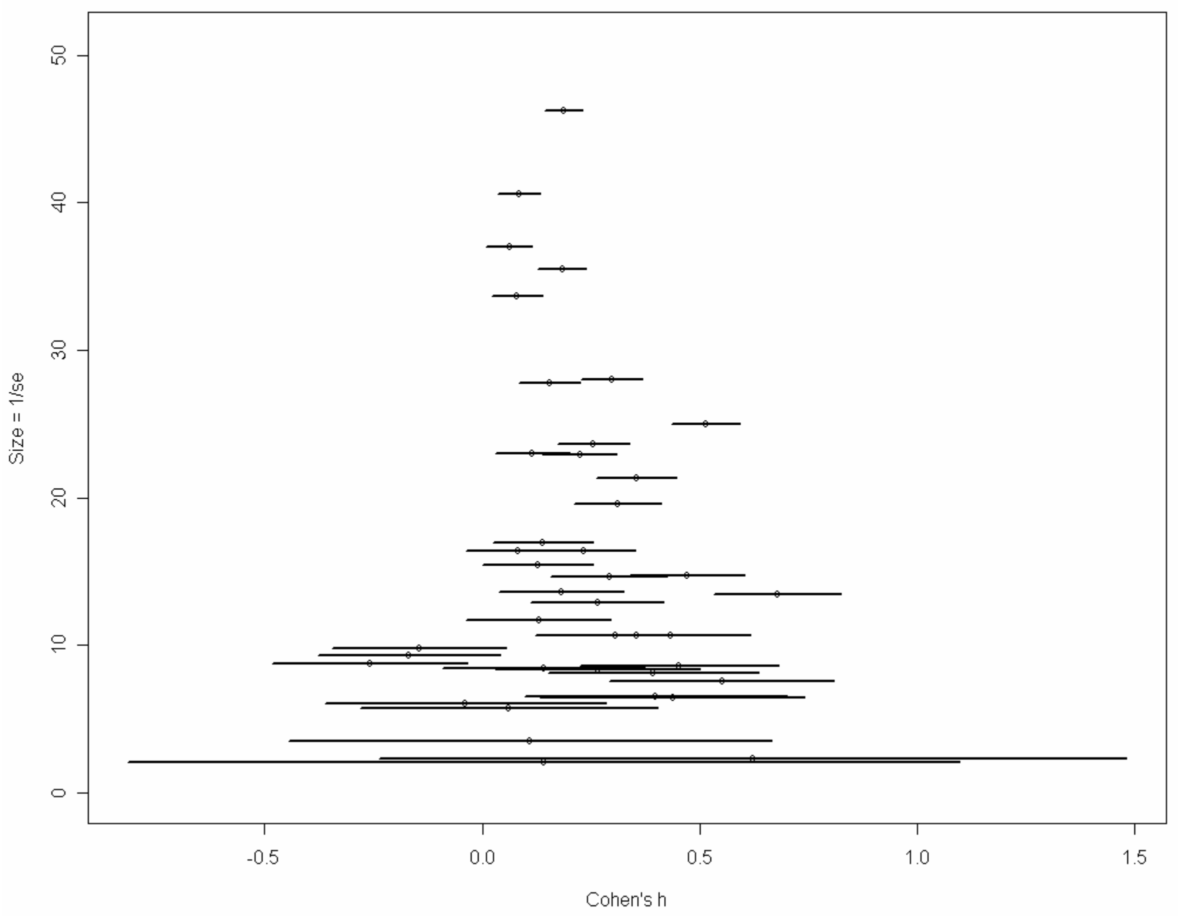


Because the significant Q statistic (Q = 161,15, df=21, p < .001) indicates that the variability across effect sizes is greater than expected from sampling error alone, a random-effects model was also estimated. For the new 22 studies the estimated random-effects mean ES is .22 (z-test value = 5.05; 95% CI = 0.13 < μ < 0.30).

Bivariate fixed-effects meta-regression indicates that the mean ES calculated for the Cairns et al. 2004 data set of ‚normal' travel plans is indeed significantly lower than the ES calculated for the Cairns et al. 2002 ‚best practice' data set (B = -,0349 ; z-test value = -2.08; 95% CI = -0.07 < μ < -0.02). However, the difference is small. It explains about 2 % of the study level variance. Thus in the last step we pooled both data sets. For the total sample of 41 studies the estimated random-effects mean ES is .23 (z-test value = 9.08; 95% CI = 0.18 < μ < 0.28).

Another important strategy to test the generalisability of an ES estimate is to assess the potential influence of publication bias. For this purpose Figure 3 presents a funnel plot for the 41 studies analysed in the present paper.

Fig. 3: Funnel plot for the 41 studies analysed



As can be seen, the plot of sample size versus ES is quite consistent with the ‚funnel' shape, expected when no publication bias is present.

Discussion and conclusion

Despite its obvious inadequacy, the traditional narrative review is still the dominating research synthesis method in the transport policy domain. The aim of our paper was to demonstrate how the combination of the concepts systematic review and quantitative meta-analysis provides a more transparent and reliable research syntheses strategy for the transport policy domain.

Within this strategy the systematic review approach provides a methodologically sound solution for the fundamental problem of each research review how to assemble a base of research findings in as unbiased a way as possible. In its core the systematic review approach consists of a set of rules explicitly guiding the search, evaluation, and selection of study findings. The strategy to fulfil these rules should be documented a-priori in a review protocol.

The Cairns et al. (2002) review provides a good example of how to conduct a systematic review in the transport policy domain. However, in appraising the methodological quality of their data, the authors do not mention one important limitation of the used work travel plan evaluation data: All the findings are based on a weak quasi-experimental research design, namely a simple one-group pre-post test design. The internal validity of this design that is its ability to provide strong causal evidence for the effectiveness of work travel plans is threatened by a variety of factors, most important its inability to check for the influence of competing alternative explanations. It would have been the reviewer's duty to remind the readers of their review of the limited ability such a weak evaluation design has to provide strong causal inferences. Thus it remains open how much of the observed car reduction can actually be attributed to the causal impact of work travel plans.

The second element of the proposed research synthesis strategy – meta-analysis – provides a quantitative way for synthesising central trends of the assembled study findings. In principle narrative synthesis as well as quantitative meta-analysis can be used for this purpose. Which option is most adequate depends on the kind of review question and the kind of data available. In their review Cairns et al. concentrate on answering three main questions: What is the average car reduction observed across the 21 'best practice' work travel plan case studies (mean effect size), are there differences in the success of the evaluated 21 work travel plans to reduce work related car use (heterogeneity of effect size distribution), and what are important factors to explain these success differences (moderators of the effect size distribution)? Cairns et al. decided to use a mainly narrative data synthesis approach for answering these review questions, however, with a quantitative element like calculating the median of the reported effect sizes across all studies as well as for specific subgroups.

From our view one can question whether this mixture of qualitative and quantitative data synthesis elements provides a methodologically adequate synthesis strategy. For example Cairns et al. use the unweighted mean/median for estimating mean effect sizes. As discussed above in the context of research synthesis these statistics are not adequate because they do not take sample size that is the precision of the single study findings into account. Cairns et al. also provide no statistical test of the calculated mean effect sizes. Thus it remains open, whether these estimates represent a true intervention effect or whether they only reflect random fluctuation.

Furthermore, Cairns et al. provide no formal analysis of the degree and nature of heterogeneity in their data. This is unfortunate because the assumed heterogeneity in work travel plans car use reduction effectiveness is the starting point of their central review question: What factors can explain this heterogeneity? This is where the time- and labour intense coding of study descriptors conducted by Cairns et al. comes in. In the central chapter of their review they try to narratively describe and summarise how a variety of organisational, site, and work travel plan descriptors are associated with the ES distribution. At this point the limitations of the narrative data synthesis approach becomes evident: It may be possible to narratively describe and summarise bi- or trivariate association, however, Cairns et al. are bound to fail when they try to narratively analyse the multivariate association between the ES distribution and the multitude of collected study descriptors. Over 40 pages they present a flood of uni- and bivariate tables, which are not only difficult to read and comprehend, but often do not make clear, how the authors arrived at their conclusions. For example the empirical evidence for their practically important conclusion that organisational or site characteristics have a neglectable impact on the ES distribution remains unclear. Our impression is that this conclusion is quite subjective, guided more by implicit a-prior assumptions than the data. We often have the impression that Cairns et al. use the data mainly as ‚empirical' support for the preconceived idea that the implementation of work travel plan is always an effective measure, no matter what kind of organisation or site, and that addressing parking is the key success factor.

Compared with this less systematic, subjective narrative synthesis approach meta-analysis provides a clear, transparent and replicable strategy for analysing central trends of quantitative study findings. Graphical techniques like the funnel plot provide a possibility to check whether a review's data base is unbiased. In the present case the arcsine transformation provides a method to calculate effect sizes for differences between proportions whose statistical power is independent of the location of the respective proportion between 0 and 1. The inverse of the effect size variance provides a weight which directly reflects the precision of the effect size found in each single case study.

In the present context, however, the greatest advantage we see in using meta-analysis is the possibility it provides for systematically analysing the degree and sources of heterogeneity. The Q-statistic provides a formal test of heterogeneity. In the present meta-analysis even after the exclusion of two out-liners the value of the Q-statistic is significant. This result indicates that the differences in the effect sizes reported by the case studies cannot be explained by sampling error alone. In such a case the application of a fixed-effects model leads to an inflation of the estimated z-values and CIs, which increases the probability of erroneously assuming a significant mean ES, which in reality only reflects random fluctuation. By explicitly adding an additional study level variance component to the study weights a random effects model enables the investigator in this case to make a more adequate summary statement about the range of likely effects. For the present set of work travel plan case studies we found a significant random-effects mean ES of .23.

In the case of a heterogeneous set of study finding the mean ES is of limited practical significance. Therefore the most important thing to do in this case is to explore systematic heterogeneity sources. The second advantage of the meta-analytical approach is that with meta-regression it provides an adequate statistical tool for such analyses. In the present case, meta-regression results lead to conclusions which differ considerably from the narrative conclusions drawn by Cairns et al. Whereas these authors come to the conclusion that organisational and site characteristics are negligible, our meta-regression results indicate a strong impact of these factors on the observed ES differences: Work travel plans implemented in public organisations, in organisations with a mainly female staff, and on sites with poor/average cycling access report the strongest ES's. Compared with the emphasis Cairns et al. put on parking as a central success factor, the meta-regression results provide only modest evidence for the importance of this factor. Organisations which give their staff financial incentives for giving up parking, report on average higher ES. Furthermore, our results indicate that characteristics of the monitoring process itself may have an impact on the found ES. A longer duration of the monitoring process seems to be associated with reporting stronger ES. In our analysis these five study descriptors explain 78 % of the observed total study-level heterogeneity.

The results of the mixed-effects meta-regression are of greater practical significance than the mean ES. Practitioners, who are confronted with the question what car reduction effect they should expect when implementing a work travel plan in their specific organisation and site, are better off if they use the estimated regression coefficients for calculating an organisation and site specific ES estimate than using the mean ES averaged over all studies. The estimated unstandardized regression coefficients on the study descriptors variables (B-weights) represent the multiplier that weights each value on a study descriptor. Thus, if the intention is to implement a work travel plan with incentives for not parking in a public organisation with difficult / average cycling access and mainly female staff bias, after a 36 months survey period a ES of .53 can be expected. However, if a private organisation on a site with difficult / average cycling access without female bias in staff intends to implement a work travel plan without providing incentives for not parking, an ES of .19 should be expected.

Reference

Anton, E. M.; Lau, J.; Kupelnick, B.; Chalmers, T. C. (1992). A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. JAMA, 268, 240-248.

Chalmers I, Altman D (eds) (1995). Systematic reviews. London: BMJ Publishing Group.

Chalmers TC, Frank CS, Reitman D. (1990). Minimizing the three stages of publication bias. JAMA, 263, 1392–5.

Clarke, M., & Oxman, A. D. (Eds.). (2000). Cochrane reviewers handbook: Version 4.1. In: ReviewManager (RevMan) (Computer program). Version 4.1. Oxford, England: The

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.

Cooper, H., & Hedges, L. V., (eds.) (1994). The handbook of research synthesis. New York: Russell Sage Foundation.

Götzsche, P. C. (1987). Reference bias in reports of drug trials. BMJ, 295, 654–656.

Hedges, L. V. (1982). Estimating effect size from a series of independent experiments. Psychological Bulletin, 82, 490-499.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando, FL: Academic Press.

Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. Psychological Methods, 1, 227 –235.

Lau, J., Antman, E. M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F., Chalmers, T. C. (1992). Cumulative meta-analysis of therapeutic trials for myocardial infarction. N Engl J Med, 327, 248-254.

Lau, J., Loanidis, J. P. A., & Schmid, C. H. (1997). Quantitative Synthesis in Systematic Reviews. Annals of Internal Medicine, 127, 820-826.

Light, R. J., & Pillemer, D. B. (1984). Summing up: the science of reviewing research. Cambridge, MA: Harvard University Press.

Ligtermoet, D. (1998). Zeven jaar vervoermanagement: synthese van ervaringen Report to Adviesdienst Verkeer en Vervoer. Report to the Netherlands Ministry of Transport, The Hague.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatments: Confirmation from meta-analysis. American Psychologist, 48, 1181-1209.

Lipsey, M. W.; & Wilson, D. (2001). Practical Meta-Analysis. Thousand Oaks, CA: Sage.

Mosteller, F., & Colditz, G. A. (1996). Understanding research synthesis (meta-analyis). Annual Review of Public Health, 17, 1-23.

Mulrow, C.D. (1987). The medical review article: state of the science. Ann Intern Med 1987; 106, 485–488.

NHS Centre for Reviews and Dissemination (CRD) (2001). Undertaking systematic reviews of research on effectiveness: CRD’s guidance for those carrying out or commissioning reviews. NHS Centre for Reviews and Dissemination, York University, CRD Report 4: second edition.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. Psychological Methods, 3, 354-379.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (eds.), The handbook of research synthesis (pp. 301-321). New York: Russell Sage Foundation.

Ravnskov, U. (1992). Cholesterol lowering trials in coronary heart disease: frequency of citation and outcome. BMJ, 305, 15–19.

Rosenthal, R., & Rubin, D. B. (1982). Comparing effect sizes in independent studies. Psychological Bulletin, 92, 500-504.

Schreffler, E. N., & Organizational Coaching (Rien J.J. van der Knaap and Pieter A.G. van den Ign) (1996). Effective TDM at worksites in the Netherlands and the US.

Shoup, D. (1997). Evaluating the effects of cashing out employer-paid parking: eight case studies Transport Policy 4(4) 201-216

Spector TD, Thompson SG. (1991). The potential and limitations of meta-analysis. J Epidemiol Community Health 1991; 45, 89-92.

Steer Davies Gleave (2001) The Take-up and Effectiveness of Travel Plans and Travel Awareness Campaigns. Report to the Department of the Environment, Transport and the Regions,

TCRP (1994): Cost Effectiveness of TDM programs: Working paper #2, COMSIS Corporation, prepared for Transit Co-operative Research Program.

Teagarden, J. R. (1989). Meta-analysis: whither narrative review? Pharmacotherapy, 9, 274–84.

Touwen, M. (1999) Travel planning in the Randstad: an evaluation based on ReMOVE. Report to Netherlands Ministry of Transport, The Hague.

Appendix I - Study References

Dataset Cairns et al.

Cairns, Sally (2002). **Making travel plans work: Research report.** Department for Transportation, p. 35 – 40.

Dataset Additional Studies

Cairns, S.; Sloman, L.; Newson, C; ANable, J.; Kirkbride, A. & Goodwin, P. (2004). **'Smarter Choices – Changing the Way We Travel'.** Chapter 3: **Workplace travel plans**, p. 49.

Anable, J.; Kirkbride, A.; Sloman, L.; Newson, C.; Cairns, S. & Goodwin, P. (2004). **Smarter Choices – Changing the Way We Travel. Case study reports**. **Birmingham City Council. Workplace Travel Plans.** Department for Transport, p. 10.

Anable, J.; Kirkbride, A.; Sloman, L.; Newson, C.; Cairns, S. & Goodwin, P. (2004). **Smarter Choices – Changing the Way We Travel. Case study reports**. **Bristol City Council. Workplace Travel Plans.** Department for Transport, p.52.

Anable, J.; Kirkbride, A.; Sloman, L.; Newson, C.; Cairns, S. & Goodwin, P. (2004). **Smarter Choices – Changing the Way We Travel. Case study reports**. **Buckinghamshire County Council. Workplace Travel Plans.** Department for Transport., p. 102, 103.

Eingereicht bei der Zeitschrift *Journal of Environmentally Psychology*

5) Twenty years after Hines, Hungerford and Tomera: A new meta-analysis of determinants of pro-environmental behaviour

Sebastian Bamberg & Guido Möser

It is now twenty years ago that Hines, Hungerford and Tomera (1986/87) published their first meta-analysis of research on responsible environmental behaviour. Goal of their meta-analysis was not only to identify variables which are strongly associated with pro-environmental behaviour, but also to determine the relative strengths of the relationships between each of these variables and pro-environmental behaviour. For answering these questions, Hines at al. used the meta-analysis methodology. Meta-analysis, a term coined by Glass (1976), is "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (p. 3). A literature search by Hines et al. resulted in a list of 128 primary studies which assessed variables in association with pro-environmental behaviour and which reported the data needed for meta-analytical purpose. The studies provide information concerning the relation between pro-environmental behaviour and three major variable categories: demographic variables, knowledge related variables, and psycho-social variables.

In the present context of special interest are the meta-analytical results concerning the bivariate relation of the psycho-social variables attitude, locus of control / self-efficacy, moral responsibility, behavioural intention and pro-environmental behaviour: Based on 9 studies Hines et al. found a mean correlation between attitude toward pro-environmental behaviour and actual pro-environmental behaviour of $r = .38$, between locus of control / self-efficacy and pro-environmental behaviour of $r = .37$ (15 studies); between moral obligation and pro-environmental behaviour of $r = .33$ (6 studies), and intention and pro-environmental behaviour of .49 (6 studies).

Before the background of their meta-analytical results, Hines et al. (1986/87) proposed a model of environmental behaviour. In this model they view intention to act and objective situational factors as direct determinants of pro-environmental behaviour. Intention itself is viewed as summarising the interplay of cognitive (action skills, knowledge of action strategies and issues) as well as personality variables (attitudes, locus of control, and personal responsibility).

In the following decade, the meta-analysis conducted by Hines et al. exerted a strong impact on further research in the field of predicting pro-environmental behaviour. By using modern statistical methods for synthesising results from different primary studies it provided strong empirical evidence for the utility of psycho-social variables as predictors of pro-environmental behaviour. This finding encouraged many researchers to continue research targeting psycho-social determinants of pro-environmental behaviours.

The present research

It is astonishing that despite the impact of this first meta-analysis, to our best knowledge no further meta-analyses of research on pro-environmental behaviour have been published since 1986. Lack of new research can not be the reason for this gap. Since the work of Hines et al. a steady stream of primary studies analysing determinants of pro-environmental behaviour has been published. A meta-analysis of these newer studies is urgently needed, not only because the length of time that has passed since the appearance of the Hines et al. meta-analysis but also because the results of this meta-analysis are based on a relatively small number of primary studies.

Thus the first goal of the present paper is to assemble a body of newer studies for an independent replication of the Hines et al. meta-analytical results. The second goal of the present meta-analysis directly ties up where the Hines et al. paper ends: We want to perform a meta-analytical test of a theoretical model of causal determinants of pro-environmental behaviour. Such a theory-driven meta-analysis more adequately reflects one main trend of environmental psychological research during the last decade: The use of theoretical models for modelling the interplay of knowledge, behavioural constraints/opportunities as well as personal values and motives in influencing the decision to behave pro-environmentally (e.g. Bamberg & Schmidt, 2003, Taylor & Todd, 1995).

This greater emphasis on modelling and testing construct relationships corresponds with a similar development in meta-analytical methodology: Apart from the traditional univariate effect sizes, researchers have started to emphasise synthesising multivariate effect sizes, especially correlation matrices, because of the increasing complexity of the research questions (e.g. Becker & Schram, 1994; Cheung, 2000; Hedges & Olkin, 1985). Just inspecting a matrix of synthesised correlations, however, may not be very informative in understanding the underlying relationships among the variables. As a result, researchers have become interested in testing theoretical models with structural equation modelling (SEM) based on the meta-analytically derived correlation matrix (Becker, 2000; Becker & Schram, 1994; Shadish, 1996; Viswesvaran & Ones, 1995). In the present paper we want to use this methodological integration of meta-analysis and SEM, which is called meta-analytic SEM (MASEM), to test a theoretical model of determinants of pro-environmental behaviour.

After presenting the integrative model of determinants of pro-environmental behaviour use as theoretical framework for our meta-analysis, the second section of the paper describes the MASEM methodology in more detail. The third section describes the search strategy as well as inclusion criteria used for assembling the body of meta-analysed studies. The main section reports the results of a two-stage procedure used for conducting the MASEM. The last section critically evaluates our results from a theoretical as well as methodological point of view.

The theoretical model

Pro-environmental behaviour is probably best viewed as a mixture of self-interest (e.g., to pursue a strategy that minimises one's own health risk) and of concern for other people, the next generation, other species, or whole eco-systems (e.g., preventing air pollution that may cause risks for others' health and/or the global climate). This mixture of self-interest and pro-social motives is also reflected by the theoretical models most frequently applied for explaining pro-environmental behaviour: Researchers who view environmental behaviour primarily as pro-socially motivated often use the Norm-Activation Model (NAM, Schwartz, 1977) as theoretical framework, whereas researchers who view self-interest as the more important motive often rely on rational choice models like the Theory of Planned Behaviour (Ajzen, 1991).

The basic premise of the NAM is that moral or personal norms are direct determinants of pro-social behaviour. Schwartz (1977) conceived moral norms as feelings of strong moral obligations that people experienced for themselves to engage in pro-social behaviour. In line with this model several primary studies provide evidence that moral norms contribute to an explanation of pro-environmental behaviours like energy conservation (Black, Stern & Elworth, 1985), recycling (Guagano, Stern & Diez, 1995), travel mode choice (Hunecke, Blöhbaum, Matthies, & Höger, 2001), and pro-environmental buying (Thøgersen, 1999). As reported above, Hines et al (1986) found a mean correlation of r = .33 between a feeling of moral obligation to preserve the environment and pro-environmental behaviour.

The formation as well as activation of a moral norm is probably based on the interplay of cognitive, emotional, and social factors (e.g., Bierhoff, 2002): In the field of pro-environmental behaviour, the awareness of and knowledge about environmental problems are important cognitive preconditions for developing moral norms. Causal attribution is probably a second important cognitive process contributing to the development of moral norms. According to Weiner (2000) attributed internal responsibility for a harmful behaviour triggers affective reactions, namely guilt feelings. Guilt is defined as a 'painful feeling of regret that is aroused when the actor actually causes, anticipates causing, or is associated with an aversive event." (Ferguson & Stegge, 1998, p.20). Guilt is an important pro-social emotion because it results in a felt obligation (moral norm) to compensate for the caused damage (Baumeister, 1998). Feelings of guilt are also closely related with social norms. A perceived mismatch between own behaviour and social norms, leads to feelings of guilt (Baumeister, 1998). Besides their impact on feelings of guilt, social norms also directly contribute to the development of moral norms. They deliver the standards what behaviour a social reference group views in a specific context as appropriate that is what is right or wrong. If an individual internalises these standards they provide the content of her/his personal moral norms.

The second theoretical framework, Ajzen's (1991) TPB, is based on a more hedonistic model of human beings. It assumes that people are motivated to avoid punishments and to seek rewards. According to this model decision making is guided by a rational evaluation of behavioural consequences. The sum of perceived positive and negative consequences determine the global attitude toward a behavioural option. Attitudes do not directly determine behaviours but only indirectly via behavioural intentions. The TPB also stresses the importance of situational constraints. People do not only calculate the personal benefit of a behavioural option, they also estimate their ability to perform this option that is their perceived behavioural control over it. Social norms are viewed as a third factor influencing decision making. In the TPB framework social norms are primarily conceptualised as social pressure that is the expectations of significant reference persons to perform or not perform a behavioural option. Fear of social exclusion is viewed as the primary motive why people tend to fulfil these expectations. As attitudes, perceived behavioural control and social norms should determine behaviours not directly but only indirectly via their impact on intentions. It is further assumed that when perceived behavioural control is a reliable predictor of objective behavioural control, it also predicts behaviour directly.

In line with the introductory statement that pro-environmental behaviour is best viewed as a mixture of self-interest and pro-social motives, it is suggestive to combine both theoretical frameworks. Thus, various researchers have proposed to introduce besides attitude, social norm and perceived behavioural control moral norm as additional independent predictor of intention (e.g. Manstead, 2000). In their analysis of the determinants of five specific pro-environmental intentions, Harland, Staats & Wilke (1999) found that the inclusion of moral norm raised the proportion of explained variance of people's intention by 1% to 10%.

Figure 1 presents our proposed integrative model graphically. As can be seen, for a more balanced representation of self-interest and pro-social motives, moral norm instead of social norm is conceptualised as third independent determinant of intention. This change is supported by reviews of TPB applications (e.g. Ajzen, 1991, Armitage & Conner, 2001) which indicate that often social norm exerts no direct effect on intention after controlling for the effects of attitude and PBC. Our integrative model ascribes social norms a more indirect role. In line with Sherif's (1936) classical study on the informational influence of social norms it is assumed that frequently people follow social norms not because they fear social pressure, but because they use social norms as information what behaviour is appropriate. Thus social norms may not only provide information whether a specific behavioural option is morally right or wrong but also whether it is beneficial or easy to perform. It is further assumed that knowledge concerning environmental problems and their solutions is also used for evaluating positive / negative consequences of a behavioural option (attitude) as well as the estimation how easy or difficult the performance of this option would be (perceived behavioural control).

Method

Data collection

Because of our goal to conduct a meta-analytical test of the above proposed integrated model of pro-environmental behaviour, we primarily search for studies testing the NAM, TPB or related models in the domain of pro-environmental behaviour. Furthermore, we focus our search on papers published since 1995 in peer-reviewed journals. As search keywords we used: recycling; energy saving; waste reduction; travel mode choice; green consumerism; reduce meat consumption; ecological behaviour; ecological behaviour & moral norm; ecological behaviour & personal norm; ecological behaviour & morality; norm activation model & ecological behaviour; theory of planned behaviour; environmental consumer behaviour; ecological consumer behaviour; green consumer behaviour; pro-environmental behaviour; environmental protection behaviour theory; pro-environmental behaviour; intention & ecological behaviour.
Besides using the internet search machine Google, the databases PsycInfo and Dissertation Abstracts, we inspected the content tables of the following journals since 1995: Advances in Consumer Research, American Behavioural Scientist, Basic and Applied Social Psychology, British Journal of Social Psychology, Communication Studies, Environment and Behaviour, Environmental Education Research, European Journal of Marketing, European Journal of Social Psychology, Gruppendynamik, Japanese Journal of Social Psychology, Journal of Applied Psychology, Journal of Applied Social Psychology, Journal of Business Research, Journal of Consumer Marketing, Journal of Cross-Cultural Psychology, Journal of Economic Psychology, Journal of Environmental Planning and Management, Journal of Environmental Psychology, Journal of Personality and

Social Psychology, Journal of Public Policy & Marketing, Journal of Social Issues, Journal of Socio-Economics, Leadership Quarterly, Marketing Theory, Perceptual and Motor Skills, Personality and Individual Differences, Population and Environment, Professional Geographer, Psychological Reports, Psychology & Marketing, Rationality and Society, Scandinavian Journal of Psychology, Social Science Research, Zeitschrift für Sozialpsychologie.

Inclusion criteria

The first step of our literature search resulted in a list of over 100 empirical papers matching our keywords. In a second step we checked whether the papers report a matrix of bivariate Pearson correlations as well as sample size. This information is needed for conducting the meta-analysis. During this step we lost a number of interesting studies because they only report multivariate results obtained from regression or SEM analyses without documenting the respective bivariate correlations. In the third step we read the parts of the remaining studies in which the analysed constructs as well as their operationalisation were described. Papers which do not analyse at least two of the constructs included in our theoretical model or where the definitions and/or measures do not fit our understanding of these constructs were also excluded. This procedure results in a list of 22 studies which reported correlation matrices of 29 independent samples. These 29 correlation matrices provide the input for the following MASEM. In the reference section the included studies are marked by an asterisk.

Conducting a MASEM

Researchers typically conduct MASEM by means of a two-stage procedure (Viswesvaran & Ones, 1995). In the first stage, the correlation matrices obtained from the primary studies have to be pooled and tested for homogeneity. After statistically testing the homogeneity of the pooled correlation matrices, one has to decide whether a fixed-effects or random-effects model is more appropriate for estimating the 'true' pooled correlation matrix.

In the literature one can find two statistical approaches for calculating the 'true' pooled correlation coefficients: the Hedges and Olkin (1985) method and the Hunter and Schmidt (1990) method. In the present paper the Hedges and Olkin method is used. In this method the correlations from each primary study are first converted into a standard normal metric by using Fisher's r-to-Z transformation. The transformed primary correlations are then used to calculate an initial pooled mean correlation, in which each primary correlation is weighted by the inverse of the within-study variance of the study from which it came (see Hegdes & Olkin, 1985, p. 231). Then the Q-test statistic of homogeneity (see Hegdes & Olkin, 1985, p. 231) is calculated for these pooled weighted correlations. Because the Q-test was developed for univariate-z values, Cheung (2000) recommends using a Bonferroni-adjusted at-least-one approach for testing the equality of elements across correlation matrices. This means that the hypotheses of homogeneity of correlation matrices will be rejected if at least one of the correlation coefficients is heterogeneous across studies.

When the heterogeneity test is insignificant, the fixed-effects model is appropriate to calculate the 'true' pooled correlation matrix. However, when the heterogeneity test is significant, the application of a fixed-effects model is inappropriate (e.g., Hunter & Schmidt, 2000). In this case potential moderators may be used to explain variability across studies or a random-effects model may be used to average the correlations. To calculate the random-effects 'true' correlation matrix, the weights for pooling the correlations for the primary studies used a variance component that incorporates both between-studies variance and within-studies variance. There are different methods for estimating the between-studies variance (e.g., Hedges & Vevea, 1998). In the present paper a non-iterative method based on the results of the Q-statistic (Hedges & Vevea, 1998, Equation 10) is applied. The between-studies variance estimate is then added to the within-study variance component. The pooled 'true' random-effects correlation matrix is recalculated with these new weights and converted back to the r metric.

One problem in synthesising correlation matrices is that studies may involve different numbers of variables, because different researchers conduct research independently. There are two common methods to handle this issue (Viswesvaran & Ones, 1995). The first method is to include only studies that contain all model variables, that is to use listwise deletion. The second method, which is the dominant method for applied researchers, is to estimate the elements of the pooled correlation matrix based on different numbers of studies that is to use pairwise deletion.

In the second stage, the pooled 'true' correlation matrix is used as input for conducting a SEM path analysis. When fitting the SEM model, a central problem one has to solve is deciding on an appropriate sample size. Because the pooled correlation matrix is usually formed by averaging across different studies based on pairwise deletion, researchers have to decide on the appropriate sample size for the analysis in SEM. Researchers have used a variety of sample sizes such as the arithmetic mean, the harmonic mean, the median or the total of the sample sizes based on the synthesised correlation coefficients.

Results

Table 1 presents the information (number of available independent primary bivarate correlation coefficients and the pooled total sample size on which these coefficients are based) extracted from the 22 studies included in our meta-analysis.

Table 1 impressively demonstrates the above mentioned missing values problem one is typically confronted with when conducting a MASEM analysis. Because our proposed integrated theoretical model contains 9 variables, 36 pooled mean correlations are necessary for conducting a MASEM test of the above proposed integrative model. As can be seen from Table 1, the information available from the 22 primary studies included in our meta-analysis varies considerably over these 36 cells: Whereas 18 independent primary correlations are available for calculating the pooled average bivariate correlations of social norm, attitude, and PBC, the calculation of pooled mean correlations for the construct attribution with social norm, guilt, attitude, and PBC is impossible, because only one correlation coefficient is available for each of these associations. The great differences in the information available are also reflected in the pooled total sample sizes on which the reported correlation coefficients are based: The pooled total sample sizes vary from n = 5.822 for the correlation of attitude and PBC to n = 175 for the correlation of attribution and social norm.

Table 1: Total Sample Size (Upper Row) And Number of Independent Correlation Matrices (Lower Row) Obtained For Each Construct

Construct	1.	2.	3.	4.	5.	6.	7.	8.	9.
1. Problem	6571								
	(12)								
2. Attribution	1196	1639							
	(4)	(5)							
3. Social Norm	1831	175	6061						
	(4)	(1)	(21)						
4. Guilt	1233	443	1233	1676					
	(2)	(1)	(2)	(3)					
5. PBC	1558	175	5229	1233	6766				
	(4)	(1)	(18)	(2)	(23)				
6. Attitude	1408	175	5077	1233	5822	6430			
	(3)	(1)	(18)	(2)	(18)	(21)			
7. Moral Norm	5280	1016	3015	1233	3165	3015	6570		
	(8)	(3)	(9)	(2)	(10)	(9)	(13)		
8. Intention	4030	1142	5242	1676	5354	5242	5006	8515	
	(7)	(3)	(17)	(3)	(18)	(17)	(11)	(24)	
9. Behaviour	4897	1115	3789	1676	4494	4158	3289	4252	8356
	(11)	(4)	(11)	(3)	(13)	(11)	(6)	(12)	(22)

Thus, Table 1 indicates a first future research task: Until now there is obviously very little systematic research analysing the contribution of internally attributed responsibility and feelings of guilt to the development of pro-environmental moral norms. Simultaneously the high rate of missing values renders the listwise deletion strategy impossible for producing the pooled correlation matrix necessary for conducting the planned MASEM of our integrated theoretical model: In the total pool of 22 studies there is not one study assessing all nine model variables. Thus we have to use the pairwise deletion strategy to estimate the elements of the pooled correlation matrix.

The lower triangular matrix presented in Table 2 reports the respective pooled correlation coefficients resulting from the pairwise deletion strategy under the fixed-effects assumption. During pooling the fixed-effects model uses the inverse of the within-study variance as weights for the primary correlation coefficients. In the cases where only one primary correlation is available, these primary correlations were directly inserted into the matrix.

For testing the homogeneity of the pooled correlation matrix we calculated for each pooled correlation coefficient the Q-statistic. Only for two (problem awareness and attitude; guilt and moral norm) of the 32 pooled correlations the Q-statistic is insignificant. For most of the remaining 30 correlations the significance value of the Q-statistic is below the critical value of $p = .0016$, which according to the Bonferroni-adjusted at-least-one approach (Cheung, 2000) indicates strong heterogeneity of the pooled correlation matrix.

As a consequence we recalculated the 32 pooled correlations under the obviously more appropriate random-effects assumption. Applying a random-effects model means that during the pooling process the inverse of the sum of the estimated within-study and between-study variance components is used as weights with which each primary correlation contributes to the average correlation. The upper part of Table 2 presents the pooled correlation matrix calculated under the random-effects assumption.

Table 2: Fisher's Z-Back-Transformed Pooled 'True' Correlation Matrix Under The Fixed-Effects (Lower Triangular Matrix) And Random-Effects Assumption (Upper Triangular Matrix)

Construct	1.	2.	3.	4.	5.	6.	7.	8.	9.
1. Problem	---	.38	.37	.54	.22	.37	.55	.33	.18
2. Attribution	.32	---	.36*	.45*	.18*	.36*	.41	.20	.18
3. Social Norm	.43	.36*	---	.46	.30	.42	.47	.40	.26
4. Guilt	.53	.45*	.49	---	.23	.50	.68	.31	.24
5. PBC	.24	.18*	.34	.29	---	.39	.39	.54	.37
6. Attitude	.37	.36*	.43	.33	.44	---	.57	.59	.38
7. Moral Norm	.53	.33	.53	.68	.44	.59	---	.53	.19
8. Intention	.30	.18	.42	.34	.62	.60	.53	---	.54
9. Behaviour	.22	.19	.28	.25	.40	.41	.37	.55	---

Note: * = no pooled correlation

This pooled random-effects correlation matrix provides the answer to our first research question: How similar are the pooled mean correlations found in our meta-analysis to those reported by the Hines et al. (1986/87)? Hines et al. reported a mean correlation of r = .37 for attitude and pro-environmental behavior, in our meta-analysis the respective random-effects mean correlation is r = .38. For self-efficacy / locus of control and pro-environmental behaviour Hines et al. reported a mean correlation of r = .37; we found a mean correlation of r = .37 for PBC and pro-environmental behaviour. Hines et al. reported a mean correlation of moral obligation and pro-environmental behaviour of r = .33; we found a mean correlation of r = .19. In the Hines et al. meta-analysis the mean correlation of intention and pro-environmental behaviour is r = .49; in our meta-analysis this mean correlation is r = .54.

Thus for three of the four associations between psycho-social variables and pro-environmental behaviour meta-analysed by Hines at al. our meta-analysis found very similar results. Only for the association of moral obligation and pro-environmental behaviour the results differ: In our meta-analysis the mean correlation of these two variables is lower.

In stage 2 of our analysis we used the pooled random-effects correlation matrix as input for a MASEM test of the structural relation of the nine variables postulated by our integrated theoretical model. The MASEM analysis was conducted with the program LISREL 8.54 (Jöreskog & Sörbom, 1996). For parameter estimation the maximum-likelihood procedure was used. As discussed above in the context of MASEM a severe disadvantage of the decision to use the pairwise deletion strategy for producing the pooled correlation matrix is the problem of deciding what the adequate sample size is. In our MASEM we use the harmonic mean of n = 1048 as sample size estimation. For assessing data-model fit the criteria recommended by Hu and Bentler (1999) are used. Their criteria include a comparative fit index (CFI) greater or equal to .96 with a standardised root-mean-square residual (SRMR) less than or equal to .10. An alternative criterion involved a root-mean-square error of approximation (RMSEA) less than .06 with a SRMR less than or equal to .10.

Figure 1: Results of the MASEM based on pooled random-effects correlations, PBC = perceived behavioral control, completely standardised path-coefficients

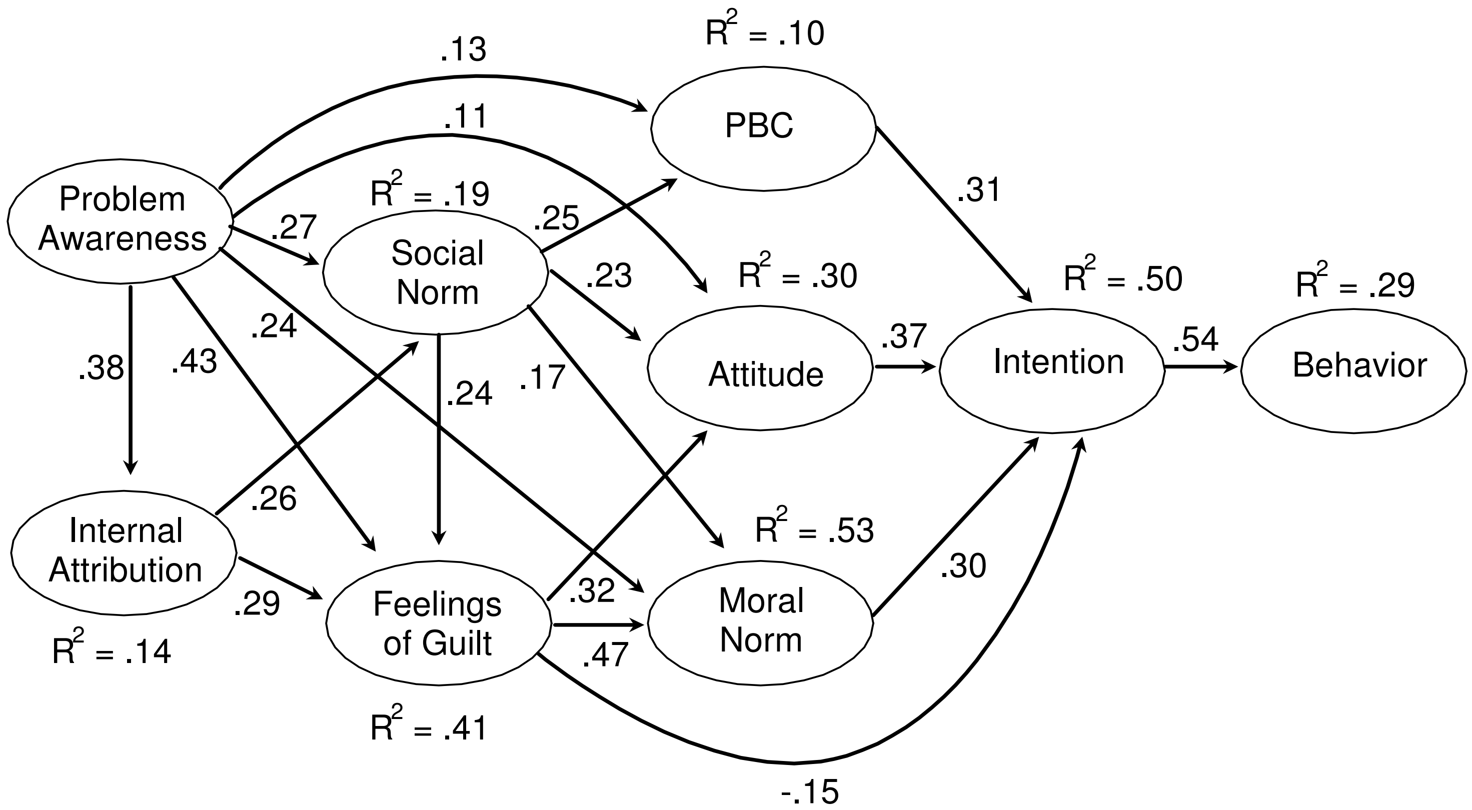


Figure 1 presents the results of the estimated MASEM (completely standardised structural coefficients and explained variances). As can be seen from Figure 2, our MASEM results confirm empirically the hypothesis derived from the integrated model that behavioural intention completely mediates the impact of all other psycho-social variables on pro-environmental behaviour. After controlling for the effect of intention PBC exerts no significant additional effect on behaviour. On average, intention explains 29 % of the variance of pro-environmental behaviour.

Also confirmed is the hypothesis that intention itself is determined by the independent impact of PBC, attitude, and moral norm. Besides these expected structural relations the LISREL modification index indicates that adding an additional direct path from guilt to intention would increase the model fit. However, the sign of this additional path is negative. Together these four constructs explain on average 50 % of variance of the intention construct.

As expected, guilt, social norm and problem awareness are substantive and positive predictors of the moral norm to behave in a pro-environmentally responsible way. Together the three constructs explain on average 53% of the variance of the moral norm construct. Also confirmed is the assumption that besides its direct as well as indirect impact (via guilt) on moral norm, social norm directly influences PBC and attitude. Inspection of the LISREL modification index indicates an unexpected direct association between feelings of guilt and attitude.

The MASEM analysis confirms the indirect, however important, role of problem awareness. This variable is directly associated with PBC, attitude, social norm, guilt and internal attributed responsibility. As expected, internal attributed responsibility is associated with social norms and feelings of guilt. However, one has to be very cautious in interpreting these two paths because they are not based on pooled correlations but only one primary correlation coefficient.

According to the criteria recommended by Hu and Bentler (1999) for assessing data-model fit, the correspondence between model implied and the actual pooled correlation matrix can be judged as acceptable (chi-square = 169.53; df = 14, $p < .001$; RMSEA = .10; CFI = .96; SRMR = .04).

Discussion and Conclusion

The goal of the present paper was a replication as well as extension of the meta-analysis on determinants of pro-environmental behaviours published 20 years ago by Hines et. al.. Extension means that we do not only report a series of meta-analytically derived mean bivariate correlations but use the pooled correlation matrices for a meta-analytical test of structural hypotheses derived from our integrative theoretical model. Statistically this test is conducted by using the pooled correlation matrix as SEM input. Such a theory-driven multivariate meta-analytical approach reflects more adequately the main trend of environmental psychology research during the last decade.

Consequently in the first section of our paper we present a theoretical model of the structural relations of nine variables frequently used as predictors of pro-environmental behaviour. This integrative model of determinants of pro-environmental behaviour serves not only as theoretical framework for deriving the research hypotheses but also guides our search of the research literature. This search results in a list of 22 primary studies providing a total of 29 independent correlation matrices with empirical information about the relationship of the nine constructs included in our integrated model. A first inspection of this pooled information shows great differences in the degree of research directed in the last decade towards these nine variables in the field of pro-environmental behaviour: Whereas a considerable number of studies have analysed the role of problem awareness / knowledge and the TPB variables social norm, attitudes, PBC and intention as behavioural predictors, the number of studies including moral norm as behavioural predictor is yet considerably lower. However, quasi no research has empirically addressed the role of responsibility attribution and emotions like feelings of guilt. Due to theoretical reasoning as well as the few data yet available underline the potential importance of these two constructs for understanding the development of pro-environmental moral norms, future research including these two constructs is urgently needed.

Homogeneity tests of the pooled correlation matrix indicate strong heterogeneity for 30 of the 32 calculated pooled correlations. Obviously the size of the correlations reported in the primary studies varies considerably across the found primary studies, probably depending on the kind of sample size or type of pro-environmental behaviour analysed. Due to this heterogeneity, from a statistical point of view the random-effects model is the more appropriate approach for calculating the 'true' pooled correlations. Comparing the pooled mean random-effects correlations obtained in our meta-analysis with those reported by Hines et al. shows for three of the four analysed psycho-social variables very similar results. The only exception is the mean correlation of moral norm and behaviour. Here the mean correlation found in our meta-analysis is considerably lower than that reported by Hines et al. (.19 vs. .33).

The results of the MASEM analysis are supportive for our postulated integrative theoretical model. The assumed mediating role of behavioural intention is confirmed. After controlling for the effect of intention none of the other variables have a significant additional effect on pro-environmental behaviour. On average, intention explains 29 % variance of self-reported pro-environmental behaviour. This result is identical with the result obtained in a huge meta-analyses of 180 empirical TPB applications across various behavioural domains conducted by Armitage and Conner (2001).

Our meta-analytical results also confirm the view of pro-environmental behaviour as a mixture of self-interest and pro-social motives. After controlling for the effect of perceived behavioural control, attitude (representing self-interest) as well as moral norm are significant independent predictors of intention. The average impact of the three predictors PBC (.31), attitude (.37), and moral norm (.30) is quite similar. This indicates that on average, the intention to perform a pro-environmental behavioural option can be viewed as a weighted balance of information concerning the three questions 'How many positive personal consequences would result from choosing this option compared to other options?', 'How difficult would the performance of this option be compared to other options?', and 'Are there reasons indicating a moral obligation for performing a specific option?'. Interesting is the unexpected additional direct, however, negative effect of feelings of guilt on intention. This result can be interpreted as a hint that without the action orientation provided by moral norms, feelings of guilt may activate avoiding coping strategy decreasing the motivation to behave pro-environmentally. Again these results underline the need of further research into the role of emotions in general as well as moral related emotion is special on the decision to choose pro-environmental options. Together the four constructs PBC, attitude, moral norm, and guilt explain 50 % of the variance of the intention construct, which is also in line with the finding of the meta-analysis conducted by Armitage and Conner (2001).

The MASEM results support our assumption that in the field of pro-environmental behaviour the formation as well as activation of a moral norm itself is determined by the interplay of cognitive, emotional, and social factors. Problem awareness, feelings of guilt, and social norms all significantly contribute to the prediction of moral norm. Together these three predictors explain 41 % variance of moral norm. Interesting is also the unexpected strong association of feelings of guilt and attitude. This result questions the conceptualisation of attitude as reflecting pure self-interest: When judging how personally beneficial the performance of a behavioural option would be, obviously people also seem to take the 'moral cost' of that option into account. The result, which is unfortunately only based on two primary correlations, that feelings of guilt are a strong predictor of moral norm as well as attitude, again underlines the potential significance of the construct feelings of guilt for future studies.

The MASEM results provide strong support for our view of social norms as more indirect determinants of intention. Besides their impact on moral norm, social norms are directly associated with the perceived degree of behavioural control as well as attitude. As discussed above, people may use social norms for judging how easy and beneficial the performance of a specific behavioural option would be. Awareness and knowledge about environmental problems seem to be a second important indirect determinant of pro-environmental behaviour. Awareness / knowledge is not only associated with the internal attribution of responsibility, social norms, and feelings of guilt, but also directly influences the degree of perceived behavioural control over as well as the attitude toward choosing a pro-environmental behaviour.

Our analysis also confirms the expected association between internal attribution, social norm, and feelings of guilt. However, these results should be interpreted very cautiously because they are not based on pooled but only on one primary correlation coefficient.

For an adequate evaluation the presented MASEM results one has to mention methodological problems with which this method is still struggling. At the moment these problems potentially pose a threat to the statistical validity of some results. A recently published simulation study (Cheung & Chan, 2005) reveals that the pairwise univariate-z method used in the present paper performs well in testing the homogeneity of correlation matrices and estimating the pooled correlation matrix. However, a major problem associated with this method concerns the determination of the sample size used for fitting the MASEM. All procedures momentarily used for this purpose like the arithmetic or harmonic mean are ad hoc solutions, not based on any statistical theory. Due to the Type I error of the chi-square test statistics, the goodness-of-fit indices, the statistical power and the standard errors of parameter estimates are all dependent on sample size used, using different sample sizes can result in different statistical inferences.

A second difficulty is that a MASEM based on the univariate-z method ignores the sampling variation across studies. After pooling the correlation matrices, researchers often use the pooled correlation matrix as the observed correlation matrix without considering the sampling variation across studies (e.g. Cheung & Chan, 2005). There are sampling variations in individual correlation matrices even when they share the same population correlation matrix. However, this sampling variation is not reflected when fitting SEM under the univariate approach in which their standard errors are ignored. Moreover, the covariation among the correlations is totally ignored in the univariate approach despite the fact that the correlations are often correlated to a certain degree (Olkin & Siotani, 1976).

The third difficulty is analysing a correlation matrix instead of a covariance matrix. Many researchers have warned about the problems of analysing the correlation matrix instead of the covariance matrix in primary research applications of SEM. Specifically, the chi-square statistics and the standard errors of parameter estimation may be incorrect.

For the first two difficulties more adequate solutions have been developed (e.g. Cheung & Chan, 2005), however, at the moment they can be only applied to homogeneous pooled correlation matrices.

Ignoring these methodological problems, what substantive conclusions can be drawn from our MASEM analysis? The positive conclusion is that in the last decade environmental psychology has made considerable progress in identifying central determinants of people's intention to choose the pro-environmental behavioural option. Research supports the conception of pro-environmental behaviour as a mixture of self-interest and pro-social motives. Thus, an adequate understanding of pro-environmental behaviour has to take both motives, self-interest as well as morality, into account. There is also progress in the understanding of the factors / processes contributing to the development as well as activation of pro-environmental moral norms. However, our analysis underlines that much more future research has to be directed toward these processes, especially those of 'moral' emotion like guilt or empathy.

As in other behavioural domains, the presently used models are less successful in explaining pro-environmental behaviour itself. The result that in the meta-analysed studies intention on average predicts only 30% variance of behaviour indicates that the processes contributing to the actual enactment of pro-environmental behavioural intention are not fully understood. At the moment the concept of implementation intention (e.g. Gollwitzer, 1999) and habit (Verplanken & Wood, 2006) are discussed as additional independent behavioural predictors.

Our meta-analysis summarises the results of correlational tests of theoretical frameworks used in the last decade for the understanding and prediction of pro-environmental behaviour. As we all know, such correlational tests do not allow causal inferences. Thus from our point of view the next decade of research on pro-environmental behaviour should concentrate more on the causal processes underlying the observed construct associations. For a better understanding of these causal processes more laboratory as well as true field experiments are needed, systematically manipulating the variables viewed as causally determining the motivation as well as actual performance of pro-environmental behaviour.

Finally we hope that 20 years will not pass again until the next meta-analysis on new research on determinants of pro-environmental behaviour will be published. From our point of view, a simple change in the editorial policy of journals publishing research on pro-environmental behaviour would considerably facilitate future meta-analyses: Editors should pay more attention that each paper reports the bivariate correlation matrix.

Reference

Ajzen, I, (1991). The theory of planned behavior. Organizational Behavior and Human Decision Processes, 50, 179-211.

*Allen, J. B., & Ferrand, J. L. (1999). Environmental locus of control, sympathy, and proenvironmental behavior: A test of Geller's Actively caring hypothesis. Environmental and Behavior, 31, 338-353.

Armitage, C., & Conner, M. (2001). Efficacy of the theory of planned behavior: a meta-analytic review. British Journal of Social Psychology, 40, 471-499.

*Bamberg, S., & Lüdemann, C. (1996). Eine Überprüfung der Theorie des geplanten Verhaltens in zwei Wahlsituationen mit dichotomen Handlungsalternativen: Rad vs. PKW und Container vs. Hausmüll. Zeitschrift für Sozialpsychologie, 32-46.

Bamberg, S., & Schmidt, P. (2003). Incentives, morality or habit? Predicting students car use for university routes with the models of Ajzen, Schwartz and Triandis. Environment and Behavior, 35, 264-285.

*Bamberg, S., Hunecke, M., Blöbaum, A. (2006). Moral Norm, Social Context and the Use of Public Transportation - Results of Two Field Studies. Unpublished manuscript.

Baumeister, R. F. (1998). The self. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), The handbook of social psychology (pp. 680-740). Boston: MacGraw-Hill.

Becker, B. J. (2000). Multivariate meta-analysis. In H. E. A. Tinsley & D. Brown (Eds.), Handbook of applied multivariate statistics and mathematical modeling (pp. 499-525). San Diego: Academic Press.

Becker, B. J., & Schram, C. M. (1994). Examining explanatory models through research synthesis. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 357-381). New York: Russell Sage Foundation.

Bierhoff, H.-W. (2002). Prosocial behavior. Psychology Press: Hove.

Black, J. S., Stern, P. C., & Elworth, J. T. (1985). Personal and contextual influences on household energy adaptations. Journal of Applied Psychology, 70, 3–21.

*Bratt, C. (1999). The impact of norms and assumed consequences on recycling behavior. Environment and behavior, 31, 630 – 656.

Cheung, M. W.-L., & Chan, W. (2005). Meta-Analytic structural equation modeling: A two-stage approach. Psychological Methods, 10, 40-64.

Cheung, S. F. (2000). Examining solutions to two practical issues in meta-analysis: Dependent correlations and missing data in correlation matrices. Unpublished doctoral dissertation, Chinese University of Hong Kong.

*Davies, J., Foxall, G. R., & Pallister, J. (2002). Beyond the intention-behaviour mythology: An integrated model of recycling. Marketing theory, 2, 29-113.

Ferguson, T. J., & Stegge, H. (1998). Measuring guilt in children. A rose by any other name has still thorns. In J. Bybee (Ed.), Guilt and children (pp. 19-74). San Diego, CA: Academic Press.

*Gärling, T., Fuji, S., Gärling, A., & Jakobsson, C. (2003). Moderating effects of social value orientation on determinants of proenvironmental behavior intention. Journal of Environmental Psychology, 23, 1-9.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5, 3-8.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. American Psychologist, 54, 493-503.

*Grob, A. (1995). A structural model of environmental attitudes and behavior. Journal of Environmental Psychology, 15, 209 – 220.

*Guagnano, G. A., Stern, P. C. & Dietz, T. (1995). Influences on Attitude-Behavior realtionships: A natural experiment with curbside recycling. Environment and Behavior, 27 , 699-718.

*Hamid, P.N. & Cheng, S.-T. (1995). Predicting antipopulational behavior. The role of molar behavorial intentions, past behavior, and locus of control. Environment and Behavior, 27, 679 – 698.

*Harland, P., Staats, H., & Wilke, H. A. M. (1999). Explaining proenvironmental intention and behavior by personal norms and the Theory of Planned Behavior. Journal of Applied Social Psychology, 29, 2505 – 2528.

*Heath, Y., & Gifford, R. (2002). Extending the Theory of Planned Behavior: Predicting the Use of Public Transportation. Journal of Applied Social Psychology, 32, 2154 – 2189.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando, FL: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. Psychological Methods, 3, 486-504.

Hines, J. M., Hungerford, H. R., & Tomera, A. N. (1986/87). Analysis and synthesis of research on responsible environmental behavior: A meta-analysis. Journal of Environmental Education, 18, 1-8.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6, 1-55.

Hunecke, Blöhbaum, Matthies & Höger, 2001

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

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative knowledge in psychology. International Journal of Selection and Assessment, 8, 275-293.

*Joireman, J.A., Lasane, T.P., Bennet, J., Richards, D. & Solaimani, S. (2001). Integrating social value orientation and the consideration of future consequences within the extended norm activation model of proenvironmental behavior. British Journal of Social Psychology, 40, 133 – 155.

Jöreskog, K.G., & Sörbom, D. (1996). LISREL 8: A user's reference guide, Chicago, IL: Scientific Software International.

*Kaiser, F. & Shimoda, T. (1999). Responsibility as a predictor of ecological behaviour. Journal of Environmental Psychology, 19, 243-253

*Kaiser, F. G. & Gutscher, H. (2003). The proposition of a general version of the theory of planned behavior: Predicting ecological behavior. Journal of applied social psychology, 33, 586 – 603.

*Kaiser, F., Wölfing, S. & Fuhrer, U. (1999). Environmental attitude and ecological behavior. Journal of Environmental Psychology, 19, 1-19.

*Knussen, C., Yule, F., MacKenzie, J. & Wells, M. (2004). An analysis of interventions to recycle household waste: The roles of past behaviour, perceived habit, and perceived lack of facilities. Journal of Environmental Psychology, 24, 237 – 246.

*Laudenslager, M. S., Holt, D. T., & Lofgren, S. T. (2004). Understanding Air Force Members' Intention to Participate in Pro-Environmental Behaviors: An Application of the Theory of Planned Behavior. Perceptual and Motor Skills, 98, 1162-1170.

*Lindsay, J. J., & Strathman, A. (1997). Predictors of Recycling Behavior: An Application of Modified Health Belief Model. Journal of Applied Social Psychology, 27, 1799-1823.

*Lynne, G. D., Casey, C. F., Hodges, A., & Rahmani, M. (1995). Conservation technology adoption decisions and the theory of planned behavior. Journal of Economic Psychology, 16, 581-598.

*Mannetti, L., Pierro, A. & Livi, S. (2004). Recycling: Planned and self-expressive behaviour. Journal of Environmental Psychology, 24, 227 – 236.

Manstead, A. S. R. (2000). The role of moral norm in the attitude-behavior relation. In D.-J.Terry & M. A. Hogg (Eds.), Attitude, behavior, and social context. The role of norms and group membership (pp. 11-30). Mahwah, New York: Lawrence Erlbaum.

*Meinhold, J. L., & Malkus, A. J. (2005). Adolescents environmental behaviors: Can knowledge, attitudes, and self-efficacy make a difference? Environemnt and Behavior, 37, 511-532.

*Mielke, R. (1985). Eine Untersuchung zum Umweltschutz-Verhalten (Wegwerf-Verhalten): Einstellung, Einstellungs-Verfügbarkeit und soziale Normen als Verhaltensprädiktoren. Zeitschrift für Sozialpsychologie, ?, 196-205.

*Nordlund, A.M. & Garvill, J. (2002). Value structures behind proenvironmental behavior. Environment and Behavior, 34, 740 – 756.

*Nordlund, A.M. & Garvill, J. (2003). Effects of values, problem awareness, and personal norm on willingness to reduce personal car use. Journal of Environmental Psychology, 23, 339 – 347.

Olkin, I., & Siotani, M. (1976). Asymtotic distribution of functions of a correlation matrix. In S. Ideka (Ed.), Essays in probability and statistics (pp. 235-251). Tokyo: Shinko Tsusho.

*Rise, Jo., Thompson, M., & Verplanken, B. (2003). Measuring implementation intentions inthe context of the theory of planned behaviour. Scandinavian Journal of Psychology, 44, 87-95.

Schwartz, S. H. (1977). Normative Influence on altruism. In L. Berkowitz (Ed.), Advances in Experimental Social Psychology, Vol. 10 (pp. 221-279). New York: Academic Press.

Shadish, W. R. (1996). Meta-analysis and the exploration of causal mediating processes: A primer of examples, methods, and issues. Psychological Methods, 1, 47-65.

Sherif, M. (1936). The psychology of social norms. New York: Harper

*Staats, H., Harland, P., & Wilke, H.A.M. (2004). Effecting Durable Change: A Team Approach to Improve Environmental Behavior in the Household. Environment and Behavior. 36, 341-367.

*Tanner, C. (1999). Constraints on environmental behaviour. Journal of Environmental Psychology, 19, 145 – 157.

Taylor, S., & Todd, P. (1995). An integrated model of waste management behavior: A test of household recycling and composting intentions. Environment and Behavior, 27, 603-630.

*Terry, D. J., Hogg, M. A., & White, K. M. (1999). The theory of planned behaviour: Self-identity, social identity and group norms. British Journal of Social Psychology, 38, 225-244.

Thøgersen, J. (1999). The ethical consumer. Moral norms and packaging choice. Journal of Consumer Policy, 22, 439–460.

Verplanken, B., & Wood, W. (2006). Interventions to break und create consumer habits. Journal of Public Policy & Marketing, 25, 1-26.

*Verplanken, B., Aarts, H., Van Knippenberg, A., & van Knippenberg, C. (1994). Attitude Versus General Habit: Antecendents of Travel Mode Choice. Journal of Applied Social Psychology, 24, 285-300.

Viswesvaran, C., & Ones, D. C. (1998). Theory testing: Combining psychometric meta-analysis and structural equation modeling. Personnel Psychology, 48, 865-885.

Weiner (2000). Intrapersonal and interpersonal theories of motivation from an attributional perspective. Educational Psychology Review, 12, 1-14.

6) Conclusion

The growing number of primary and secondary analyses for problems in the social sciences offers the practitioners a wide range of facts to build their decisions upon. However, these facts do not always rest upon unanimous results. The fact that results of studies in a similar area of research often vary greatly causes particular problems if the implementation of certain measures in a distinct field is intended. One goal of this work was to show the techniques and the methods of research synthesis, such as systematic review and meta-analysis. These allow for a summary of different empirical results from individual primary and secondary analyses and to apportion those results according to various causal criteria. Another goal was to review the applicability of different procedures of research synthesis for the measurement of the effectiveness of applied methods in the field of transport policy. We looked at three different studies in this area and were able to identify and to test the limits and the possibilities of methods of research synthesis.

This conclusion is arranged in three sections. In the first section, the results of the individual studies are summarised and reviewed critically. The second section is an analysis of the strengths and the weaknesses of the methods of research synthesis, which we used in this study. Finally, in the third section the applicability of our results for future studies are to be highlighted. Furthermore, we will look into the possibilities of amelioration of the applied methods, especially of meta-analysis.

The first study *'Meta-Analysis: An Alternative to narrative reviews for synthesising social science research?'* gave an overview of the most important methods of meta-analysis. A focus was on the practical application of these methods (i.e. a synthesis of empirical findings) in the fields of social and political sciences. With the techniques of meta-analysis it is possible to gather data from independent empirical studies (primary and secondary analysis) and to explain possible differences in the results of various studies to clarify why the results differ greatly among various studies. Meta-analyses are conducted almost the same way as any other classical research process. The aforementioned process includes collection of data as well as the interpretation of results, see for example Cooper and Hedges (1994) or Higgins and Green (2005). Methods have been designed which are now the standard for researchers when applying and publishing a meta-analysis. The standard methods serve as guidelines how meta-analyses are put into practice. Meta-analytical techniques of synthesis can be adopted succesfully in areas of research in which great numbers of primary and secondary analyses are implemented. A vast number of examples of meta-analyses can be found in the fields of medicine, psychology or education. Furthermore, the number of primary and secondary analyses which are conducted in other scientific fields is growing. This is a good basis for meta-analytical methods, see Wagner and Weiß (2005). Not only is it possible to review the published empirical data via meta-analysis and to explain differing results, it is also possible to draw exact and precise conclusions. The latter aspect of meta-analysis is highly relevant for practical application because a reliable and authoritative basis for decision making can be established.

The methods of meta-analysis are now highly sophisticated and applicable in many different fields. The most common procedures are *standardised mean difference, odds ratios and correlation coefficients*. Whereas the correlation coefficient is known in many different scientific fields as a measure of assessing a relationship between two variables, the method of *standardised mean difference* is first and foremost used for the synthesis of *Randomised Control Trials*. Medicine often uses *Odd Ratios* for risk evaluation. There is one thing all aforementioned methods have in common. The results from primary and secondary studies are not simply added up nor is an arithmetic mean calculated. The number of cases is included in the calculation of the overall effect. This way the effect which is calculated on the basis of all aggregated studies estimates the actual effect of the population. However, there is more to meta-analysis than just the objectivity and the verifiability of the results. Moreover, it allows for discrete and relatively uncomplicated ways of estimating and critically reviewing the results.

Weaknesses and strengths of meta-analysis will be discussed later. Nevertheless, I want to state the most common criticism of meta-analysis. Above all the lack of comparability between studies is criticised ('Apples and Oranges' Problem, see for example Hall, Tickle-Degnen, Rosenthal and Mosteller (1994)). This lack of comparability is caused by the differing definitions of independent variables, applied methods and characteristics of individual samples. Another cause of bias is the publication of results in selected media only (so called 'Publication Bias', see for example Hall, Tickle-Degnen, Rosenthal and Mosteller (1994)). However, there are techniques which, though they can not estimate these given weaknesses, at least have the capacity to control them to a certain extent.

Concerning the main interest of investigation of the second study, *'Are Soft Policy Measures Effective in Reducing Peoples' Car Use? A Meta-Analytical Review of Research Evidence*', various behaviour oriented soft policy measures with the aim of reducing private car use were chosen. On the one hand transport policy was chosen as area of investigation due to the problems in the field of traffic policy making. On the other hand the high number of yet conducted primary and secondary analysis served as a rich basis four our purposes. On the grounds of a meta-analytical procedure by Cohen (1988) we tried to measure effectiveness of previously implemented soft policy measure interventions in the field of traffic policy making.

Among others, the intensive use of cars in industrialised countries led to environmental pollution and other cumulative traffic problems like traffic jams. Inter alia, the stated problems were the reason for the introduction of a number of different measures which aimed to promote the intensive use of environment-friendly means of transportation. There were two central areas of intervention: Hard policy measures and soft policy measures. Part of the former are the improvement of infrastructure (creation of more bus stops, construction of bicycle paths, reduction of existing parking places, introduction of traffic-calmed areas) and the renewal and expansion of already existing public transportation. The latter is concerned with the development of measures which enhance the willingness of the individual to make more use of environment-friendly transportation. Very important strategies are educational advertising campaigns, attempts of stimulation at the workplace and in schools. Further tactics are monetary incentives on the one hand and penalties (disincentives) on the other. So far the main focus lay on the introduction of hard policy measures. Despite great capital investment, the main goal of habitual change from the usage of cars to environment-friendly transportation could not be reached.

Many studies were conducted which focused on finding the most appropriate and the most efficient measures among the choice of measures given above. The results were not very satisfying however. The accounts of a relation between interventions (especially in the area of hard policy measures) and the actual increase of the usage of environment-friendly transportation ranged from at best slightly positive to negative. These rather pessimistic results could be an explanation for the great number of studies which looked into the subject in recent years. Previously the main focus lay on the measures in the field of hard policy measures. This is the reason why in this study we concentrated on the area of soft policy measures.

The data for the meta-analytical investigation which we conducted came from 141 primary and secondary studies in the area of transport policy making. The studies had been carried out worldwide over the last fifteen years. The common goal of the studies was to evaluate the effectiveness of different measures of intervention in the area of soft policy measures, particularly with regards to the reduction of car use. The investigations tried to find out to which extent systematic interaction has the potential to lead to environment-friendly behaviour. Some examples of the interceding measures were/are education about the consequences of pollution caused by car use or education about advantages of using trains as a means of transportation. It was to be investigated out if these *soft policy measures* lead to the avoidance of environmentally unfriendly transportation habits.

All studies were of different regional origin or differed in quantity of collected data or in design. Moreover, the results of individual interventions, i.e. the intensity of their effects, differed greatly in some cases. A simple comparison of the studies did not clarify the question which intervention showed an effect, if they showed an effect at all. The implementation of systematic review techniques allowed for a first systematic comparison of studies. A first step was to differentiate by regional provenance. In addition, the studies were divided into three subgroups depending on the type of intervention. The statistical basis of primary and secondary analyses was proper for further investigations which used meta-analysis. In order to analyse the differences in percentage rate of car use before and after the implementation of an intervention, the approach of Cohen, as reported in Rosenthal (1994) was used. Thereby the overall effectiveness of applied interventions in the field of soft policy measures aiming at the reduction of car use could be evaluated. In all 141 studies a significant average overall effect of 0.15 was estimated (pre-post difference). This small but positive effect is of considerable significance in the field of transport policy making. Transportation experts acknowledged that so far the investment in improving infrastructure in the area of hard policy measures has been unsuccessful. This brings out the relevance of the calculated overall effect. Statements by transport experts support this insight into the area of interest.

A separate analysis of the sub-groups of transport interventions (work travel plans, school travel plans, and personalised travel planning / travel awareness campaigns / public transport marketing) showed that work travel plans were most effective. With an effect of 0.24 work travel plans had the comparably greatest impact on people's transport habits concerning the reduction of car usage. However, the three sub-groups vary among different parts of society. Work travel plans correspond to members of public or private enterprises whereas school travel plans are solely aimed at schools. The third type, personalised travel planning / travel awareness campaigns / transport marketing, incorporates all members of society and thus covers the greatest number of people. Personalised travel planning / travel awareness campaigns / transport marketing have a relatively small effect (0.11). However, because of the integration of a great part of society, the effect of car use reduction caused by the above stated measures is remarkable. The two interventions work travel plans and school travel plans show great differences in the results of individual studies. Therefore they can be described as statistically heterogeneous. The identification of possible moderators for the explanation of wide ranging results of individual primary analyses were demonstrated in the following article on work travel plans.

The third article '*Are Work Travel Plans Effective? - Systematic Review and Meta Analyses in the Transport Policy Domain*' draws a comparison between our own results and the results obtained by Cairns, Davies, Newson and Swiderska (2002) and Cairns, Sloman, Newson, Anable, Kirkbride and Goodwin (2004). We used the same data set Cairns et al. (2002) and Carins et al. (2004) used in their narrative reviews and conducted a quantitative meta-analysis which concentrated on the effectiveness of the soft policy measure work travel plan. Work travel plans are a special type of soft policy measure intervention, a bundle of measures to motivate staff to commute in a more sustainable way, especially to reduce single occupied vehicle usage.

It was found that measures which are obviously similar to one another in similar circumstances had a rather heterogeneous impact on single occupied vehicle usage by the company staff. This led to the conclusion that external factors (for example regional or organisational characteristics and differences in the design of the work travel plans) have a dominant impact on people's behaviour as has been reported by Cairns et al. (2002) and Cairns et al. (2004).

The analysis of the results of the narrative reviews showed that the great number of external factors which have to be taken into consideration (because they have an impact on the results) make an objective comparison of such studies impossible. A weighting of external factors could not be carried out successfully because neither size nor direction could be quantified via narrative reviews. Generally these circumstances lead to implementations which are hardly comprehensible or to a rather subjective emphasis of the study. Moreover, crucial differences in sample size, in effectiveness of measures and in the design of studies have not even been taken into consideration yet. This is due the fact that these parameters are hard to put into a qualitative evaluation. Therefore a first step has to be the identification of external factors and a systematic categorisation. These two procedures seem inevitable if further evaluations are expected to present objective results. During this preparatory stage all reported external factors from many different studies were collected regardless of their individual relevance. They were then grouped under the following categories: *Characteristics of the monitoring process*, *organisation's staff*, *site characteristics*, *location* and *description of the implemented work travel plans*. Within the 46 investigated work travel plans only 21 reported all moderators. This is significant for how systematically inconsistent the underlying reviews were.

Techniques of the meta-analysis made it possible to quantitatively describe the effect of the moderators (i.e. magnitude and direction of the impact of external factors).

Regardless of the heterogeneity of the data a mean effect of 0.27 (fixed effects model) was found throughout the 21 studies. According to Cohen (1988) this constitutes a small to medium effect on the reduction of car use via work travel plans. In comparison to the relatively small to negative effects that were found in the field of hard policy measures, our results are remarkable. Combining the moderators 'incentives for not parking in a public organisation which has been difficult to average bicycle access' and 'predominantly female employees' the obtained result was even higher, 0.56. Cohen (1988) classified this as a medium effect. In a private organisation which had an average distribution of male/female employees and did not give incentives for not parking but had almost similar bicycle access to the above mentioned public organisation an effect of only 0.19 was found. According to this, it can be seen as a minor effect. For practitioners who are concerned with the development of work travel plans, this provides two useful implications. On the one hand, the estimated effect can be calculated more accurately on the other hand the overall effect can be maximised through the variation of flexible factors. The three areas 'site factors, organisational factors and characteristics of the monitoring process' have the greatest impact on the effectiveness of work travel plans. The results from meta-analyses differ significantly from results obtained in narrative reviews. Cairns et al. (2002) and Carins et al. (2004) found through narrative reviews that parking is the essential factor for the success of a work travel plan. In contrast our meta-analysis concluded out that parking was a negligible factor.

The fourth article, '*Twenty Years after Hines, Hungerford and Tomera: A new meta-analysis of determinants of pro-environmental behaviouriour*' looked at psychosocial factors on environment-friendly behaviour. Whereas only facets of environment-friendly behaviour were quantitatively analysed in the previous articles, the complete set of data from peer reviewed articles since Hines, Hungerford and Tomera (1986/1987) were now considered. On the basis of a confirmatory structural equations model we looked for a causal relationship between psychosocial variables and environment-friendly behaviour. Singular aspects of environment-friendly behaviour are *travel mode choice*, *recycling* and *general ecological behaviour*. The structural equation model was developed on the basis of the *Theory of Planned Behavior* by Ajzen (1991) and the *Norm Activation Model* by Schwartz (1977). These constitute the most widely accepted and used proposals for the explanation of environment-friendly behaviour. The main goal was to test the causal determinants (Problem Awareness; Attribution; Social Norm; Feelings of Guilt; PBC; Attitude; Moral Norm; Intention and Behavior) of the model theories at hand on the basis of meta-analysis. Furthermore, we wanted to be able to quantify the causal connections between different factors which influence environment-friendly behaviour.

In order to conduct a meta-analytical summary of individual studies it was necessary to report the relations between determinants as bivariate correlations. Out of the total number of 128 studies, only 22 overall received 29 independent correlation matrices. During a first step, the 29 independent correlation matrices were grouped in a table of pooled weighted correlations by means of meta-analysis. In a second step the estimated true correlations were used as an input parameter for a structural equation model. The combination of the two procedures meta-analysis and structural equation modeling are referred to as MASEM.

The results of the MASEM analysis confirmed the structural relation between the theoretically postulated psychosocial determinants of environmentally friendly behaviour. The estimated effect that psychosocial variables directly impact the intention of environmentally friendly behaviour could be observed. However, a direct influence on behaviour through the psychosocial variables could not be found. Practitioners may draw the conclusion that there is only a mean positive relation between the intention to act environment friendly and the actual behaviour (be it environmentally friendly or not). Overall a combination of meta-analysis and structrual equations modeling allowed for a differentiated evaluation of the relation between psychosocial determinants.

In the following part I am going to critically analyse the strengths and weaknesses of the applied methods of scientific synthesis. In doing so, I will focus on the scientifically substantial problems.

Systematic review and meta-analysis have one clear advantage over narrative reviews. This is the possibility to systematically analyse great numbers of primary and secondary analyses. By the use of a rigidly systematic procedure a high degree of verifiability is achieved. The newly developed concept of analysis gave way for an easy integration of the results of recent studies as well as an improvement of the statistical power of the target dimension. The great number of existing primary and secondary studies in the field of transport policy making could be an ideal area of application for methods of systematic review and meta-analysis provided that all necessary data is reported. Unfortunately, this is not the case in the majority of studies. Relevant parameters (such as case numbers, effects etc.) are often not reported at all or are not reported in enough details. Therefore a large number of the data base had to be abolished despite its high thematic relevance. It could not be included in the study. Hence it was not available for a meta-analysis.

Singular findings, which differ immensely from the rest, can only be explained in the context of the specific circumstances. This is where the analysis of moderators plays an important role. A precondition for this is however that information on geographical, organisational and sociocultural particularities is at hand. Furthermore, the length and the date of the intervention should be well established. The above-mentioned problem especially matters in the area of soft policy measures because this is where relevant information is often not satisfyingly (completly) reported. However, using the extensive data base in the field of work travel plans, it could be shown that influences of the moderators could be used to account for the differences among individual results from primary and secondary analyses.

In the second and third analysis a simple before-after study design without having a control group was generally applied. These results proved to be somehow problematic. Trend effects throughout the procedure could hardly be controlled. Moreover, one company, which implemented a work travel plan, moved from a city to a rural area during the implementation of the work travel plan. This led to the falsification of results in this special case. The change of infrastructure caused an increase of car use, due to which the net effect of the work travel plan lost its verifiability. With a randomised control group, we would have been able of calculating a net effect. This indicates that if possible, a randomised control group should be incorporated into the study.

The major criticism of meta-analysis is that different studies are directly compared to one another although they often have distinct general conditions. Since consistent methodologies are seldom, anyone who intents to conduct a meta-analysis has to face this so called '*apples and oranges problem'*. However, by allocating studies into subgroups with similar general conditions the comparability of measures could be improved drastically. To do so, moderators were introduced which helped subdivide the studies. We successfully did so in study three '*Are Work Travel Plans Effective? - Systematic Review and Meta Analyses in the Transport Policy Domain*'.

Practitioners in the field of transport policy making, now have the possibility to extract relevant information from a great number of available studies. This way they can match general conditions better and more easily. Hence, meta-analysis provides a clear basis for decision making. This was not possible when only applying narrative reviews.

Several meta-analyses have revealed that most empirical primary and secondary studies were conducted in English speaking countries. More research is needed, especially in non-English speaking countries in order to get a better overall impression of the field and to obtain a better basis for meta-analysis.

During my work, we tried to show how the results of many individual studies can be summarised to one overall effect by using the methods and techniques above. We then gave three authentic examples in order to demonstrate how the techniques and methods can be used to attain insights into the global state of investigation.

Meta-analyses have the potential to lead to further progress of knowledge. There are two main reasons for that. On the one hand is the steady growth of the number of primary and secondary studies which are dedicated to transport policy. This also complicates the situation because the higher the number of different results, the harder it is to clearly define the state of research. On the other hand, meta-analysis significantly enhances the state of research by identifying new explanatory factors. The analysis of moderators is an important part of this because it makes the description and explanation of different empirical findings possible.

7) References for the Introduction and the Conclusion Sections

Ajzen, I. (1991). The theory of planned bejavior. *Organizational Behavior and Human Decision Processes, 50,* 179 – 211.

Bamberg, S., & Schmidt, P. (2003). Incentives, morality or habit? Predicting students car use for university routes with the model of Ajzen, Schwarz and Triandis. *Environment and Behavior, 34,* 264 - 285.

Cairns, S., Davies, A., Newson, C. & Swiderska, C. (2002). *Making travel plans work: Research report.* ((former) Department for Transport, Local Government and the Regions (DTLR)). London.

Cairns, S., Sloman, L., Newson, C., Anable, J., Kirkbride, A. & Goodwin, P (2004). *'Smarter Choices - Changing the Way We Travel'*. (Final report of the research project: The influence of soft factor interventions on travel demand. Research report for the Department for Transport). London. Retrieved December 1, 2005, from http://www.dft.gov.uk/stellent/groups/dft_susttravel/documents/page/dft_susttravel_029722.pdf.

Cochrane Muscoloskeletal Group (2006).. *What is a Systematic Review?* Retrieved February 3, 2006, from http://www.cochranemsk.org/cochrane/review/default.asp?s=1

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

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

Collins, John A., & Fauser, Bart C.J.M. (2004). Balancing the strengths of systematic and narrative reviews. *Human Reproduction Update, 11*, 103 - 104.

Energy Efficiency and Conservation Authority (2006). School Travel Plans. (EECA). Retrieved March 7, 2006, from http://www.eeca.govt.nz/transport/school-travel-plans/ index.html.

European Commission. (2001, September 12). *A transport policy for Europe's citizens*. Retrieved February 3, 2006, from http://europa.eu.int/comm/ transport/white_paper/doc/ip_2001_1263_en.pdf

Hall, J. A., Tickle-Degnen, L., Rosenthal, R., & Mosteller, F. (1994). Hypotheses and Problems in Research Synthesis. In H. Cooper, & L. V. Hedges (Eds.), *The Handbook of Research Synthesis* (p. 20). New York: Russell Sage Foundation.

Higgins J.P.T., & Green S. (Eds.). (2005). *Cochrane Handbook for Systematic Reviews of Interventions 4.2.5* (Rev. ed.). Chichester, UK: John Wiley & Sons.

Hines, J.M., Hungerford, H.R., & Tomera, A.N. (1986/1987). Analysis and synthesis of research on responsible environmental behavior: A meta-analysis. *Journal of Environmental Education, 18,* 1 – 8.

Lipsey, M. W., & Wilson, D. (2001). *Practical Meta-Analysis.* Thousand Oaks, CA: Sage.

Rosenthal, R. (1994). Parametric Measures of Effect Size. In Cooper, H., & Hedges, L. V. (Eds.), *The Handbook of Research Synthesis* (pp. 235, 237, 238). New York: Russell Sage Foundation.

Schwartz, S. H. (1977). Normative Influence on altruism. In L. Berkovitz (Ed.), *Advances in Experimental Social Psychology* (pp. 221 – 279). Vol. 10, New York.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752 – 760.

Steer Davies Gleave (October, 2003). *Evaluation of first yellow bus pilots schemes.* Retrieved Januar 8, 2006 from http://www.dft.gov.uk/pgr/sustainable/schooltravel/research/evaluationoffirstyellowbuspi5749

Wagner, M., & Weiß , B. (2005). Meta-Analyse als Methode der Sozialforschung. In A. Dieckmann (Ed.), *Methoden der Sozialforschung, Kölner Zeitschrift für Soziologie und Sozialpsychologie, Sonderheft 44/2004* (pp. 479 – 504). Köln: VS Verlag für Sozialwissenschaften.

8) Appendix I: A meta-analysis of the impact of new yellow school buses on pupils transport to school

I present here the meta-analytically analysed results from five studies; reported in in the study *Evaluation of First yellow bus pilots schemes, prepared for the Department for Transport* by Steer Davies Gleave (2003). Using the meta-analysis techniques described above, there is a slightly negative impact on public transport to schools, using Cohen's h as effect size statistic, a fixed effect of -0,0572 (p =0,0248, -0,1071 < CI < -0,0073). A random effects shows no significant findings, random effect -0,0311 (p=0,4543, -0,1125 < p < 0,0503). The random effects variance component is 0,002593. The results are homogeneous (p = 0,1997, Q = 5,9923, df = 4), so that the results from the fixed effects model are appropriate to use, see SPSS-Output. However, as the authors wrote, the impact is negative see Table: Effect Sizes Yellow Bus Scheme.

Table: Effect Sizes Yellow Bus Scheme

ES ID	Study Reference	Type of Inter-vention	No SOV before	No SOV after	No Car per 100 before	No Car per 100 after	*h*	w
1	Wrexham secondary - 2003	2	2.615	2.463	81,0	77,0	-0,0983	1000,606
2	Runnymede secondary - 2003	2	1.459	1.349	62,0	62,0	0,0000	434,572
3	Wrexham primary - 2003	2	80	47	55,0	66,0	0,2256	18,194
4	Hebden Bridge Primary - 2003	2	179	146	60,0	64,0	0,0824	49,967
5	Wrexham post 16 - 2003	2	161	70	79,0	81,0	0,0500	39,217

Lit.: Steer Davies Gleave (2003). **Evaluation of First Yellow Bus Pilots Schemes**. Prepared for Department of Transport, London, p. 21 – 23, 83.

SPSS-Output *Yellow Bus Scheme (2003)*:

```
Run MATRIX procedure:

*****  Meta-Analytic Results  *****

------- Distribution Description ---------------------------------
          N      Min ES      Max ES    Wghtd SD
      5,000       -,098        ,226        ,062

------- Fixed & Random Effects Model -----------------------------
         Mean ES    -95%CI    +95%CI        SE         Z         P
Fixed     -,0572    -,1071    -,0073     ,0255   -2,2452     ,0248
Random    -,0311    -,1125     ,0503     ,0415    -,7482     ,4543

------- Random Effects Variance Component ------------------------
v    =    ,002593

------- Homogeneity Analysis -------------------------------------
          Q          df          p
     5,9923      4,0000      ,1997

Random effects v estimated via noniterative method of moments.
```

Calculated with the *SPSS for Windows Meta-Analysis Macros*, written by Wilson, David B. (2006). http://mason.gmu.edu/~dwilsonb/ma.html.

Erklärung

Hiermit versichere ich, daß ich diese Dissertation – mit Ausnahme der in Koautorenschaft verfaßten und oben genauer spezifizierten Teile – selbständig verfaßt, keine anderen als die angegebenen Quellen und Hilfsmittel verwendet und sämtliche Stellen, die den benutzten Werken im Wortlaut oder dem Sinne nach entnommen sind, mit Quellenangaben kenntlich gemacht habe.

Gießen, im Jahr 2006

Unterschrift Guido Möser

Spezifizierung der Koautorenschaften

1. Artikel (Chapter 2):

Der Artikel *Meta-analysis: An alternative to narrative reviews for synthesising social science research?* wurde zusammen mit Herrn Prof. Dr. Peter Schmidt verfasst. Ich bin Erstautor des Artikels.

Eingereicht bei der Zeitschrift *European Societies*

Guido Möser:

1) Erstellen der theoretischen Konzeption und Schreiben des Theorieteils

2) Erstellen der kompletten englischen Fassung

3) Erstellen des empirischen Teils

4) Überarbeitung des theoretischen und empirischen Teils

Peter Schmidt:

1) Erarbeitung der theoretischen Konzeption

2) Überarbeitung des theoretischen und empirischen Teils

………………………………… …………………………………

(Guido Möser) (Prof. Dr. Peter Schmidt)

Gießen, den :……………………………

2. Artikel (Chapter 3):

Der Artikel *Are ‘Soft’ Policy Measures Effective in Reducing Peoples’ Car Use? A Meta-Analytical Review of Research Evidence* wurde in zusammenarbeit mit Herrn PD Dr. Sebastian Bamberg verfasst. Ich bin Erstautor des Artikels.

Eingereicht bei Zeitschrift *Journal for Environmental Psychology*

Guido Möser:

1) Konzeption und Struktur des Beitrags

2) Erstellung der ersten theoretischen Fassung

3) Identifikation, Sammlung und Analyse der relevanten Literatur

4) Erhebung der Daten

5) Durchführung der empirischen Untersuchung und Analysen

6) Diskussion der Schlussfolgerungen

Sebastian Bamberg:

1) Diskussion Konzeption und Struktur des Beitrags

2) Revision der ersten theoretischen Fassung

3) Kontrolle der Analysen und Überarbeitung der Diskussion

………………………………… …………………………………

(Guido Möser) (PD Dr. Sebastian Bamberg)

Gießen, den :……………………………

3. Artikel (Chapter 4):

Der Artikel *Are Work Travel Plans Effective? – Systematic Review and Meta Analysis in the Transport Policy Domain* wurde in Koautorenschaft mit Herrn PD Dr. Sebastian Bamberg verfasst. Ich bin Koautor des Artikels.

Eingereicht bei Zeitschrift *Transportation*

Sebastian Bamberg:

1) Konzeption und Struktur des Beitrags

2) Erstellung der ersten Fassung des theoretischen Teils und der Diskussion

3) Beschreiben der Analyseergebnisse

Guido Möser:

1) Überabreitung der ersten Fassung

2) Identifikation, Sammlung und Analyse der relevanten Literatur

3) Erhebung der Daten

4) Durchführung der empirischen Untersuchung und Analysen

5) Aufbereitung der Ergebnisse der Analysen für die Endfassung

6) Überarbeitung der Diskussion

………………………………… …………………………………

(PD Dr. Sebastian Bamberg) (Guido Möser)

Gießen, den :……………………………

4. Artikel (Chapter 5):

Der Artikel *Twenty years after Hines, Hungerford, and Tomera: A new meta-analysis of determinants of pro-environmental behavior* wurde in Koautorenschaft mit Herrn PD Dr. Sebastian Bamberg verfasst. Ich bin Koautor des Artikels.

Eingereicht bei Zeitschrift *Journal of Environmental Psychology*

Sebastian Bamberg:

1) Konzeption und Struktur des Beitrags

2) Erstellung der ersten Fassung des theoretischen Teils und der Diskussion

3) Beschreiben der Analyseergebnisse

Guido Möser:

1) Überabreitung der ersten Fassung

2) Identifikation, Sammlung und Analyse der relevanten Literatur

3) Erhebung der Daten

4) Durchführung der empirischen Untersuchung und Analysen

5) Aufbereitung der Ergebnisse der Analysen für die Endfassungen

5) Überarbeitung der Diskussion

………………………………… …………………………………

(PD Dr. Sebastian Bamberg) (Guido Möser)

Gießen, den :……………………………