A meta-evaluation of climate policy evaluations: findings from the freight transport sector

(article starts on next page)

A meta-evaluation of climate policy evaluations: findings from the freight transport sector

Lina Trosvik, Johanna Takman, Lisa Björk, Jenny Norrman & Yvonne Andersson-Sköld

View supplementary material ⌂

Published online: 08 Feb 2023.

Submit your article to this journal ⌂

Article views: 76

View related articles ⌂

View Crossmark data ⌂

A meta-evaluation of climate policy evaluations: findings from the freight transport sector

Lina Trosvik [a,b], Johanna Takman [b,c], Lisa Björk [a], Jenny Norrman [b] and Yvonne Andersson-Sköld [a,b]

[a]Swedish National Road and Transport Research Institute (VTI), Gothenburg, Sweden; [b]Department of Architecture and Civil Engineering, Chalmers University of Technology, Gothenburg, Sweden; [c]Swedish National Road and Transport Research Institute (VTI), Stockholm, Sweden

ABSTRACT

Knowledge about how implemented policy instruments have performed is important for designing effective and efficient policy instruments that contribute to reductions of greenhouse gas emissions. This paper carries out a meta-evaluation of ex-post evaluations of climate policy instruments in the freight transport sector. By analysing the outcomes and quality of evaluations, the aim is to identify whether estimated effects of policy instruments can be compared between evaluations and if the results are appropriate to use for evidence-based decision making. To analyse these aspects, commonly applied evaluation criteria are assessed and classified according to an assessment scale. We confirm that few ex-post evaluations are carried out and that there is a gap between evaluation theory and how ex-post policy evaluations are performed in practice, where evaluation criteria recommended in policy evaluation guidelines are found to often be neglected in evaluations. The result is a lack of systematic climate policy evaluation which hinders reliable conclusions about the effect of policy instruments. There is a need for more systematic monitoring and evaluation of implemented policy instruments and we suggest that evidence-based decision making can be improved by adjusting current policy evaluation guidelines and by introducing an evaluation obligation.

1. Introduction

Climate change is one of the world's most challenging problems and the reduction of greenhouse gas (GHG) emissions is vital to avoid detrimental effects on the environment and society (Intergovernmental Panel on Climate Change (IPCC), 2022). The IPCC of the United Nations (UN) published its sixth assessment report in 2021 with the overarching message that observed increases in GHG concentrations unequivocally are caused by human activities and that "global warming of 1.5°C and 2°C will be exceeded during

the twenty-first century unless deep reductions in CO_2 and other greenhouse gas emissions occur in the coming decades" (IPCC, 2021).

Following the Paris Agreement in 2015, where the goal was set to limit global warming to well below 2°C and preferably to 1.5°C compared to pre-industrial levels (UN, 2022), several countries have adopted targets of GHG emissions reductions within the transport sector to help curb climate change. For example, as part of the European Green Deal to reduce GHG emissions by at least 55% by 2030 compared to 1990 levels, the European Commission (EC) has proposed several policy measures in the "Fit for 55 Package" to target the emissions within the transport sector (EC, 2021, 2022). The transport sector has the highest reliance on fossil fuels of all sectors and accounts for about one-quarter of the global CO_2 emissions (International Energy Agency (IEA), 2021a, 2022a). Of the GHG emissions from the transport sector, freight transport (heavy duty trucks, rail, and shipping) accounts for about one-third of the emissions (IEA, 2021b).[1] Furthermore, global GHG emissions from freight transport have increased by about 37% over the last two decades (IEA, 2022a) and the demand for freight transports are forecasted to increase (IEA, 2022a; International Transport Forum (ITF), 2019).

To achieve necessary reductions of GHG emissions from freight transports, the use and implementation of policy instruments that can encourage a transition towards renewable fuels, a modal shift to less GHG-intensive transport modes, and an improvement of the operational and technical energy efficiency are needed (IEA, 2021b). To ensure that the targets of GHG reductions will be achieved at the lowest cost for society, the implemented policy instruments should be effective and efficient. However, although there exist numerous climate policy instruments in the freight transport sector, evaluations of them have been found to be lacking in many ways (ITF, 2022; Takman et al., 2020; Takman & Gonzalez-Aregall, 2021). For example, occasional evaluations are more common than regular monitoring of policy instruments and evaluations often lack quantifications of policy instruments' effects (ITF, 2022), which limits the understanding about the performance of currently implemented policy instruments and the continuous adaptation and improvement of them.

Policy evaluations are important for understanding how well policy instruments work and whether they have been successful in achieving their targets. They can provide useful information for policy makers on which policy instruments to implement in the future as well as how to improve or correct already implemented ones. However, previous literature has found that there is a gap between evaluation theory and how ex-post policy evaluations are performed in practice (Huitema et al., 2011) and that policy evaluations often use different types of methods which makes comparisons between them difficult (Harmelink et al., 2008; Haug et al., 2010). Furthermore, since policy evaluations may uncover critical problems of the evaluated policy instrument, which may call for legislative repeal, there is a risk of selective or biased policy evaluations (Bovens et al., 2008; Mastenbroek et al., 2016; Schoenefeld & Jordan, 2017). For example, Huitema et al. (2011) find that governmental bodies, which often have a specified policy agenda, are less likely to challenge established goals in policy evaluations than other actors, and Hildén (2011) finds that independent evaluations, which have greater possibilities to contribute to reflexive learning, are less likely to enter the policy cycle. To draw reliable conclusions about the performance of policy instruments and to ensure evidence-based decision making, it is essential that evaluations have adequate methodological quality and legitimate analyses.

The purpose of this study is to increase the knowledge about evaluations of climate policy instruments in the freight transport sector by reviewing policy evaluations and carrying out a meta-evaluation.[2] By examining which types of policy instruments that are evaluated and the outcomes and quality of evaluations, the aim is to identify whether estimated effects of policy instruments can be compared between evaluations and if the results are appropriate to use for evidence-based decision making. To analyse these aspects, commonly applied evaluation criteria are assessed and classified according to an assessment scale for each included evaluation.

Despite an increased implementation of climate policy instruments (Michaelowa et al., 2018) and an increased recognition of the value of policy evaluations (Fujiwara et al., 2019), there are relatively few studies that systematically compile and assess the effects and results of climate policy evaluations in practice. There are a few exceptions, such as Haug et al. (2010), Huitema et al. (2011), Auld et al. (2014), and Fujiwara et al. (2019), which provide systematic reviews of ex-post climate policy evaluations. What these studies have in common, including with this paper, is that they all use the method of meta-analysis. However, they differ from this paper since they mainly focus on the evaluation *outcomes*, whereas the *quality* of the evaluations is not considered to a large extent. Hence, this study contributes to the literature by increasing the knowledge about both the quality of policy evaluations and the effects of climate policy instruments. By analysing how evaluations fulfil recommendations in evaluation guidelines, this study contributes with knowledge and recommendations about how future evaluations should be designed to make their conclusions more accessible for policy makers and facilitate comparisons of policy instruments. In addition, most previous studies, except Auld et al. (2014), focus on evaluations of policy instruments implemented in European countries, whereas the scope of this paper includes policy instruments implemented also elsewhere in the world. There are, to the authors' knowledge, no studies that systematically compile and assess the effects and results of climate policy evaluations in the freight transport sector. Since previous studies of evaluations of climate policy instruments, except Fujiwara et al. (2019), are about a decade old, this study also contributes with updated knowledge.

2. Policy evaluation literature

To be able to review policy evaluations and carry out a meta-evaluation of them, it is first relevant to recognise and define what policy evaluation is and what defines a "good" policy evaluation. Crabb and Leroy (2008) define policy evaluation as "a scientific analysis of a certain policy area, the policies of which are assessed for certain criteria, and on the basis of which recommendations are formulated" (p.1). Vedung (2017) states that policy evaluation is a "careful, retrospective assessment of merit, worth, and value of the administration, output and outcome of government interventions, which is intended to play a role in future practical action situations" (p.3). According to the EC (2017a), which provides guidelines for policy evaluations, an ex-post policy evaluation should be an evidence-based judgement of the extent to which a policy instrument fulfils certain evaluation criteria. The evaluation should consider *why* and *how much* something has changed on account of the policy instrument, rather than just assessing *what* has happened. Furthermore, it should look for causality between the policy instrument and the observed

changes, and it should be carried out after a time period long enough to allow for changes to be identified and measured (EC, 2017a). The Organisation for Economic Co-operation and Development (OECD) also provides guidelines for policy evaluation, in which they include six evaluation criteria that should be used to support consistent high-quality evaluations with a common framework (OECD, 2021). They state that the criteria should be related to the aim and context of the specific evaluation and should not be applied in a fixed way for all evaluations. Furthermore, the interpretation of the evaluation criteria and the sources of evidence may be different depending on when the evaluation is made in relation to the policy instrument's implementation (OECD, 2021). It is therefore important to adapt the criteria for the specific evaluation and consider which of them that are possible to evaluate at different points in time. The most common evaluation criteria are described below in section 3.2.

3. Method

3.1. Systematic search for policy evaluations

This study uses a systematic search for policy evaluations. The method of systematic review facilitates the identification of all relevant research evidence that fulfils certain criteria set out in a search protocol, which can reduce the risk for a biased search (Adelle et al., 2012; Moher et al., 2009).

The search was divided into white and grey literature, where white literature refers to papers published in peer-reviewed journals and grey literature to literature produced by institutions not controlled by commercial publishers, such as governments, academia, businesses, and industry (Gokhale, 1997). The search strategy used in this paper, which is based on the methodologies described in Tsafnat et al. (2014) and Moher et al. (2009), can be described by the following steps: (1) Preparation: decision of databases and keywords to be used in the search, (2) Retrieval of items: searching in databases with the aim to find all relevant items, (3) Screening: removing duplicates, then screening titles and abstracts to remove irrelevant items, (4) Eligibility: screening full text for relevance and removing items that not fulfil inclusion criteria, (5) Snowball: following citations of included items to find additional items.

Whereas there are several databases of peer-reviewed articles that can be used to search for white literature, the search for grey literature is not as straightforward. Instead, the search for grey literature was based on a database of climate policy instruments in the freight transport sector compiled by two of the authors of this study in a previous research project (see Takman et al. (2020) and a description of the database in the supplemental online material).[3] The search procedure, search terms for each of the literature types, and the review process and its limitations are described in more detail in the supplemental online material.

The following six inclusion criteria were used for all identified studies: First, the evaluated policy instrument must be aimed at reducing GHG emissions in the freight transport sector (although it can cover additional sectors). Second, the policy instrument must be implemented as a public tool employed to correct for market failures and/or to reach objectives in society, thus private measures are excluded. Third, it must be an evaluation of actual outcomes of ongoing or terminated policy instruments, thus ex-ante evaluations

are excluded. Fourth, the evaluation must include an analysis of at least one of the six outcome evaluation criteria (described below in Table 1), thus status reports and other descriptive reports are excluded. Fifth, the evaluation must evaluate the impact on GHG emissions, although the impact does not have to be expressed in quantitative terms, and sixth, the publication year of the evaluations must be sometime over the period 2000–2021.

Table 1. Quality criteria based on criteria included in Mastenbroek et al. (2016), Huitema et al. (2011), and Crabb and Leroy (2008), and outcome criteria based on policy evaluation guidelines by the EC (2017a, 2017b) and the OECD (2021).

Criteria	Definition
Quality criteria	
Internal validity	Using the same data again, can the results be replicated? Is there enough information provided in the evaluation to be able to replicate the results (data sources and descriptions of the method)?
Reliability	Are references and data sources clearly presented and described? Are the variables in the data explained?
Robust methodology	Is the choice of methodology well-motivated and are potential weaknesses with the method mentioned/discussed?
Complexity	Are side effects and causality analysed (in relation to the outcome variables and the scope of the evaluation)?
Outcome criteria	
Effectiveness	This criterion involves an examination of the interventions' effects and the extent to which it achieves (or progresses towards achieving) its objectives. In cases where the intervention does not achieve its objectives, the effectiveness analysis should include an identification of factors hindering progress. The extent to which the observed effects can be linked to the intervention should also be analysed. Examples of questions to answer in the evaluation to fulfil this criterion include: Is the intervention achieving its objectives? What have been the effects of the intervention?
Efficiency	This criterion considers the relationship between the resources used for the intervention and the resulting effects and changes generated by the intervention. The evaluation of this criterion involves an examination of the extent to which the intervention delivers results in a timely and cost-effective way. Examples of questions to answer in the evaluation to fulfil this criterion include: How well are resources being used? To what extent has the intervention been cost-effective? To what extent are the costs of the intervention justified, given the effects it has achieved?
Relevance	This criterion involves an examination of the extent to which the objectives of the intervention are adequately defined, realistic and feasible, and whether they respond to the needs and problems in society. Examples of questions to answer in the evaluation to fulfil this criterion include: Is the intervention doing the right things? How well do the objectives of the intervention correspond to the needs?
Coherence	This criterion includes concepts of complementarity, harmonisation, and co-ordination. It involves an examination of how well the intervention works together with other interventions and actions. This may include internal coherence (i.e. coherence within institutions and with other interventions with similar objectives) and external coherence at different levels (i.e. coherence with other interventions and coherence with national and international obligations). Examples of questions to answer in the evaluation to fulfil this criterion include: How well does the intervention fit? To what extent is the intervention coherent internally and externally?
Impact	This criterion considers the ultimate significance, going beyond the effectiveness criterion and the immediate results, and involves an examination of the extent to which the intervention generates more transformative holistic effects. Such effects may include social, environmental, and economic effects or indirect consequences of the intervention, or enduring changes in systems or norms. An example of a question to answer in the evaluation to fulfil this criterion is: What difference does the intervention make?
Sustainability	This criterion involves an examination of whether the benefits (e.g. economic, social, or environmental benefits) of the intervention are likely to continue over the medium and long term. An example of a question to answer in the evaluation to fulfil this criterion is: Will the benefits last?

3.2. Meta-evaluation method

To analyse the evaluations included in the meta-evaluation, their content was compiled by using a template comprising information about the evaluation, the evaluated policy instrument, and the evaluation criteria. More specifically, the compiled information about the evaluation includes authors, title, abstract, type of study (white, grey), journal name (if white literature), and publication year. More specific information about the evaluation includes the affiliation of authors, whether the evaluation was commissioned, and the purpose of the evaluation. The template also includes the name of the evaluated policy instrument, which transport modes that are affected, the country in which the policy instrument is implemented, the time period that the policy instrument has been in force, and the targets, purpose and scope of the policy instrument.

The template also includes evaluation criteria related to the evaluations' outcomes and quality. Table 1 presents the definitions of the six most common evaluation criteria related to the analysis of results and findings, here referred to as outcome criteria, and the four most commonly discussed evaluation criteria measuring the quality of evaluations, here referred to as quality criteria.

The six outcome evaluation criteria included in this study are chosen based on recommendations in policy evaluation guidelines by the EC (2017a) and the OECD (2021). The evaluation criteria in these policy evaluation guidelines overlap, except for the criteria of "EU added value" and "sustainability" which are specific for the EC (2017a) and the OECD (2021) guidelines, respectively, of which the former is excluded in this study since it only applies for policy instruments implemented in the European Union (EU). The four quality criteria included in this study is based on criteria used by Mastenbroek et al. (2016), Huitema et al. (2011), and Crabb and Leroy (2008) and were chosen to measure the replicability of evaluations and the robustness and complexity of their methods. Other criteria included in previous studies, such as the description of the scope, the external validity, and the usefulness of the evaluation are less relevant to include for the scope of this study.

To analyse and compare the content of the policy evaluations, the assessments of the evaluation criteria were described both in a qualitative way and through an assessment scale. The assessment scale was used as a tool to compare how different evaluation criteria have been addressed across evaluations. For each included policy evaluation, each evaluation criterion was classified according to the scale as follows: it was marked with a dash symbol if the evaluation criterion was not analysed in the policy evaluation; it was marked with an empty square symbol if the policy evaluation analyses/discusses the evaluation criteria shortly (i.e. only parts of the aspects in the definition of the criterion are analysed/discussed); and it was marked with a black square symbol if the policy evaluation analyses/discusses the evaluation criteria in detail (i.e. all aspects in the definition of the criterion are analysed/discussed).[4]

4. Results and discussion

4.1. Overview of included policy evaluations

The total number of search hits was 2198, with the majority being white literature. After reviewing the studies' title and abstract, there were 293 studies potentially relevant to

include, and after reviewing their full text, 20 evaluations were found to fulfil the inclusion criteria and are included in the meta-evaluation. The most common reasons for exclusion are that the study does not focus on policy instruments or the transport sector, that it is not an ex-post evaluation, or that it does not examine the effects on GHG emissions. Many of the potentially relevant studies that evaluate policy instruments either focus on passenger transport, effects on other environmental issues than GHG emissions, such as air pollution, or on the policy instruments' implementation, compliance, or enforcement (rather than the policy outcomes). For example, among the search hits, there were several evaluations of low emission zones, congestion charges, and urban freight and city logistic policies, but since those primarily are aimed at reducing air pollution or congestion, rather than GHG emissions, they are not included in this analysis. Furthermore, many of the potentially relevant studies are ex-ante studies or simulations of potential future policy instruments.

Compared to previous studies, the number of included evaluations is low.[5] However, the inclusion criteria in this study are more specific, for example including the policy instruments' effect on GHG emissions and the focus on freight transport, which explain the lower number of evaluations. The disadvantage of few included evaluations is that the findings are not possible to generalise, but the advantage is that the evaluations can be analysed more comprehensively and provide more detailed findings.

Table 2 presents an overview of the included evaluations, sorted by the type of policy instrument evaluated. Most of the evaluations evaluate policy instruments affecting several sectors, of which freight transport is one of them, and few of the evaluations focus *only* on the effects in the freight transport sector. The most common types of evaluated policy instruments are different types of EU Directives or programmes (six evaluations) and taxes (six evaluations), followed by biofuel policies (two evaluations), and larger and heavier vehicles (two evaluations). There are also evaluations evaluating several policy instruments in the same study (two evaluations), one evaluation of a voluntary programme and one evaluation of a vehicle access restriction. A review of the evaluations' findings and conclusions is presented in the supplemental online material.

The motivations or purposes of the evaluations are seldom mentioned, except for evaluations of EU Directives or programmes where it is a requirement stated in the Directives. This raises the question of why some types of policy instruments are more frequently analysed than others. The evaluability of policy instruments (the extent to which a policy instrument can be reliably evaluated) may be one aspect affecting which policy instruments that are evaluated, which can be affected by the data availability and stakeholder interests of the intended use of the evaluation (Bovens et al., 2008; Davies & Payne, 2015).

4.2. Classification of evaluation criteria according to the assessment scale

The classification of evaluation criteria is presented in Figure 1, and Figure 2 summarises the number of evaluations classified according to the three levels on the assessment scale. Of the quality criteria, Figure 2 shows that the least frequently analysed/discussed criterion is complexity. More specifically, five evaluations have not included an analysis/discussion of side effects or causality of the policy instrument, and only three evaluations provide a more detailed analysis/discussion. The other quality criteria have a relatively

Table 2. Overview of the policy evaluations included in the meta-evaluation.

	Authors (publication year)	Type of study	Type of policy instrument	Purpose of policy instrument	Transport modes affected (country affected)	Years in force (evaluated years)	Affiliation of authors	Purpose of evaluation	Policy instrument found successful*
1	McKinnon (2005)	White	Larger and heavier vehicles: increasing maximum truck weight	Reduce truck traffic growth, promote fuel efficiency and cleaner vehicles	Heavy duty trucks (United Kingdom)	2001–ongoing (2001–2003)	University	Not mentioned	Yes
2	Palander (2017)	White	Larger and heavier vehicles: increasing maximum truck weight	Improve environmental emission efficiency for road transport in the Finnish forest industry	Heavy duty trucks (Finland)	2013–ongoing (2013–2014)	University	Not mentioned	Yes
3	Andersson (2019)	White	Carbon tax	Reduce GHG emissions	Transport sector (Sweden)	1991–ongoing (1960–2005)	University	Not mentioned	Yes
4	Shmelev and Speck (2018)	White	Carbon tax	Reduce GHG emissions	Transport sector (Sweden)	1991–ongoing (1961–2012)	Consultants	Not mentioned	Mixed results
5	Bernard and Kichian (2019)	White	Carbon tax	Reduce GHG emissions	Road transport sector (Canada)	2008–ongoing (1987–2016)	University	Not mentioned	Mixed results
6	Elgie and McClay (2013)	White	Carbon tax	Reduce GHG emissions	Road transport sector (Canada)	2008–ongoing (2008–2012)	University	Not mentioned	Yes
7	Best et al. (2020)	White	Carbon tax	Reduce GHG emissions	Road transport sector (several countries)	1990–ongoing (1990–2017)	University	Not mentioned	Mixed results
8	Aydin and Esen (2018)	White	Environmental taxes	Improve environmental problems and reduce GHG emissions	Road transport sector (several countries)	1995–ongoing (1995–2013)	University	Not mentioned	Mixed results
9	Canada's Ecofiscal Commission (2016)	Grey	Biofuel policies: production subsidies and biofuel mandates	Support the production and consumption of biofuels	Road transport sector (Canada)	Mid-2000s – ongoing (2010–2015)	University, consultants	Not mentioned	Mixed results
10	Navius Research (2020)	Grey	Biofuel policies: blending mandates, carbon pricing, low-carbon fuel standards	Reduce GHG emissions from the transport sector	Road transport sector (Canada)	2010–ongoing (2010–2019)	Consultants	Commissioned: provide an annual report of biofuels	Yes
11	Transport & Mobility Leuven et al. (2014)	Grey	EU Directive or programme: Directive 92/6/EEC (use of speed limitation devices)	Improve road safety and reduce GHG emissions	Heavy duty trucks (European Union)	1992–ongoing (1995–2011)	Consultants	Commissioned: required by the Directive	Yes
12	Ricardo Energy & Environment and TEPR (2015)	Grey	EU Directive or programme: Directive 2009/33/EC (the Clean Vehicles Directive)	Stimulate clean and energy-efficient vehicles by requiring procurers to consider environmental impacts	Light and heavy duty trucks (European Union)	2010–ongoing (2010–2014)	Consultants	Commissioned: required by the Directive	Mixed results

13	EC (2016)	Grey	EU Directive or programme: Directive 92/106/EEC (the Combined Transport Directive)	Shift road transport to more environmentally friendly modes	Modal shift of road freight transport to maritime transport and rail (European Union)	1992–ongoing (1992–2015)	European Commission	Commissioned: required by the Directive	Mixed results
14	ECORYS Nederland BV (2007)	Grey	EU Directive or programme: Marco Polo Programme I	Enhance intermodality, reduce congestion, and improve environmental performance of freight transport	Modal shift of road freight transport to maritime transport and rail (European Union)	2003–2006 (2003–2006)	Consultants	Commissioned: required by the Programme	Yes
15	Europe Economics (2011)	Grey	EU Directive or programme: Marco Polo Programme I & II	Enhance intermodality, reduce congestion, and improve environmental performance of freight transport	Modal shift of road freight transport to maritime transport and rail (European Union)	2003–2013 (2003–2010)	Consultants	Commissioned: required by the Programme	No
16	CE Delft et al. (2015)	Grey	EU Directive or programme: Directive 2009/28/EC (the Renewable Energy Directive)	Support production and promotion of energy from renewable sources	Transport sector (European Union)	2009–ongoing (2009–2014)	Consultants	Commissioned: required by the Directive	Yes
17	Creutzig et al. (2011)	White	Several policy instruments: Fuel efficiency standards, renewable fuel policies	Reduce GHG emissions	Road transport sector (several countries)	No specific year (no specific year)	University, research institute	Not mentioned	No
18	Touratier-Muller et al. (2019)	White	Several policy instruments: Voluntary reduction programme, mandatory compliance certification about freight CO_2 impact	Reduce CO_2 emissions by promoting synergies between shippers and carriers and improving information availability	Shippers and carriers (France)	2008–ongoing (information not available)	University	Not mentioned	Mixed results
19	Bynum et al. (2018)	White	Voluntary programme: SmartWay Transport Partnership	Improve fuel efficiency and reduce environmental impacts	Heavy duty trucks (U.S.A.)	2004–ongoing (2004–2015)	Government Agency: US. EPA	Not mentioned	Yes
20	Yusuf (2018)	White	Vehicle access regulation: freight vehicle access restriction policy	Improve safety and reduce delays, energy use, and emissions	Heavy duty trucks (Indonesia)	2011–ongoing (2010–2012)	University	Not mentioned	Mixed results

* Policy instruments are classified as successful (not successful) if the evaluation finds that the policy instrument has (not) been effective in reducing GHG emissions and recommends that the policy instrument should be continued (terminated or substantially improved). Evaluations are classified as "mixed results" if the evaluation finds that the policy instrument has reduced GHG emissions, but that it is not enough to reach targets and that improvements need to be made.

10 L. TROSVIK ET AL.

	Quality criteria				Outcome criteria					
	Internal validity	Reliability	Robust methodology	Complexity	Effectiveness	Efficiency	Relevance	Coherence	Impact	Sustainability
1	■	■	■	□	□	—	□	□	—	□
2	□	□	□	□	□	—	—	—	—	□
3	■	■	■	■	■	□	□	—	■	—
4	—	□	□	—	□	—	—	—	—	—
5	■	■	■	□	■	—	□	□	—	—
6	■	■	■	□	□	—	□	□	□	—
7	■	■	■	□	■	—	□	—	□	—
8	■	□	■	□	□	—	□	—	□	—
9	■	■	■	■	■	■	■	■	■	□
10	□	—	□	—	□	—	—	—	—	—
11	■	■	■	□	■	□	■	□	□	□
12	■	■	■	■	■	■	■	■	□	□
13	—	□	□	□	□	□	■	■	□	—
14	□	■	□	—	□	□	■	□	□	□
15	■	■	■	□	□	□	■	□	■	■
16	□	□	□	□	□	■	■	□	□	□
17	□	□	—	□	□	□	■	■	■	□
18	□	□	□	—	□	—	—	—	—	—
19	—	—	—	—	□	—	□	□	—	□
20	□	■	□	□	□	—	—	—	—	—

■ = Higher detail, □ = Lower detail, ▬ = Not included

Figure 1. Classification of the evaluation criteria for each included evaluation.
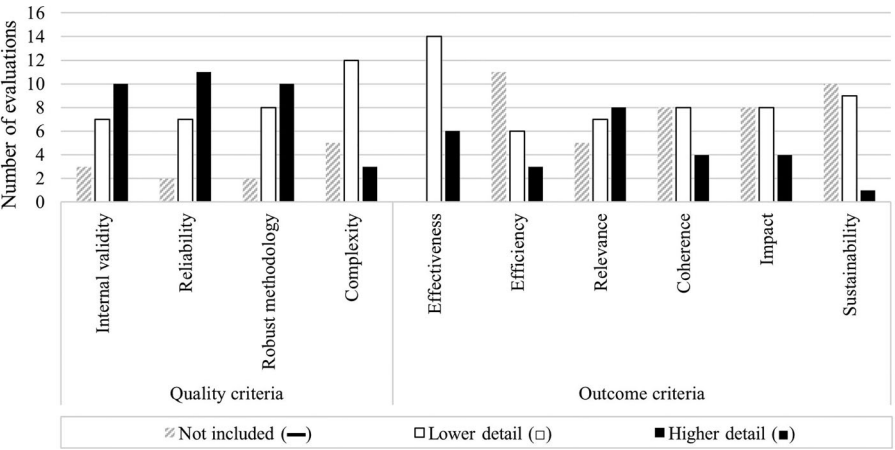

Figure 2. Number of evaluations classified according to the three levels on the assessment scale for each evaluation criteria.

similar distribution of the number of evaluations being classified at the different levels on the assessment scale, where around half of the evaluations are classified as having higher detail and around two evaluations do not include any discussion about the criteria. For the outcome evaluation criteria, Figure 2 shows that effectiveness is the most frequently analysed criterion (which is expected because it was one of the inclusion criteria). Most of the evaluations have analysed the criterion with a lower detail and only six evaluations have higher detail, including whether the observed effects can be linked to the policy instrument. The relevance criterion has the highest number of highly detailed analyses, and efficiency and sustainability are the criteria least frequently analysed in evaluations. The coherence and impact criteria have the same number of evaluations for the three levels on the assessment scale, with eight not analysing the criteria and four analysing the criteria in detail.

The most common reason for classification as lower detail of the evaluation criterion internal validity is inadequate descriptions of surveys or questionnaires used in the evaluations. For example, some evaluations used surveys to collect information to examine the effects of the policy instruments, but many lack information about aspects such as which questions that were asked and how respondents were chosen. For the evaluation criterion reliability, the most common type of missing information is a description of included variables. The evaluation criterion robust methodology addresses the motivation for the choice of method, and it was classified as high detail if the evaluation both motivates why the specific method is the most appropriate in addition to discussing potential methodological weaknesses. Even though a classification as lower detail does not necessarily imply that the methodology is of low quality, it still implies an uncertainty concerning the quality of the method as it is hard to appreciate its motivation and potential weaknesses.[6] These findings are in line with Turnpenny et al. (2009), who find that a challenge within policy evaluations has been how to conceptualise and measure the quality of evaluations, and with Harmelink et al. (2008) and Haug et al. (2010) which find that policy evaluations often use different types of methods which may make comparisons between them difficult.

The evaluation criterion complexity addresses whether the evaluation uses a method that allows an analysis of causality and potential side effects of the policy instruments. When evaluating effects of policy instruments, it is relevant to analyse whether observed effects are linked to the implementation of the policy instrument, or if there may be alternative explanations (EC, 2017a; OECD, 2021). However, this study finds that the most common method to analyse effects is to use statistics to describe or calculate effects, which involves a risk of non-causal interpretations. The complexity criterion is the least frequently classified as highly detailed in this study, indicating that causality rarely can be established. Drawing conclusions about policy instruments' effects, without discussing alternative explanations, may lead to misleading results and difficulties for policy makers in interpreting results.

The efficiency criterion is also relevant for the possibilities to compare effects of policy instruments. For example, a policy instrument could be found to lead to substantial reductions of GHG emissions but be very expensive in terms of costs for society, and the same reductions could potentially be achieved more efficiently. Hence, to understand how GHG emissions can be reduced to the lowest cost for society, an analysis of policy instruments' efficiency is highly relevant. However, this criterion is one of the least commonly evaluated outcome criteria among the included evaluations in this paper.

4.3. Affiliation of authors and methodological choices

To connect the findings in Figure 1 to the affiliation of authors, Figures 3 and 4 show the share of evaluations classified at the three levels of the assessment scale for each evaluation criteria sorted by affiliation by authors. The affiliation of authors is divided into university, consultants, and other, where other includes research institutes, government agencies, the EC, and evaluations having authors with different affiliations (e.g. one university-affiliated author and one consultant). For example, for the evaluation criterion internal validity, Figure 3 shows that 50% of the evaluations were classified as highly detailed, where 30% are written by authors affiliated to universities, 15% by consultants, and 5% by others. About the same distribution can be seen for the criteria reliability and robust methodology. For the complexity criterion, 60% of the evaluations are classified as lower detail, where 35% are written by university-affiliated authors. Figure 4 shows the corresponding results for the outcome criteria. Of the effectiveness analyses that have a higher detail, it is the same share of evaluations with authors having university affiliation as being consultants. For the criteria efficiency, relevance, coherence, and impact,

Figure 3. The share of evaluations for each quality criterion classified according to the three levels on the assessment scale, sorted by author affiliation.

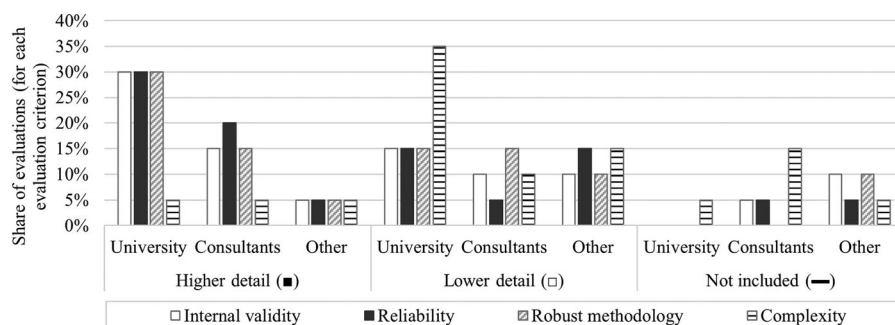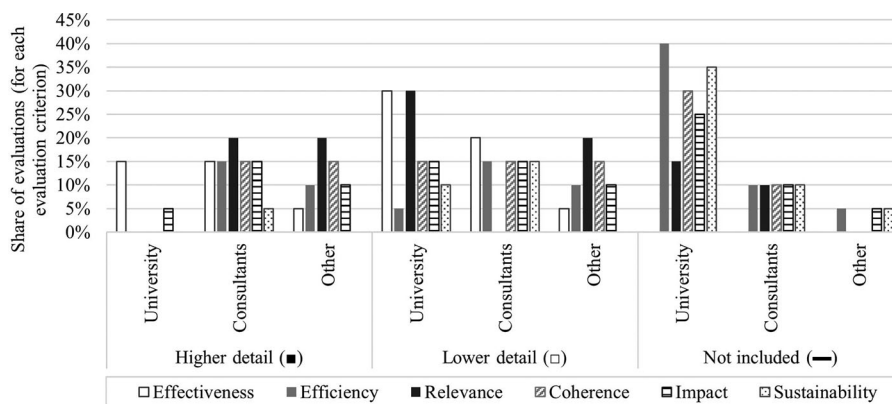**Figure 4.** The share of evaluations for each outcome criterion classified according to the three levels on the assessment scale, sorted by author affiliation.

consultants provided a detailed analysis for about 15–20% of the evaluations, whereas university-affiliated authors commonly did not present an analysis of these.

The main findings from these figures are that evaluations written by authors with university affiliation are found to include a higher level of detail about the quality criteria and fewer outcome evaluation criteria compared to authors affiliated as consultants. One explanation for these results is that consultants more often write reports (grey literature) with a broader scope than university-affiliated authors who more often write articles with a narrower scope to be published in peer-reviewed journals (white literature). A deeper analysis of the affiliation of authors is provided in section 4.4.2.

The meta-evaluation also comprises information about the evaluations' methodological choice and whether evaluations classify evaluated policy instruments as successful.[7] Nine evaluations conclude that the evaluated policy instrument has been effective in reducing GHG emissions and an additional nine of the evaluations present mixed results about the effectiveness of their evaluated policy instruments and argue that the policy instrument has achieved GHG emission reductions, although these are not enough to reach targets. Two evaluations find that the policy instrument has not been effective in reducing GHG emissions and argue that the policy instrument should be terminated or substantially improved. There is almost no difference between white and grey literature in the conclusions about the policy instruments' successfulness. Of the nine evaluations that have concluded the evaluated policy instrument(s) to be successful, four have a higher detail level on the criteria internal validity, reliability, and robust methodology, only two have a higher detail on the effectiveness criterion, and only one has a higher detail level on the complexity criterion. The most common type of method is to use statistics to make calculations or descriptive analyses to examine the effects of policy instruments, for which eight are grey literature and five are white literature. Six evaluations use econometric approaches, and these are only white literature. Literature review and survey/interviews are used by eight and seven evaluations, respectively, and is more commonly used in grey literature, especially for survey/interviews. Due to uncertainties related to the methodological quality and weaknesses in terms of lacking causality analyses, the conclusions about successfulness from these evaluations should be interpreted with caution (see section 4.4.1 for a deeper analysis).

4.4. Case studies of how studies approach different evaluation criteria

With the aim to better understand how different evaluation criteria are approached in evaluations and to assess how the fulfilment of different criteria can be improved in future evaluations, this section reviews certain criteria in more detail.

4.4.1. The complexity criterion

The evaluation by Bynum et al. (2018), which examines how a voluntary policy model contribute to fuel efficiency and reduced environmental impacts from freight transport, is classified as not having analysed the complexity criterion. By reviewing the literature and providing statistics of CO_2 emission reductions, they conclude that the evaluated policy instrument is effective. However, side effects and causality are not discussed, which weakens the relatively strong conclusion of the policy instrument being effective. For example, observed effects are only provided for firms that have entered

the programme but lacks a comparison with firms that did not. Therefore, a relevant discussion could be whether there is a risk for a self-selection bias, i.e. that firms that already have taken steps towards reducing CO_2 emissions may be more likely to enter the programme than others. Another relevant analysis could be to compare observed effects between included firms with firms outside of the programme to examine whether the same change can be seen in the whole sector or whether the effects are likely an effect of the programme. Furthermore, the evaluation mentions that a mandatory programme that sets GHG and fuel efficiency standards recently was implemented, but no discussion of its potential effects is included. In addition, no potential side effects from the policy instrument are discussed, where for example, factors such as changes in freight transport supply, rebound effects, or modal shifts could be relevant to discuss.

The evaluation by McKinnon (2005), which evaluates the effects of increasing the maximum legal weight of trucks in the UK, is classified as having analysed the complexity criterion at lower detail. By providing statistics of traffic levels, transport costs, and emissions, the paper finds that the evaluated policy instrument has yielded economic and environmental benefits and that there is a motivation for a further increase in the weight limit. Although McKinnon (2005) shortly discusses the potential impact of other policy instruments, potential side effects, such as a shift of freight from rail to road, are not discussed.

The study by Andersson (2019), which assesses the environmental efficiency of the Swedish carbon tax on transport fuels, is classified as having analysed the complexity criterion at higher detail. By using an econometric approach and a synthetic control method to estimate the effects in the transport sector, Andersson (2019) finds that the policy instrument has been successful in significantly reducing CO_2 emissions. The method of constructing a "synthetic Sweden" (a control group of OECD countries that resemble Sweden) represents the counterfactual of not implementing a carbon tax (Andersson, 2019), which allows for a causality analysis and an examination of the extent to which the observed effects can be linked to the intervention. In addition, the paper examines alternative explanations for the results and discusses side effects such as risks for carbon leakage through cross-border shopping of fuel.

All three studies above conclude that the evaluated policy instruments have been successful, yet the evidence supporting their conclusions differ. Conclusions about a policy instrument being successful without consideration of side effects or causality may lead to overestimated or underestimated effects and misleading recommendations for policy makers. Of the nine evaluations in this study that have concluded that the evaluated policy instrument(s) are successful, only one is classified as having a higher detail level on the complexity criterion. Though causality can be difficult to establish, evaluations should nevertheless include discussions of alternative explanations to the observed effects and discuss potential side effects that may influence the results to better inform evidence-based decision making.

4.4.2. Objectivity and affiliation of evaluators

As policy evaluations may uncover critical problems of the evaluated policy instrument, which may call for legislative repeal, there is a risk of selective or biased policy evaluations due to political pressures. This risk is increased when the evaluator has a governmental connection or when governmental actors commission organisations to conduct

evaluations (Bovens et al., 2008; Mastenbroek et al., 2016; Schoenefeld & Jordan, 2017). Other sources of influence, for example lobbying or economic interests may of course also influence how evaluations are carried out and which results that are reported. As described in section 4.3, there is almost no difference between white and grey literature in terms of the results regarding the policy instruments' successfulness. This indicates that there is no difference in judgements made of policy instruments' successfulness. However, depending on the interpretation of what a successful policy instrument is, evaluations that are classified as having mixed results about the effectiveness could in many cases instead be classified as not successful. For example, although evaluations classified as mixed results may observe reductions of GHG emissions, some conclude that the stringency of the policy instrument should increase, which in a way is equivalent to the policy instrument not being successful in terms of reaching the targets in its current design. For example, Ricardo Energy & Environment and TEPR (2015) find that the Clean Vehicles Directive has had a very limited effect on CO_2 emissions, but still recommends that the policy instrument should be retained because there are no alternative policy instruments identified. Another example is CE Delft et al. (2015), which evaluate the Renewable Energy Directive and find that the growth rate of renewable energy was lower than necessary to achieve the target in the transport sector, but still recommend that the Directive should be maintained due to policy stability. In both of these cases, the evaluations were commissioned by the EC.

In Bynum et al. (2018), the evaluators are affiliated with the same agency that launched the evaluated policy instrument. Although this may involve benefits such as the authors having an inside knowledge for how the policy works, there is also a risk for a lack of independence and less critical evaluations (Schoenefeld & Jordan, 2017). There are some examples in Bynum et al. (2018) that raises a question of such lack of independence. First, the authors only present barriers for reducing GHG emissions that are addressed by the evaluated policy instrument, while not mentioning other barriers not addressed by the policy instrument. Second, the evaluation concludes that the programme has observed significant behavioural changes, but states that these cannot be quantified. Third, the evaluation provides statistics of reduced CO_2 emissions for only two categories, while the CO_2 emissions for other categories are not evaluated. Despite weaknesses in evidence, the authors conclude that the evaluated policy instrument is effective. These examples could be an indication of a situation referred to as a confirmation bias, where only the evidence that support the policy instrument are presented, and which Schoenefeld and Jordan (2017) argue is a risk in internal evaluations.

If evaluations' analyses of the quality criteria are lacking, there is a higher risk for biased evaluations due to reduced replicability and verification of evidence. The finding that evaluations written by authors with university affiliation more often include a higher level of detail about the quality criteria compared to authors affiliated as consultants suggests that these more often take measures to reduce the risk for biased evaluations. As previously mentioned, university-affiliated authors more often aim to publish articles in peer-reviewed journals (white literature), which potentially can explain the higher level of detail about the quality criteria. Being able to publish an evaluation in peer-reviewed journals could be a guideline for policy evaluators for how quality criteria should be implemented.

Although these findings cannot be generalised due to the low number of included evaluations, they are in line with findings of earlier studies. Few of the included evaluations state the motivation or purpose of the evaluation, discuss the extent to which established political targets are appropriate, or whether there are any competing interests involved, which is a finding in line with Turnpenny et al. (2009), who found that few evaluations address or question underlying political motivations or the framing of the evaluations. Additionally, both Haug et al. (2010) and Huitema et al. (2011) found that most evaluations analyse the effectiveness and goal achievement of the policy instrument, but that few analyse reflexive learning.

4.4.3. How variations in quality criteria may affect comparability of results

An example of a difficulty in comparing results of evaluations is when evaluations come to different conclusions. For example, of the two evaluations of the Swedish carbon tax included in this study, one finds that the tax is an effective and efficient policy instrument (Andersson, 2019) and the other finds no significant effect and concludes that policy makers should not rely entirely on taxation to achieve environmental targets (Shmelev & Speck, 2018). Since Andersson (2019) describes the method in detail, discusses methodological weaknesses, compares results with earlier studies, and includes a causality analysis as well as an identification of potential side effects, the evaluation has been classified as having higher detail on both robust methodology and complexity. Shmelev and Speck (2018) include almost no description of the method, do not discuss potential weaknesses, and do not include any complexity analysis, which is why the evaluation criteria robust methodology and complexity have been classified as the lowest level of detail. Based on this information, the findings by Andersson (2019) seem more reliable compared to those by Shmelev and Speck (2018) since the method is explained in more detail, is possible to replicate, and includes an analysis of whether the observed effects can be linked to the policy instrument. This type of finding highlights the importance of comparability between evaluations to ensure evidence-based decision making. It also points to how the criteria can be used to determine how to interpret contradicting results on the same policy.

5. Conclusions and recommendations

To increase the knowledge about evaluations of climate policy instruments in the freight transport sector, this study reviews policy evaluations and carries out a meta-evaluation. The aims are to identify whether estimated effects of policy instruments can be compared between evaluations and if the results are appropriate to use for evidence-based decision making, which can contribute with important implications for policy makers aiming at reducing GHG emissions from freight transport. Based on the results and discussion, this study comes to the following conclusions and recommendations:

- *Few policy evaluations were found to evaluate the effects on GHG emissions of climate policy instruments in the freight transport sector*. This is of high concern given the ambitions political goals with respect to cutting GHG emissions and the numerous climate policy instruments that, at least partly, are designed to address emission reductions.

Instead, many evaluations analyse effects on other environmental issues, such as air pollution, or the policy instruments' implementation, compliance, or enforcement.

• *An obligation to evaluate policy instruments should be considered*. This study shows that due to insufficient transparency of methodological choice and data sources, as well as insufficient analyses of the policy instruments' performance, many evaluations are not suitable to inform evidence-based decision making or for comparing the performance of different policy instruments. To ensure that climate targets are reached efficiently, there is a need for more systematic monitoring and evaluation of implemented policy instruments that could inform adaptations or removals of existing policies and introduction of new. This could potentially be achieved through a policy evaluation obligation. Such an obligation should at least involve the following conditions: First, policy instruments must have clear targets that can be assessed. Second, the availability of data needed to evaluate policy instruments must be reviewed at the time of policy implementation. If the required data is unavailable, measures should be taken to collect such data to increase the evaluability of the policy instrument. Third, clear guidelines and evaluation procedures are needed for evaluators (see below).

• *Evaluations guidelines for evaluators should be improved*. Although current evaluation guidelines (by the EC and the OECD) include relevant evaluation criteria, they could be improved to enable more consistency and comparability across evaluation studies. Methodological recommendations are currently missing in the guidelines by the OECD (2021) and partly missing in the guidelines by the EC (2017a). The OECD (2021) does not include any specific methodological recommendations in the evaluation guidelines, while the EC (2017a) states that the most appropriate method should be used to assess impacts, that evaluators should be transparent about the methodological choices, and that limitations of data or methodologies, as well as risks of unintended consequences, should be transparent and clearly mentioned. Publishers of policy evaluation guidelines should consider including the following recommendations: First, to be able to compare climate policy instruments in the transport sector and understand whether they contribute to achieving targets to the lowest cost for society, all evaluations must at least include an assessment of their effectiveness and efficiency. Second, to ensure that the estimated effects of policy instruments are not overestimated or underestimated, analyses of the policy instruments' side effects and the causality between the implementation and the effects should be included. Third, the evaluating actor must be unbiased, and any competing interests must be clearly stated. Fourth, to be able to draw reliable conclusions about policy instruments' performance, conclusions in evaluations should be based on appropriate evidence where the credibility and interpretation of the evidence should be clear. Therefore, guidelines of policy evaluations should consider including recommendations of well-motivated choice of methodology, comparisons with related studies, and discussions of potential weaknesses.

Policy evaluations are highlighted as an important part of ensuring evidence-based decision making, and they are often discussed as something that we need more of. Yet, this study has pointed out that it may be just as important to ensure that evaluations have sufficiently high quality and address aspects related to causality and efficiency. This study concludes that there is too little systematic climate policy evaluation in the freight

transport sector to be able to compare results, draw reliable conclusions and support evidence-based policy making. To understand how climate policy evaluations can become more systematic and of higher quality, relevant areas for future research may include investigating barriers and conditions for the evaluability of climate policy instruments, the connection between ex-ante and ex-post evaluations, and which evaluation methods that are suitable for evaluating different types of policy instruments.

Notes

1. Rail and shipping include both passenger transport and freight transport.
2. Here, a meta-evaluation is defined as a systematic review of evaluations to determine the quality of their methods and findings (Cooksy & Caracelli, 2005).
3. The policy database is available by contacting the corresponding author of this study.
4. A classification as lower detail does not necessarily imply that the evaluation has low quality, it only indicates that the criterion was not discussed or motivated in detail.
5. This can be compared with 165 evaluations in Auld et al. (2014), 236 evaluations in Fujiwara et al. (2019), 262 evaluations in Haug et al. (2010), and 259 evaluations in Huitema et al. (2011).
6. Regarding the criteria of robust methodology and reliability, it is beyond the scope of this study to analyse the quality of evaluations' methods and whether relevant and complete data sources has been used. Such analyses would require that all included evaluations have well-described methodologies and that the authors of this study have good knowledge within all methodologies used in the included evaluations as well as research areas. Thus, to avoid misleading or biased analyses of these criteria, this study reviews the description and motivation for the chosen method and the presentation of data sources and references.
7. Figure S1 in the supplemental online material presents a summary of the evaluations' conclusions about the policy instruments' successfulness. Figure S2 in the supplemental online material shows the number of evaluations using different types of methods.

Disclosure statement

Funding

ORCID

Lina Trosvik 🔟 http://orcid.org/0000-0002-8511-882X
Johanna Takman 🔟 http://orcid.org/0000-0001-7688-4808
Lisa Björk 🔟 http://orcid.org/0000-0002-9878-6923
Jenny Norrman 🔟 http://orcid.org/0000-0003-2849-7605
Yvonne Andersson-Sköld 🔟 http://orcid.org/0000-0003-3075-0809

References

Adelle, C., Jordan, A., & Turnpenny, J. (2012). Proceeding in parallel or drifting apart? A systematic review of policy appraisal research and practices. *Environment and Planning C: Government and Policy*, *30*(3), 401–415. https://doi.org/10.1068/c11104

Andersson, J. J. (2019). Carbon taxes and CO2 emissions: Sweden as a case study. *American Economic Journal: Economic Policy*, *11*(4), 1–30. https://doi.org/10.1257/pol.20170144

Auld, G., Mallett, A., Burlica, B., Nolan-Poupart, F., & Slater, R. (2014). Evaluating the effects of policy innovations: Lessons from a systematic review of policies promoting low-carbon technology. *Global Environmental Change*, *29*, 444–458. https://doi.org/10.1016/j.gloenvcha.2014.03.002

Aydin, C., & Esen, Ö. (2018). Reducing CO2 emissions in the EU member states: Do environmental taxes work? *Journal of Environmental Planning and Management*, *61*(13), 2396–2420. https://doi.org/10.1080/09640568.2017.1395731

Bernard, J. T., & Kichian, M. (2019). The long and short run effects of British Columbia's carbon tax on diesel demand. *Energy Policy*, *131*, 380–389. https://doi.org/10.1016/j.enpol.2019.04.021

Best, R., Burke, P. J., & Jotzo, F. (2020). Carbon pricing efficacy: Cross-country evidence. *Environmental and Resource Economics*, *77*(1), 69–94. https://doi.org/10.1007/s10640-020-00436-x

Bovens, M., 't Hart, P., & Kuipers, S. (2008). The politics of policy evaluation. In M. Michael, R. Martin, & E. G. Robert (Eds.), *The Oxford handbook of public policy* (pp. 319–335). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199548453.003.0015

Bynum, C., Sze, C., Kearns, D., Polovick, B., & Simon, K. (2018). An examination of a voluntary policy model to effect behavioral change and influence interactions and decision making in the freight sector. *Transportation Research Part D: Transport and Environment*, *61*, 19–32. https://doi.org/10.1016/j.trd.2016.11.018

Canada's Ecofiscal Commission. (2016). *Course correction: It's time to rethink canadian biofuel policies*. https://ecofiscal.ca/wp-content/uploads/2016/10/Ecofiscal-Commission-Course-Correction-Biofuels-Report-October-2016.pdf

CE Delft, Ricardo-AEA, Ecologic Institute, E-Bridge, & REKK. (2015). *Mid-term evaluation of the renewable energy directive – a study in the context of the REFIT programme*. (Publication code: 15.3D59.28). https://www.ecologic.eu/sites/default/files/publication/2015/ce_delft_3d59_mid_term_evaluation_of_the_red_def-1.pdf

Cooksy, L. J., & Caracelli, V. J. (2005). Quality, context, and use: Issues in achieving the goals of metaevaluation. *American Journal of Evaluation*, *26*(1), 31–42. https://doi.org/10.1177/1098214004273252

Crabb, A., & Leroy, P. (2008). *The handbook of environmental policy evaluation*. Routledge.

Creutzig, F., McGlynn, E., Minx, J., & Edenhofer, O. (2011). Climate policies for road transport revisited (I): evaluation of the current framework. *Energy Policy*, *39*(5), 2396–2406. https://doi.org/10.1016/j.enpol.2011.01.062

Davies, R., & Payne, L. (2015). Evaluability assessments: Reflections on a review of the literature. *Evaluation*, *21*(2), 216–231. https://doi.org/10.1177/1356389015577465

EC. (2016). *Refit Ex-Post Evaluationof Combined Transport Directive 92/106/EEC*. Commission Staff Working Document SWD(2016) 141 final.

EC. (2017a). *Better regulation guidelines. Commission staff working document SWD(2017) 350 final*. European Commission.

EC. (2017b). *Better regulation toolbox*. European Commission. https://ec.europa.eu/info/sites/default/files/better-regulation-toolbox.pdf

EC. (2021). *Communication from the Commission to the European Parliament, the Council, the European economic and social committee and the committee of the regions – 'Fit for 55': delivering the EU's 2030 climate target on the way to climate neutrality*. COM(2021) 550 final. European Commission, Brussels.

EC. (2022). *European Commission - strategy - priorities 2019-2024 - a European green deal*. https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal

ECORYS Nederland BV. (2007). *Evaluation of the Marco Polo programme (2003-2006) framework contract for mid-term and ex-post evaluations* (Lot2- reference: TREN/A1/17-2003).

Elgie, S., & McClay, J. (2013). Policy commentary/commentaire BC's carbon tax shift is working well after four years (attention Ottawa). *Canadian Public Policy*, *39*(Supplement 2), S1–S10. https://doi.org/10.3138/CPP.39.Supplement2.S1

Europe Economics. (2011). *Evaluation of the Marco Polo programme 2003*-2010 *Final Report*. Europe Economics, London.

Fujiwara, N., van Asselt, H., Bößner, S., Voigt, S., Spyridaki, N.-A., Flamos, A., Alberola, E., Williges, K., Türk, A., & ten Donkelaar, M. (2019). The practice of climate change policy evaluations in the European union and its member states: Results from a meta-analysis. *Sustainable Earth*, *2*(1), 1–16. https://doi.org/10.1186/s42055-019-0015-8

Gokhale, P. (1997). Grey literature varieties—definitional problems. In *Third international conference on grey literature: perspectives on the design and transfer of scientific and technical information*, 13-14 November 1997, Luxemburg.

Harmelink, M., Nilsson, L., & Harmsen, R. (2008). Theory-based policy evaluation of 20 energy efficiency instruments. *Energy Efficiency*, *1*(2), 131–148. https://doi.org/10.1007/s12053-008-9007-9

Haug, C., Rayner, T., Jordan, A., Hildingsson, R., Stripple, J., Monni, S., Huitema, D., Massey, E., van Asselt, H., & Berkhout, F. (2010). Navigating the dilemmas of climate policy in Europe: Evidence from policy evaluation studies. *Climatic Change*, *101*(3), 427–445. https://doi.org/10.1007/s10584-009-9682-3

Hildén, M. (2011). The evolution of climate policies–the role of learning and evaluations. *Journal of Cleaner Production*, *19*(16), 1798–1811. https://doi.org/10.1016/j.jclepro.2011.05.004

Huitema, D., Jordan, A., Massey, E., Rayner, T., van Asselt, H., Haug, C., Hildingsson, R., Monni, S., & Stripple, J. (2011). The evaluation of climate policy: Theory and emerging practice in Europe. *Policy Sciences*, *44*(2), 179–198. https://doi.org/10.1007/s11077-011-9125-7

IEA. (2021a). *Data and statistics - CO_2 emissions by sector, World 1990-2019*. Retrieved February 17, 2022, from https://www.iea.org/data-and-statistics/data-browser/?country = WORLD&fuel = CO2%20emissions&indicator = CO2BySector

IEA. (2021b). *IEA - reports - tracking transport 2021*. Retrieved February 17, 2022, from https://www.iea.org/reports/tracking-transport-2021

IEA. (2022a). *Topics - transport*. Retrieved February 17, 2022, from https://www.iea.org/topics/transport

IEA. (2022b). *Policies – policies database*. Retrieved March 18, 2022, from https://www.iea.org/policies

IPCC. (2021). *Climate change 2021: The physical science basis. Contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change* [Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)]. Cambridge University Press.

IPCC. (2022). *Summary for policymakers* [H.-O. Pörtner, D.C. Roberts, E.S. Poloczanska, K. Mintenbeck, M. Tignor, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem (eds.)]. In: Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, B. Rama (eds.)]. Cambridge University Press. In Press.

ITF. (2019). *ITF transport outlook 2019*. OECD Publishing. https://doi.org/10.1787/transp_outlook-en-2019-en

ITF. (2022). *Mode choice in freight transport. ITF research reports*. OECD Publishing.

Mastenbroek, E., van Voorst, S., & Meuwese, A. (2016). Closing the regulatory cycle? A meta evaluation of ex-post legislative evaluations by the European Commission. *Journal of European Public Policy*, *23*(9), 1329–1348. https://doi.org/10.1080/13501763.2015.1076874

McKinnon, A. C. (2005). The economic and environmental benefits of increasing maximum truck weight: The British experience. *Transportation Research Part D: Transport and Environment*, *10*(1), 77–95. https://doi.org/10.1016/j.trd.2004.09.006

Michaelowa, A., Allen, M., & Sha, F. (2018). Policy instruments for limiting global temperature rise to 1.5° C–can humanity rise to the challenge? *Climate Policy*, *18*(3), 275–286. https://doi.org/10.1080/14693062.2018.1426977

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ*, *339*(1), b2535. https://doi.org/10.1136/bmj.b2535

Navius Research. (2020). *Biofuels in Canada 2020 - tracking biofuel consumption, feedstocks and avoided greenhouse gas emissions*. Navius Research, Vancouver.

OECD. (2021). *Applying evaluation criteria thoughtfully*. OECD Publishing. https://doi.org/10.1787/543e84ed-en

Palander, T. (2017). The environmental emission efficiency of larger and heavier vehicles–a case study of road transportation in Finnish forest industry. *Journal of Cleaner Production*, *155*, 57–62. https://doi.org/10.1016/j.jclepro.2016.09.095

Ricardo Energy & Environment, & TEPR. (2015). *Ex-post evaluation of directive 2009/33/EC on the promotion of clean and energy efficient road transport vehicles: final report*. European Commission, Brussels.

Schoenefeld, J., & Jordan, A. (2017). Governing policy evaluation? Towards a new typology. *Evaluation*, *23*(3), 274–293. https://doi.org/10.1177/1356389017715366

Shmelev, S. E., & Speck, S. U. (2018). Green fiscal reform in Sweden: Econometric assessment of the carbon and energy taxation scheme. *Renewable and Sustainable Energy Reviews*, *90*, 969–981. https://doi.org/10.1016/j.rser.2018.03.032

Takman, J., & Gonzalez-Aregall, M. (2021). A review of public policy instruments to promote freight modal shift in Europe: Evidence from evaluations. Working Paper 2021:6, Swedish National Road & Transport Research Institute (VTI).

Takman, J., Trosvik, L., & Vierth, I. (2020). *Triple F etableringsprojekt - Omvärldsanalys Policy*. Triple F Rapport nummer: 2020.2.13.

Touratier-Muller, N., Machat, K., & Jaussaud, J. (2019). Impact of French governmental policies to reduce freight transportation CO2 emissions on small-and medium-sized companies. *Journal of Cleaner Production*, *215*, 721–729. https://doi.org/10.1016/j.jclepro.2019.01.052

Transport & Mobility Leuven, CE Delft, TRT, & TNO. (2014). *Evaluation study on speed limitation devices: Final report*. Publications Office of the European Union.

Tsafnat, G., Glasziou, P., Choong, M. K., Dunn, A., Galgani, F., & Coiera, E. (2014). Systematic review automation technologies. *Systematic Reviews*, *3*(1), 1–15. https://doi.org/10.1186/2046-4053-3-74

Turnpenny, J., Radaelli, C. M., Jordan, A., & Jacob, K. (2009). The policy and politics of policy appraisal: Emerging trends and new directions. *Journal of European Public Policy*, *16*(4), 640–653. https://doi.org/10.1080/13501760902872783

UN. (2022). United Nations framework convention on climate change - process and meetings - the Paris Agreement. Retrieved February 18, 2022, from https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement

Vedung, E. (2017). *Public policy and program evaluation*. Routledge.

Yusuf, N. (2018). The impact of freight vehicle access restriction on the sustainability of Jakarta intra urban tollway system. *Planning Malaysia*, *16*, 50–62. https://doi.org/10.21837/pm.v16i5.410