Are Effects of Travel Feedback Programs Correctly Assessed?

Satoshi Fujii
Tokyo Institute of Technology, Japan

Sebastian Bamberg
University of Giessen, Germany

Margareta Friman
Karlstad University, Sweden

Tommy Gärling
Göteborg University, Sweden

Abstract

It has been claimed that travel feedback programs (TFPs) are effective in reducing levels of car-use related congestion, noise, and air pollution. The paper examines this claim, noting that many evaluation studies have employed less than optimal research designs. Demonstrating the use of meta-analysis, the results of 15 Japanese TFPs show that the effect sizes estimated for the frequently used research design lacking adequate control groups differ from the effect sizes estimated for research designs including adequate control groups. In addition, estimates of the homogeneity of treatment effects appear to differ, thus suggesting that inferences of causes of the effectiveness of a TFP vary with research design.

Key words: Travel behavior, Behavioral change, Travel feedback programs

Introduction

The urgent economic, social, and environmental problems being experienced worldwide due to increasing trends in car ownership and use have been frequently noted and documented (e.g., Black, 2001; Crawford, 2000; Goodwin, 1996; Hine & Grieco, 2003; Whitelegg, 2003). Various policy measures that aim at reducing the levels of car-use related congestion, noise, and air pollution have therefore been proposed and implemented. Some of these policy measures focus on changing or reducing demand for private car use. They are generally referred to as either *mobility management* or *travel demand management* (TDM) measures (Kitamura et al., 1997; Pas, 1995).

Table 1 presents major TDM measures that are considered to be or are being implemented in many urban areas worldwide. Several distinctions can be made based on classifications that have been proposed (see Loukopoulos, 2007, for an overview). One distinction is between pull measures that make alternative modes relatively more attractive and push measures that make a chosen mode relatively less attractive or even prohibited. Another distinction is that between measures that aim at changing travel behavior by changing available travel options (e.g., road or congestion pricing that increases monetary costs and decreases congestion), and those that aim at changing travel behavior by changing car users' beliefs, attitudes, values, and cognitive skills (i.e., informational and educational measures) without changing travel options. In this paper we focus on these latter type of policy measures, sometimes referred to as "soft" measures (Jones & Sloman, in press). Soft measures are in fact likely to be a necessary ingredient of any transport policy aimed at changing car use,

---

Insert Table 1 about here

A basic tenet in recent theoretical approaches to understanding travel behavior change (e.g., Bamberg et al., 2008; Gärling et al., 2002) is that behavior is goal-directed. A behavioral

change is therefore hypothesized to be initiated by a change goal related to attitudes and values, followed by the formation of a plan for how to change the behavior. Execution of the plan would in general require some acquired cognitive skills. The set goal may be conceived of as *feedforward* that in conjunction with *feedback* regulates the behavior change. *Negative* feedback informs about the discrepancy between the current state (e.g., frequency of car travel) and the change goal (e.g., reduction in frequency of car travel). *Positive* feedback informs about the distance from an undesirable goal (e.g., costs exceeding the household budget). Feedback may also have hedonic effects that increases or decreases subjective well-being.

Consistent with this theoretical stance, *travel feedback programs* (TFPs) have been developed (Fujii & Taniguchi, 2005), including individualized marketing (Brög et al., 2003), travel blending (Rose & Ampt, 2001), and personal travel planning (Jones & Sloman, in press). All TFPs share the common feature that targeted car users receive feedforward and feedback. Feedforward includes travel information (e.g., time tables, maps informing about alternative travel options for commuting or shopping). Feedback refers to information about behavioral consequences, for instance, CO_2 emissions caused by car use. TFPs differ with respect to location, technique, and procedure (see Table 2). There are three types of locations where TFPs have been implemented: residential areas, schools, and workplaces. TFPs in residential areas typically target daily travel behavior of any household member, whereas TFPs in schools and workplaces are typically confined to commuting trips. TFPs in schools may be implemented as part of the school curriculum.

Insert Table 2 about here

TFPs use several different techniques. These techniques differ with respect to whether they motivate travel behavior change, whether they provide customized information, whether they request setting goals of changing travel behavior, and whether they request plans to be made for how to change travel behavior. For example, individualized marketing does not provide motivational support (Brög, 1998), while travel blending does (Ampt & Rooney, 1999; Rose & Ampt, 2001). A TFP that involves planning includes a request that participants make plans for how to change their travel behavior. As an example, Fujii and Taniguchi (2005) proposed a TFP, implemented in several cities in Japan (see Fujii & Taniguchi, 2006), that required participants to form a behavioral plan for how to change their travel behavior. A final issue is whether the TFP provides customized information. Typical TFPs such as travel blending and individualized marketing do, but some less elaborated TFPs do not. For example, a TFP implemented in Obihiro, Japan provided participants with non-customized information about the bus service and requested that they made a behavioral plan for how to use the bus (Taniguchi & Fujii, 2007).

TFP procedures also differ. For instance, individualized marketing involves two or three contacts to conduct a survey of travel as well as of intentions to change travel behavior, to provide customized information, and to provide further customized information if necessary (Brög, 1998). Travel blending involves four contacts (Ampt & Rooney, 1999; Rose & Ampt, 2001): to motivate a travel behavior change, to conduct a travel diary survey, to provide customized comments, and to provide additional customized comments. The less elaborated TFP includes a single stage. For instance, the TFP in Obihiro, Japan (Taniguchi & Fujii, 2007) provided participants with a single questionnaire and non-customized information. The questionnaire included a request that participants formulate a behavioral plan for how to change their travel behavior.

TFPs have been implemented in several cities in Australia, Germany, Sweden, the UK, the United States, and Japan (see reviews in Department for Transport, UK, 2004a, 2004b; Fujii & Taniguchi, 2006). Individualized marketing has produced a reduction in car use up to 14% (South Perth, Australia) and not less than 2% (Breisgau-Hochschwartwald, Germany). Travel blending implemented in Australia and the United States produced a reduction in car use up to 15% (Adelaide, Australia) and not less than 9% (Brisbane, Australia). TFPs implemented in UK cities such as Gloucester, Bristol, and Nottingham have reduced car use by 7 to 15% in urban areas, and by 2 to 6% in rural areas. Transport for London implemented four different pilot TFPs, called "personalized journey planning" under the brand name TravelOptions, which reduced car use by 5 to 11% (Transport for London, UK, 2004a, 2004b). Note that the above percentages of car use reduction are regional-based total reduction of car use.

The effectiveness of TFPs implemented in Japan until 2003 was reported in Fujii and Taniguchi (2006). Thirteen Japanese TFPs reduced car use or CO_2 emissions by 0% to 40% with an average of 18%. Note that the reported percentages were based on the targeted sample, thus they differ from those presented in the reports from Department for Transport, UK (2004a, 2004b) that were based on aggregated data.

Despite this seemingly impressive evidence of the effectiveness of TFPs, transportation researchers (e.g., O'Fallon & Sullivan, 2003; Richardson, 2003; Stopher & Bullock, 2003; Stopher, Alsnih, Bullock, & Ampt, 2004) warn that without further critical analyses the evidence may paint a too optimistic picture. The main target of these criticisms is the methodological quality of the evidence for justifying the conclusion that TFPs are effective in reducing car use. Most of this evidence is based on evaluation studies that were not executed as rigorously as required. Möser and Bamberg (2007) analysed the methodological quality of 141 TFP evaluation studies with quite sobering results. All studies

use a weak experimental design lacking adequate control groups. The inability of this evaluation design to control nuisance factors drastically decreases the internal validity of causal inferences from the evaluation data (e.g., Shadish, Cook, & Campbell, 2002). Thus, the most important task of future studies of the effectiveness of TFPs is to launch evaluations that use more powerful experimental designs.

Another concern relates to the methods that have been used for synthesizing the results from available TFP evaluation studies. Most reviews of the effectiveness of TFPs use for this purpose narrative techniques, frequently in conjunction with quantitative methods adequate for disaggregate data analyses but not for synthesising aggregated study results (e.g., Transport for London, UK, 2004a, 2004b). The scientific literature on research synthesis increasingly questions whether such an approach provides a defensible way of research synthesis (e.g., Button & Kerr, 1996). Instead this literature recommends as a more adequate method of research synthesis quantitative meta-analyses developed in psychology and medicine for synthesizing the results of a body of independent studies. Bamberg and Möser (2007 and Möser and Bamberg (2007) provide a first demonstration of how to use quantitative meta-analysis techniques for synthesising the results of a body of TFPs. For 141 quasi-experimental TFP evaluation studies included in the analysis, these authors report a standardized random-effects mean effect size of 0.11. Translated back to the original metric, this indicates that on average the implementation of TFPs results in a 5% increase of the proportion of trips not conducted by car.

The Present Study

The aim of the present study is to show how evaluations of the effectiveness of TFPs may vary depending on research design chosen for the evaluations. A first section discusses the advantages and disadvantages of three different research designs that may be used. A second section describes the meta-analyses techniques that will be used for synthesizing the

results of a set of TFP evaluation studies. The third section presents the results of the analyses. In the last section the implications of the results for future evaluation studies are discussed.

Assessments of Effects of TFPs

In the evaluations of TFPs, different research designs have been used in order to evaluate their effects. Important methodological features include whether the research designs entail (1) disaggregated-level experiments including experimental and control groups, (2) a broad band of measures of beliefs, attitudes, cognitive skills, and actual travel behavior (car use, public transport use) in a targeted sample, or (3) aggregated-level observations of traffic volumes, regional modal shares of car and public transport, frequency and duration of traffic congestion, and number of passengers in bus or trains. In this connection a distinction can be made between research aiming at assessing factors that increase the effectiveness of TFPs and assessment of whether an implemented TFP has reached some targeted goal of car-use reduction, reduced emissions, or increased public transport use. In the latter case it is essential to obtain broad-band aggregated-level measures whereas the use of control groups is not always necessary; in the former case the use of control groups is essential for making valid causal inferences whereas broad-band aggregated-level measures are important for generalizing the results.

In this paper our focus is on disaggregated-level experiments aimed at assessing causes of the effectiveness of TFPs. Our argument is that control groups are needed, and in the following section we demonstrate empirically that erroneous conclusions are drawn if data from control groups are ignored.

Before-After Comparisons

Measuring travel behavior before and after the implementation of a TFP may appear to be adequate. This is referred to as a Treatment Group Pre-Post Test Only (TPP) experimental design (Shadish et al., 2002). However, a TPP design fails to eliminate confounding of

spontaneous changes over time (e.g., in experience, attitudes, or values), external factors (e.g., seasonal factors, changes in household economy), reactive effects of repeated measurements, and systematic changes in measuring instruments (Cook & Campbell, 1979; Shadish et al., 2002). If the sample is selected to include only an extreme group (as may be the case with habitual car drivers), regression towards the mean is expected when an imperfectly reliable measurement is repeated. A spurious regression effect thus threatens valid conclusions. In a similar vein, a before measurement would increase the effect of a TFP if it primes a positive attitude or informs the participants what is expected from them.

Experimental-Control Group Comparisons

Including control groups in an Only Post Test Control (OPC) experimental design with only after-measurements alleviates some of the problems pertaining to TPP designs at the same time as it introduces others. It needs to be stressed that OPC designs only allow valid causal inferences when participants are randomly assigned to experimental and control groups and when there is no differential mortality (i.e, selectivity; see Heckman, 1979). Thus, the possibility to draw causal inferences from OPC designs depends on the equivalence of experimental and control groups prior to the treatment. Randomly assigning participants to experimental and control groups is a powerful way to guarantee this equivalence. However, random assignment is only effective for large sample sizes. For small samples it would too frequently by chance result in non-equivalent groups that bias the estimation of the treatment effect. Generally, the randomized OPC design is still the best research design for assessing cause-effect relationships. It is easy to execute and, because it uses only a posttest, is relatively inexpensive. It is not susceptible to the above mentioned biases threating the internal validity of the TPP designs such as spontaneous changes, influences of external factors, reactive effects of repeated measurements, and systematic changes in measuring instruments. The only threat of internal validity associated with the OPC design is selection-mortality. This threat is

especially salient if there are differential rates of dropouts in experimental and control groups. If the treatment or control conditions are noxious or negative (e.g., a painful medical treatment), differential mortality may become a serious threat to valid inferences. In the case of TFPs, a mortality difference between control and experimental groups might emerge when response rate depends on willingness to change travel behavior is activated by the TFP intervention.

Non-randomized assignment of participants to experimental and control groups or the occurrence of differential mortality during the study changes the OPC design from true experimental to quasi-experimental. The internal validity of a quasi-experimental OPC design is considerably lower than that of an experimental OPC design.

One advantage of including pre-tests in an OPC research design is the possibility that this makes it possible to establish that the groups are equivalent before treatment starts. If the pre-test indicates that the groups are non-equivalent prior to the treatment, expanding the OPC design to the Pre-test-Posttest Control (PPC) design allows the use of the pre-test scores for statistically controlling the effects of the non-equivalence. However, this can never completely eliminate the threat to internal validity caused by non-randomization or differential mortality. This is so because the pre-test does not necessarily reflect all or not even the most important sources of non-equivalence, and thus cannot completely correct the biases caused by non-equivalence. A second advantage of the PPC design consists in the possibility to use the pre-test scores as co-variates when analyzing the effect of the intervention. This increases the power of the statistical tests of the treatment effects. However, it should be noted that the PPC design does not completely eliminate biases due to reactive effects of repeated measurements and differential mortality. Mortality may be different between experimental and control groups. Taking this into account, an OPC design with random assignment would be preferable to a PPC design with non-random assignment. Yet, if random assignment is not plausible due

to practical limitations, a PPC design may be used to correct biased conclusions from an OPC design with non-random assignments.

Empirical Comparisons of Evaluation Studies

In this section we describe the techniques for meta-analysis (Hedges & Olkin 1985; Rosenthal 1991; Cooper & Hedges 1994; Lipsey & Wilson 2001) that we will use.

Mean Effect Size

In meta-analysis, a mean effect size across the available single intervention studies is estimated, and subsequently the homogeneity of the effect size is assessed. The mean effect size is commonly obtained by weighting the individual effect sizes by the reciprocal of the effect size variance (Hedges & Olkin, 1985),

$$\overline{ES} = \frac{\sum_{j=1}^{k} w_j ES_j}{\sum_{j=1}^{k} w_j}, \qquad (1)$$

where ES_j is effect size obtained in study j, w_j is weight for ES_j , and k is the number of studies.

The standard deviation for $\overline{ES}$ ($\sigma(\overline{ES})$) is

$$\sigma(\overline{ES}) = \sqrt{\frac{1}{\sum_{j=1}^{k} w_j}}. \qquad (2)$$

The significance of $\overline{ES}$ can be tested by

$$z(\overline{ES}) = \overline{ES} / \sigma(\overline{ES}) = \frac{\sum_{j=1}^{k} w_j ES_j}{\sum_{j=1}^{k} w_j} \Bigg/ \sqrt{\frac{1}{\sum_{j=1}^{k} w_j}}. \qquad (3)$$

Note that when the calculated effect sizes are based on small sample sizes, there will be an upward bias. Sample sizes of the present studies are small according to Hedges (1992). The following correction for this bias will be used:

$$ES_j' = \left\{1 - \frac{3}{4N-9}\right\} ES_j \qquad (4)$$

where N is the total study sample size to be used in all the studies to be analyzed in the meta analysis.

Regarding homogeneity of the treatment effect across studies, Hedges' (1992) *Q*-test is commonly used to test whether the observed variance in effect sizes is larger than expected due to sampling error,

$$Q = \sum_{j=1}^{k} \frac{\left(ES_j - \overline{ES}\right)^2}{\sigma^2\left(ES_j\right)}. \qquad (5)$$

Under the null hypothesis of homogeneity, *Q* has a chi-square distribution with *k*-1 degrees of freedom.

If *Q*-test rejects the hypothesis of homogeneity of effect sizes across studies, a random effect size model, that accounts for the errors associated with sampling from populations that themselves have been sampled from a superpopulation, should be used for assessing $\overline{ES}$ (see Hedges & Vevea, 1998). The error term, therefore, contains variability arising from differences between studies in addition to within-study variability. Standard errors in the random-effects model are, therefore, larger than in the fixed case, which makes significance tests of combined effects more conservative.

Effect Sizes and Weights for Different Evaluation Designs

The method to calculate effect sizes (ES_j) and weights (w_j) of individual studies for the meta-analysis are different in different research designs. In a Treatment Group Pre-Post Test Only (TPP) design the respective effect size from the single evaluation study, $ES^{TPP}{}_{j,}$ is defined as follows (cf. Becker, 1988):

$$ES^{TPP}{}_j = (M^{after,E}{}_j - M^{before,E}{}_j) \,/\, SD^{P}{}_{j..} \qquad (6)$$

where $M^{after,E}{}_j$ and $M^{before,E}{}_j$ are the before- and after-measurements in the experimental group in study j, and $SD^P{}_j$ is the pooled standard deviation of the experimental group in study j. The standard deviation of this effect size ($ES^{TPP}{}_j$), necessary for calculating the single study weight is defined as

$$SD^{TPP}{}_j = \sqrt{\left\{\frac{2(1-\rho_j)}{n_j}\right\} + \frac{ES^{TPP}{}_j}{2n_j}}. \quad (7)$$

where ρ_j is the correlation of the before and after measurements in Study j and n_j is the sample size of the experimental group in study j. The weight for study j is calculated by

$$w^{TPP}{}_j = 1/(SD^{TPP}{}_j)^2. \quad (8)$$

Second, the effect size $\boldsymbol{ES^{OPC\_noB}}$ in an OPC design is defined as (c.f. Becker, 1988):

$$ES^{OPC\_noB}{}_j = (M^{after,E}{}_j - M^{after,C}{}_j) \,/\, SD^P{}_j. \quad (9)$$

where $M^{after,E}{}_j$ and $M^{after,C}{}_j$ are the after-measurement means for the experimental and control groups in study j, and $SD^P{}_j$ is the pooled standard deviation of the after-measurements for both groups. The standard deviation of this effect size, $ES^{OPC}{}_j$, necessary for calculating the single study weight is defined as

$$SD^{OPC}{}_j = \sqrt{\left(\frac{ne_j + nc_j}{ne_j nc_j}\right) + \frac{(ES^{OPC}{}_j)^2}{2(ne_j + nc_j)}}. \quad (10)$$

where ne_j and nc_j are the sample sizes of the experimental and control group in study j. The weight for study j is calculated by

$$w^{OPC}{}_j = 1/(SD^{OPC}{}_j)^2. \quad (11)$$

Finally, the effect size in a PPC design, $\boldsymbol{ES^{PPC}{}_j}$, can be calculated by the following equation (c.f. Becker, 1988):

$$ES^{PPC}{}_j = \{(M^{after,E}{}_j - M^{before,E}{}_j) - (M^{after,C}{}_j - M^{before,C}{}_j)\} / SD^{P}{}_{j..} \quad (12)$$

where $M^{after,E}{}_j$ and $M^{before,E}{}_j$ are the after- and before-measurement means for the experimental group and $M^{after,C}{}_j$ and $M^{before,C}{}_j$ are those for the control group in study j. SD_P is the pooled standard deviation for before-measurements in both groups. The standard deviation of this effect size, $ES^{PPC}{}_j$ is defined as

$$SD^{PPC}{}_j = \sqrt{\left(\frac{ne_j + nc_j - 2}{ne_j + nc_j - 4}\right)\left(\frac{2(1-\rho_j)(ne_j + nc_j)}{ne_j nc_j} + \left(ES^{PPC}{}_j\right)^2\right) - \left(ES^{PPC}{}_j\right)^2}. \quad (13)$$

where ne_j and nc_j are the sample sizes of the experimental and control group in study j and ρ_j is the pooling correlation across control and experimental groups of the before- and after-measurement in study j. The weight for study j is calculated by

$$w^{PPC}{}_j = 1/(SD^{PPC}{}_j)^2. \quad (14)$$

Translation of Effect Sizes to the Original Metric

Effect sizes given by eq. (1) are dimensionless. Transforming effect sizes to the original metric (e.g. trip rate per week) is important in order to assess the reduction of car use due to TFPs. The effect size given by eq. (1) can be translated into the original metric with the following equation

$$\overline{ES}^{original} = \overline{ES} \times SD^{before,C} \quad (15)$$

where $SD^{before,C}$ is the pooled standard deviation of dependent variable that can be obtained with the following equations;

$$SD^{before,C} = \sum_{j}^{k}\left(nc_j - 1\right)SD^{before,C}{}_j \Big/ \sum_{j}^{k}\left(nc_j - 1\right) \quad (16)$$

where, $SD^{beforeC}{}_j$ is the standard deviation of pre measurement in study j.

Mean of the dependent variable in the original metric before any experimental manipulation can be estimated by the pooled mean of the pre measurement in the control groups as follows

$$M^{before,C} = \sum_{j}^{k}\left(nc_j - 1\right)M^{before,C}{}_j \Big/ \sum_{j}^{k}\left(nc_j - 1\right) \quad (17)$$

The estimated value of the dependent variable after experimental manipulation ($M^{after,C}$) in the original metric can be obtained in as follows

$$M^{after,C} = M^{before,C} + \overline{ES}^{original} \quad (18)$$

Results

Access was obtained to data from 15 Japanese studies evaluating TFPs with pretest-posttest control (PPC) designs, where participants were randomly assigned to experimental and control groups. In these studies the average number of car trips per week was used as the central outcome variable for the evaluation of the TFP. The average correlation between the pre-test and post-test scores of weekly car trips was $\rho_j = 0.77$. The sample sizes of the control- and experimental groups included in our meta-analysis as well as the pre- and post-test means and standard deviations of the outcome variable in the experimental and control groups are reported in Table 3. A caveat is that most of the 15 evaluation studies are based on small sample sizes, and possibly as a consequence of these small sample sizes, in 8 of the 15 studies

the pre-test means raise doubts about the equivalence of experimental and control groups prior to the intervention. However, the differences in pre-test means between experimental and control groups, taking into account the weights of respective study, did not reach statistical significance ($z = 0.60$, $p = .55$). Therefore, even though the groups may not in all the studies be equivalent prior to the intervention, the total effect size taking into account the weights that are calculated using eq. (1) would not be biased. In the case when different TFPs were conducted at the same site (e.g., Sapporo or Ryugasaki), the different experimental groups were compared with the same control group. Even though a meta-analysis using TFPs in the same site does not violate the statistical assumptions of the meta-analysis, the fact the different experimental groups were compared with the same control group violates the assumption of statistically independent observation units. Therefore, the reported standard errors (SE) and z-values may be biased. Still, given the large number of observations, one should be justified in assuming that the control groups are independent.

Table 4 shows mean effect sizes ($\overline{ES}$) related to the three different research designs Treatment Group Pre-Post Test (TPP), Only Post-Test Control (OPC) and Pre-Post Test Control (PPC). $\overline{ES}$ for TPP was calculated with eqs. (6) and (8), $\overline{ES}$ for OPC with eqs. (9) and (11), and $\overline{ES}$ for PPC with eqs. (12) and (14).

Insert Table 3 about here

As can be seen in Table 4, for all three research designs, the *Q*-tests did not rejected the hypothesis of homogeneity across effect sizes, even though the test was marginally significant for the TPP design. Therefore, effect sizes can be assessed by fixed rather than random effects.

Insert Table 4 about here

The effect size for TPP was significant. The mean effect size -0,121 is equivalent to an average decrease in rate of car trips from 6.91 to 6.03 per week (a 12.7% reduction) in the original metric. The marginal significant mean effect size of –0.109 for the OPC design is equivalent to an average decrease in rate of car trips from 6.93 to 6.12 per week (a 11.4% reduction), and the marginal significant mean effect size -0.165 for the PPC design is equivalent to an average decrease in rate of car trips from 6.93 to 5.72 per week (a 17.2% reduction)

Discussion

Policy measures are needed that reduce current levels of car-use related congestion, noise, and air pollution. It is important to determine whether these measures are effective. It is also important in the development of such measures to learn why they are effective, so that they can be made cost-effective. For this latter purpose adequate methods permitting causal inferences are essential.

This paper has proposed an appropriate research design for making causal inferences concerning what factors make Travel Feedback Programs (TFPs) effective. Of three possible research designs, it is argued that Only Post-Test Control (OPC) designs are generally the most adequate if based on adequate sample sizes and if differential mortality is prevented. A more

convenient design is the Treatment Group Pre-Post Test (TPP) design. However, this frequently used design is a weak experimental design which does not permit strong causal inference. Thus, we recommend restricting the use of the TPP design to cases where the application of an OPC design is impossible. Including pre-test measures extended the OPC to the PPC design is another alternative. The main benefit of this extension is that it permits an explicit test of equivalence of experimental and control groups prior to the intervention as well as the occurrence of differential mortality. By reducing the error variance in the outcome variable, the use of pre-test scores as co-variates would also increase the power of statistical tests for detecting an intervention effect.

The present empirical analyses are based on a set of 15 TFP evaluation studies all using a PPC design. In the first step this body of 15 evaluation studies was used to demonstrate how to apply quantitative meta-analytical techniques for calculating an adequate estimate of the mean TFP impact across these single evaluation studies. Estimating the mean impact of TFP across a body of evaluation studies provides a more stable and generalizable estimate.

The second focus of the paper was analyses of the influence that different research designs on the estimated mean effect size. Fortunately, the analyses show that the direction of the mean effect sizes derived from the three different research designs is the same: All three effect sizes indicated that in the experimental group the introduction of a TFP reduces the average number of weekly car trips. However, at the same time the results also indicate a substantive difference in the average impact of TFP when using the effect sizes obtained from the three different research designs. Whereas the effect sizes derived from the OPC design showed a mean effect size of 0.11, the mean effect size derived from the PPC design was 0.17. Translated back to the original metric, this corresponds to a difference in the estimated average car use reduction caused by the implemented TFPs from 6.93 to 6.03 (a 11.4 % reduction) in the case of OPC-based effect sizes compared with a reduction from 6.93 to 5.72 (a 17.2 % reduction) in the case of the PPC-based effect sizes.

For the present set of evaluation studies, using the effect sizes from different research designs also influenced the statistical inference from the quantitative meta-analysis. Generally, TPP-based effect sizes seem to be associated with an underestimation of the standard error which leads to inflated *z*-values. Thus, using TPP-based effect sizes would result in the wrong conclusion that a significant experimental–control group difference exists when in fact this difference is random. In the present study control-group based effect sizes result in a more conservative statistical inference. In both cases the *z*-values only indicate a marginally significant control-experimental group difference.

The difference between the results of the OPC- and PPC-based effect sizes demonstrate the strong influence that non-equivalent experimental and control groups may have on the effect size estimates. As discussed above, the pre-test means indicated for at least 8 of the included 15 studies a substantive difference between experimental and control groups in the pre-tests. The general trend of the difference was a lower pre-test number of car trips in the control groups. The non-equivalence of control and experimental groups is probably the main reason for the low mean OPC effect size. The lower number of car trips in the control groups may thus artefactually reduce the true effect of the TFPs. The PPC-based effect sizes correct for this difference in the pre-tests, thus result in a stronger mean effect size. Although in some single studies pre-measures were different between groups, as mentioned above, there were no overall significant differences between experimental and control groups in the meta-analysis. Therefore, it may be concluded that generalizable estimates of effect sizes of TFP could be obtained by meta-analysis rather than from single studies. Such an advantage of meta-analysis in estimating effect sizes is more pronounced when the research are based on rather small samples studies like in the present study.

To summarize, whereas from a methodological point of view, the superiority of true experimental designs like the OPC design in providing effect sizes with a high internal validity is undisputable, in practice the advantages of experimental designs can only be realized if the

equivalence of experimental and control groups prior to the intervention is guaranteed. In the context of a planned experimental design non-equivalence can arise when sample sizes are too small or when differential mortality occurs. Practically, differential mortality should be prevented by using data-collection methods and settings reducing the drop-out rates.

The focus so far has been on effect sizes. Another important measure is the homogeneity index Q if the aim is to disentangle factors that make TFPs more effective. If this measure is statistically significant, it indicates that such factors exist, calling for additional analyses. The present results do not show that any of the Q values were significant, although they differed substantially depending on research design. There is thus a possibility that different research designs also may lead to different conclusions concerning causal factors.

The lesson from this exercise is still that randomly assigned control groups should be included in evaluations of TFPs, in particular if the aim is to make causal inferences of what factors increase their effectiveness. But this may also hold if policy decisions are based on estimates of effect sizes. Since employing control groups is inconvenient or sometimes not feasible, a solution may be to compare intact groups given that they are large and can be assumed to not differ. Future research may also try to disentangle nuisance factors rendering conclusions from a TPP design invalid. One possible such factor is the length of the time lag between pretest and posttest. Increasing this time lag would in general augment the probability of spontaneous changes as well as the influence from external factors.

Acknowledgment

Financial support for this research was obtained from the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), #17360244, 2007 to Satoshi Fujii, from the EU (Project no. 518368, Sixth Framework Programme) to Sebastian Bamberg, and from the Swedish Governmental Agency for Innovation Systems, grant #2004-02974 to Margereta Friman. We thank Takao Tanaka and Haruna Suzuki for assistance in compiling the data.

Address correspondence about this article to Satoshi Fujii, Tokyo Institute of Technology, Department of Civil Engineering, 2-12-1, Ookayama, Meguro-ku, Tokyo, Japan, 152-8552, tel & fax: +81-3-5734-2590, e-mail: fujii@plan.cv.titech.ac.jp.

References

Ampt, E., & Rooney, A., 1999. Reducing the impact of the car—a sustainable approach. Travel Smart Adelaide, 23rd Australasian Transport Forum, Perth, 29 September–1 October.

Bamberg, S., Fujii, S., Friman, M., & Gärling, T. (2008). *Soft transport policy measures: Do they work, and why do they work?* Manuscript submitted for publication.

Bamberg, S., & Möser; G. (2007). Why are work travel plans effective? – Comparing conclusions from narrative and meta-analytical research synthesis. *Transportation,*

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology, 41*, 257-278.

Black, W. R. (2001), An unpopular essay on transportation, *Journal of Transport Geography*, 9, 1-11.

Brög, W. (1998). *Individualised Marketing: Implications for TDM*. CD-ROM of Proceedings of 77th Annual Meeting of Transportation Research Board.

Brög, W., Erl, E., & Mense, N. (2003). *Individualised marketing: Changing travel behaviour for a better environment*. Paper presented at the TRIP research conference: The economic and environmental consequences of regulating traffic, Hillerød, 2-3 February.

Button, K., & Kerr, J. (1996, August). Synthesizing the results of quantitative case studies in transport analysis. Paper presented at 36th congress of the European Regional Science Association, ETH, Zürich, Switzerland.Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Chicago: Rand McNally.

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

Crawford, J. H. (2000). *Carfree Cities*, Utrecht, the Netherlands, International Books.

Department for Transport, UK (2004a). Smarter choices: changing the way we travel. London, UK.

Department for Transport, UK (2004b). Personalised travel planning demonstration programme, presented at Personalised Travel Planning: End of Programme Conference 2004, Bristol, UK.

Fujii, S., & Taniguchi, A. (2005). Reducing family car use by providing travel advice or requesting behavioral plans: an experimental analysis of travel feedback programs. *Transportation Research D*, 10, 385–393.

Fujii, S. & Taniguchi, A. (2006) Determinants of the effectiveness of travel feedback programs—a review of communicative mobility management measures for changing travel behavior in Japan, *Transport Policy*, *13*, pp. 339-348.

Gärling, T., Eek, D., Loukopoulos, P., Fujii, S., Johansson-Stenman, O., Kitamura, R., Pendyala, R., & Vilhelmson, B. (2002). A conceptual analysis of the impact of travel demand management on private car use. *Transport Policy*, *9*, 59-70.

Goodwin, P. B. (1996). Simple arithmetic. *Transport Policy*, *3*, 79-80.

Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica* **47** (1), 153-161.

Hedges (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.

Hine, J., & Grieco, M. (2003). Scatters and clusters in time and space: implications for delivering integrated and inclusive transport. *Transport Policy*, *10*, 299-306.

Jones, P., & Sloman, L. (in press). Encouraging behavioral change through marketing and management: What can be achieved? In K. W. Axhausen (Eds.),. *Moving through nets: The physical and social dimensions of travel*. Oxford: Elsevier.

Kitamura, R. (1988). An evaluation of activity-based travel analysis. *Transportation*, 15, 9-34.

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

Loukopoulos, P. (2006). A classification of travel demand management measures. In T. Gärling & L. Steg (Eds.), *Threats to the quality of urban life from car traffic: Problems, causes, and solutions* . Amsterdam: Elsevier.

Möser, G., & Bamberg, S. (2007). The effectiveness of soft transport policy measures: A critical assessment and meta-analysis of empirical evidence. *Journal of Environmental Psychology*, in press.

Pas, E. I. (1995), The urban transportation planning process. In S. Hanson (Ed.), *The geography of urban transportation* (pp. 53–77). Amsterdam: Elsevier.

Rose, G., & Ampt, E. (2001). Travel blending: An Australian travel awareness initiative. *Transportation Research D, 6*, 95–110.

Rosenthal, R.(1991). *Meta-Analytic Procedures for Social Research*. Newbury Park, CA: Sage.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

Steg, L. (2003). Factors influencing the acceptability and effectiveness of transport pricing. In: J. Schade & B. Schlag (Eds.), *Acceptability of transport pricing strategies* (pp.187-202). Amsterdam: Elsevier.

Taniguchi, A & Fujii, S. (2007) Promoting public transport using marketing techniques in mobility management and verifying their quantitative effects, *Transportation*, 34 (1), pp. 37-49.

Whitelegg, J. (2003), Transport in the European Union: time to decide, in N. Low and B. Gleeson (eds.), *Making Urban Transport Sustainable*, Basingstoke, UK, Palgrave Macmillan, pp. 115-131.

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.

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, August, 26th – August 30th, ETH Zürich, Switzerland.

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.

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

Table 1. Travel demand management measures (adapted from Steg, 2003).

TDM measure	Examples
Physical change measures	– improving public transport
	– improving infrastructure for walking and cycling
	– park & ride schemes
	– land use planning to encourage shorter travel times
	– technical changes to make cars more energy-efficient
Legal policies	– prohibiting car traffic in city centers
	– parking control
	– decreasing speed limits
Economic policies	– taxation of cars and fuel
	– road or congestion pricing
	– kilometer charging
	– decreasing costs for public transport

Table 2. Common features of travel feedback programs and features on which they may differ.

Feedback and feedforward information		
Location	**Technique**	**Procedure**
residential area (all trips)	*motivational support*	*single stage*
workplace (commute)	*customized information*	*multistage (travel diary*
school (commute)	*request goal setting*	*survey, feedback)*
	request plan formation	

Table 3.Descriptive for 15 TFP evaluation studies.

TFP cases	N (exp. group)	N (control group)	M (exp. pretest)	M (exp. posttest)	M (control pretest)	M (control posttest)	SD (exp. pretest)	SD (exp. posttest)	SD (control. pretest)	SD (control posttesty)
2003 Sapporo GIS-based TFP	26	21	6.28	5.56	5.43	5.05	5.48	4.83	6.55	6.22
2003 Sapporo Paper-based TFP	24	21	5.21	4.75	5.43	5.05	3.05	2.59	6.55	6.22
2005 Ryugasaki TFP with feedback comments	83	67	6.32	5.83	5.41	5.31	6.73	3.7	7.33	0.11
2005 Ryugasaki TFP without feedback comments	70	67	4.27	4.04	5.41	5.31	4.62	5.56	7.33	0.11
2005 Fukuoka home-visit TFP	103	72	8.33	6.93	7.56	8.82	12.04	10.43	9.17	18.34
2005 Takasaki new comer TFP	108	28	6.08	5.58	6.15	5.13	2.6	2.78	2.78	3.33
2005 Ryugasaki new comer TFP	21	25	4.95	5.3	3.68	5.08	3.18	3.28	3.43	3.12
2003 Kawanishi pt user TFP without behavioral feedback	108	52	7.61	7.09	6.81	8.75	6.86	6.58	6.37	7.86
2003 Kawanishi non pt user TFP without behavioral feedback	16	10	9.33	11.08	9.1	9.57	8.82	14.19	9.29	8.75
2003 Kawanishi non pt user TFP with ticket and without behavioral feedback	17	10	12.34	10.15	9.1	9.57	11.97	14.58	9.29	8.75
2003 Kawanishi TFP without non behavior change intention	18	19	9.71	9.45	10.8	12.9	4.88	8.49	7.56	11.2
2003 Kawanishi pt user TFP with behavioral feedback	106	52	6.28	6.07	6.81	8.75	6.28	6.79	6.37	7.86
2003 Kawanishi non pt user TFP with behavioral feedback	15	10	12.9	13.53	9.1	9.57	9.71	12.16	9.29	8.75
2003 Kawanishi non pt user TFP with ticket & behavioral feedback	16	10	8.89	6.56	9.1	9.57	6.04	5.97	9.29	8.75
2003 Kawanishi TFP with behavioral feedback without non behavior change intention	16	19	10.94	10.22	10.8	12.9	8.75	8.73	7.56	11.2

n = sample size, M = mean value (rate of trip per week), SD = standard deviation for the dependent variable, exp. = experimental group, control = control group.

Table 4. Mean effect sizes, 95% confidence interval, standard errors, z values for effect size, and homogeneity measure from meta-analysis.

	$\overline{ES}$	-95% *CI*	+95% *CI*	$SE(\overline{ES})$	$z(\overline{ES})$		Q		original metric		
									before (frequency per week)	after (frequency per week)	reduction rate
Treatment Group Pre-Post Test (TPP) design											
	-0.121	-0.100	-0.162	0.021	-5.68	(p<.001)	22.92	(p= .061)	6.91	6.03	12.7%
Only Post-Test Control (OPC) design											
	-0.109	-0.088	-0.150	0.06	-1.82	(p = .069)	13.01	(p = .525)	6.91	6.12	11.4%
Pre-Post-Test Control (PPC) design											
	-0.165	-0.144	-0.206	0.085	-1.95	(p = .052)	4.48	(p = .991)	6.91	5.72	17.2%

Note: $\overline{ES}$ = mean effect size; *CI* = Confidence Interval; SE($\overline{ES}$) = Standard error of $\overline{ES}$; $z(\overline{ES})$ = z-value for effect size; Q = homogeneity measure