Evaluation and Program Planning 69 (2018) 109–117

Contents lists available at ScienceDirect

Evaluation and Program Planning

journal homepage: www.elsevier.com/locate/evalprogplan

Public project success as seen in a broad perspective. Lessons from a meta-evaluation of 20 infrastructure projects in Norway

Gro Holst Volden

Norwegian University of Science and Technology, 7491 Trondheim, Norway

ARTICLE INFO

Keywords:
Project evaluation
Project success
Evaluation framework
OECD-DAC criteria

ABSTRACT

Infrastructure projects in developed countries are rarely evaluated ex-post. Despite their number and scope, our knowledge about their various impacts is surprisingly limited. The paper argues that such projects must be assessed in a broad perspective that includes both operational, tactical and strategic aspects, and unintended as well as intended effects. A generic six-criteria evaluation framework is suggested, inspired by a framework frequently used to evaluate development assistance projects. It is tested on 20 Norwegian projects from various sectors (transport, defence, ICT, buildings). The results indicate that the majority of projects were successful, especially in operational terms, possibly because they underwent external quality assurance up-front. It is argued that applying this type of standardized framework provides a good basis for comparison and learning across sectors. It is suggested that evaluations should be conducted with the aim of promoting accountability, building knowledge about infrastructure projects, and continuously improve the tools, methods and governance arrangements used in the front-end of project development.

1. Introduction

The purpose of this study is (1) to demonstrate the importance of and need for a broad evaluation approach to measure success in large infrastructure projects, and (2) to test an evaluation methodology that is commonly applied in projects and undertakings in low-income countries, but now on projects in a more complex context in a high-income country.

1.1. Broad evaluation of public projects

Governments in high-income countries invest vast amounts of funds each year in infrastructure and other large public projects, such as roads and railways, public buildings, defence acquisitions and ICT infrastructure. The number and scale of such projects grow over time (Flyvbjerg, 2014). Even in a small country such as Norway, annual investments in large public projects amount to USD 6 billion per year not including petroleum sector investments (Norwegian Ministry of Finance, 2015).

Samset (2003) argues that in order to be truly successful, public investment projects must not only perform well operationally, but also tactically and strategically. Correspondingly, Baccarini (1999) defines two levels of project success, i.e. project management success (which concerns delivery), and product success (which concerns the outcome). However, whereas operational project success is highlighted by practitioners as well as academics (the problem of cost overruns being particularly well documented in the literature, cf. Morris & Hough, 1991; Flyvbjerg, Skamris Holm, & Buhl, 2003; van Wee, 2007), tactical and strategic success is often ignored, possibly because it challenges the way analysts think and has political aspects to it (Samset & Christensen, 2017).

Although Norway, as many other OECD countries, has been assigned a high level of evaluation maturity in national government (Jacob, Speer, & Furubo, 2015), systematic evaluations of public investment projects with respect to their outcomes are rarely conducted (Samset & Christensen, 2017; Rambøll & Agenda Kaupang, 2016). One exception is the transport sector, where benefit-cost analyses are performed to documents the projects' value-for-money (not so much ex-post, but before projects are submitted for government approval). However, many authors argue that benefit-cost efficiency is too narrow as measure of projects' tactical and strategic success (House, 2000; Heinzerling & Ackerman, 2002). This view is supported by the fact that benefit-cost efficiency rarely affects the priority ranking of road projects in Norway, which implies that decision-makers pursue other goals (Nyborg, 1998; Eliasson, Börjesson, Odeck, & Welde, 2015). Project success is clearly multi-faceted, and an evaluation can only be relevant to various stakeholders if it comprises a broader set of criteria.

This paper presents a generic framework for broad evaluations of large public projects. It is inspired by the criteria recommended by the Organisation for Economic Cooperation and Development's

E-mail address: gro.holst.volden@ntnu.no.

https://doi.org/10.1016/j.evalprogplan.2018.04.008
Received 22 September 2017; Received in revised form 14 March 2018; Accepted 29 April 2018
Available online 30 April 2018

Development Assistance Committee (OECD-DAC), which are much used in development assistance. The present study aims to demonstrate that the criteria are well-suited for infrastructure projects in industrial countries as well.

1.2. The case of Norway

Several authors have highlighted the crucial role of the front-end phase of projects (Williams & Samset, 2010; Morris, 2013; Samset & Volden, 2015). Many aspects that later create problems are typically present already at the project definition stage. In public projects, the Government as ultimate project owner should ensure the necessary quality-at-entry of project proposals and plans. This was done in Norway year 2000, when a scheme requiring external quality assurance of the decision basis was introduced for projects with an estimated investment cost exceeding USD 90 million. The scheme includes: (1) quality assurance of the choice of concept before the Cabinet decision to start a pre-project, and (2) subsequent quality assurance of the project management basis and cost estimate before the project is submitted to Parliament for approval and funding. Quality assurance is performed by external experts that are pre-qualified by the Ministry of Finance (Volden & Samset, 2017).

As of today more than 200 projects have been subject to quality assurance up-front, of which some 90 have so far been completed. There is evidence that this has improved the Norwegian Government's basis for decisions regarding major public investments (Kvalheim, Christensen, Samset, & Volden, 2015) and that the projects keep within their budgets (Welde, 2017). Nevertheless, the projects should also be evaluated ex-post, to verify how they actually perform in a broad perspective. In this study we test the suggested OECD-DAC evaluation framework on 20 Norwegian projects that were quality-assured in their front-end phase. The findings regarding these projects' performance are interesting in themselves, but the main purpose of this paper is to discuss the experiences with the evaluation framework and the evaluation process, as basis for improving and consolidating them.

The framework was first tested on four projects and the results presented to the Norwegian Ministry of Finance. The subsequent 16 evaluations included in this study were conducted on request from the Ministry. Ex-post evaluation has thus already become an integrated part of the project governance scheme and is likely to be used to further improve the quality assurance scheme. With time, a database is built, which allows for quantitative analyses of success at different levels across sectors and project types.

The paper starts with a presentation of the theoretical framing and the chosen evaluation framework. Then we present our methodology and data (the 20 projects), before we provide a synthesis of the findings in terms of the projects' success on various levels, and a discussion of the experiences with the evaluation framework and how it has been applied. Finally we offer some conclusions and discuss future extensions of the study.

2. Theoretical framing

Evaluation is the systematic investigation of the effectiveness of a project or other intervention. An evaluation requires evaluation expertise and rigorous application of scientific methods, while at the same time being focused on solving practical problems and being useful to project sponsors, decision-makers and other stakeholders (Rossi, Lipsey, & Freeman, 2004).

Evaluation became particularly relevant in the U.S. in the 1960s associated with the Kennedy and Johnson administrations and the social programs implemented at the time. Its aim was to learn from successes and failures and improve forward planning. It spread subsequently to other countries and different sectors, particularly to international development aid, where the effectiveness of investments and policy was contested.

Evaluation may be conducted at different stages during a project's lifetime. Each stage raises different questions to be answered, and correspondingly different evaluation approaches are needed. This would involve the assessment of i) the need for the project, ii) project design and logic/theory, iii) the implementation of the project, iv) its outcome or impact (i.e., what it has actually achieved), and v) its cost and efficiency (Rossi et al., 2004).

All projects are explicitly or implicitly based on an assumed set of causal relationships between inputs, project activities, outputs, and outcome. Several authors argue the merits of using this so-called logic model (McLaughlin & Jordan, 1999; Samset, 2003), also referred to as the programme theory (Chen, 1990; Weiss, 1997; Rogers, Petrosino, Huebner, & Hacsi, 2000) as representation of the project to help visualize important aspects, and especially when preparing for an evaluation. It helps clarify for all stakeholders: the definition of the problem, the overarching goals, and the capacity and outputs of the project (McLaughlin & Jordan, 1999). Further, looking at the different components of a project in relation to the overall societal objective, it allows for illumination of potential misalignments. Experience has shown that projects' logic is often unclear (Karlsen & Jentoft, 2013) and that goal hierarchies are characterized by a range of errors (Samset, Andersen, & Austeng, 2014). A critical assessment of the project's logic model might enable the evaluator to reveal a weak or faulty logic before any empirical evidence has been gathered (Brousselle & Champagne, 2011). In recent years, new versions of theory-based evaluation have emerged, such as realistic evaluation (Pawson & Tilley, 1997) and the theory of change (ToC) (Connell & Kubisch, 1998; Sullivan & Stewart, 2006).

3. A six-criteria evaluation framework

The chosen evaluation framework in the Norwegian context is presented below. As a general requirement, an overall evaluation framework ought to measure the success of projects in a broad perspective. It should be flexible enough to accommodate all types of projects, and sufficiently standardized to allow for comparisons between projects.

The starting point of each project evaluation should be the mapping of the logic model. The logical framework methodology is used, which focuses on the hierarchy of agreed goals, and identifies external risks on each level (Samset, 2003). The methodology was originally developed for USAID (Rosenberg, Posner, & Hanley, 1970), but its use spread rapidly by the UN, to aid administrations in a number of countries, and later to the OECD and the EU Commission.

In the Norwegian quality-assured projects, a logic model in the form of a goal hierarchy already exists, but it must be checked for consistency, and if necessary upgraded by the evaluator. When possible, the evaluator should also thoroughly investigate the goodness of the underlying theories (i.e. apply a truly theory-based approach), using existing literature and expert statements. The resulting model should be on the form illustrated in Fig. 1.

The overall evaluation criteria should be developed from the logic model. Since projects are de facto established to fulfil a certain purpose (Project Management Institute, 2013), one must ask *whether the intended*

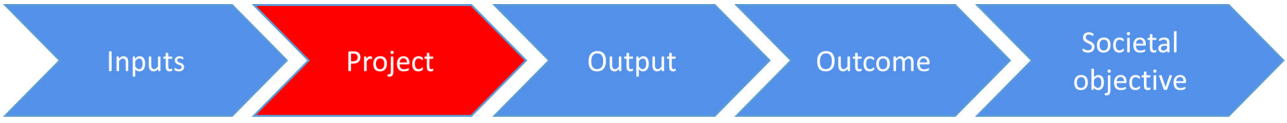

Fig. 1. The logic model for a project.

results have been attained. Operational, tactical as well as strategic goal achievement should be assessed (i.e. output, outcome and societal objective). Furthermore, one should study any *side effects* that can be attributed to the project, i.e. impacts beyond those defined by the project owner (Sartorius, 1991; Gasper, 2000; Bakewell & Garbutt, 2005). Finally, it is important to note that public funding is a limited resource, and we must ensure that the funds are spent wisely. Therefore, regardless of whether efficiency and value-for-money are stated as project goals, an evaluation should always ask about this.

A standardized set of five evaluation criteria is much used, by the UN and other institutions and aid organizations, and has been endorsed by the OECD-DAC (OECD, 1991, 2002). Evaluation according to this framework highlights i) the need for the project (relevance), ii) whether the uses of resources and time are reasonable (efficiency), iii) whether stated goals are achieved (effectiveness), iv) what other positive or negative effects may occur because of the project (other impacts), and v) whether the positive effects persist after the conclusion of the project (sustainability).

As noted by Picciotto (2013), development projects are not so different from projects in developed countries. The five criteria reflect hard-won lessons of experience, and have by and large replaced prior approaches that focused only on inputs and outputs. They can be used equally at project, programme and policy level, and are aligned with the results-oriented stance favoured by most countries. Other sectors have introduced variants of the criteria (see, for example, European Commission (2013) concerning socio-economic development in Europe; ALNAP (2006) concerning humanitarian projects; and European Commission (2015) concerning regulations). A thorough review of the five criteria, which was performed by a group of professional evaluators (Chianza, 2008), concluded that the criteria work well and in particular that they satisfy Michael Scriven's Key Evaluation Checklist for programme evaluation (see Scriven, 2015 for the most recent version). However, Chianza suggested some improvements, the most important being to widen the interpretation of some criteria, such as relevance and sustainability not only covering the project owner perspective, and to define efficiency not only in the narrow sense (cost and time efficient delivery of the project), but also in terms of value-for-money. We agree with Chianza, and as a consequence have chosen to include benefit-cost efficiency as a sixth criterion of the model. For such economic analyses, we follow the standard method, as presented by, for example, Boardman, Greenberg, Vining, & Weimer, 2011. An implication is that the efficiency criterion focuses only on cost and time efficient project delivery.

Our definitions of the six criteria, and the level of success which they represent, are presented in Table 1. Their relation to the logic model is illustrated in Fig. 2.

The purpose of our evaluations is to give an *overall* picture of public project success. With budget limitations, we cannot be too ambitious regarding the methodological rigour when responding to each criterion. Experimental designs are rarely realistic for any of the evaluation criteria, and certainly not for the strategic ones. Rather, we must use simplified evaluation designs, economic data collection methods, and triangulation between various data sources and methods of analysis, quantitative as well as qualitative, to ensure validity and reliability. This is discussed further in the next section.

4. Research setting and research methodology

4.1. The projects

This study regards the evaluation of Norwegian public investment projects that have been subject to formal quality assurance before being approved for funding. In total, more than 200 major projects have been through the government's quality-at-entry scheme since the year 2000, representing primarily roads (53%), buildings (18%), railway (9%), ICT (11%) and defence (9%). Since the subsequent detailed planning and implementation period of such large projects is extensive, only 90 projects have been completed so far. Of these, 40 have been in operation for at least five years, and thus considered ready for evaluation.

A total of 20 projects was evaluated during 2012–2017 and are included in this meta-evaluation: eight road projects, five buildings, three railway projects, two ICT projects, and two defence projects (Table 2). The projects were chosen in chronological order, and constitute a relatively representative picture of quality-assured projects in their operational phase (50%). In addition to the sample projects, Table 2 shows which evaluators were involved. They represent consultancies in Norway and Sweden, and researchers from the Concept Research Programme, all considered independent of the projects and the implementing agencies.

4.2. The evaluation process

The six criteria framework does not guarantee high-quality evaluation by titself. In addition, evaluation skills, independent evaluators, appropriate data collection and analysis methods, etc. is required.

Each evaluation followed a defined process, which consisted of six steps, based on Samset's (2003) project evaluation textbook and also aligned with Michael Scriven's Key Evaluation Checklist (2015). In *Step 1* the Concept Research Programme selected the project to be evaluated, and sought acceptance from the responsible ministry (e.g. the Ministry of Transportation in the case of road projects). The ministries could, in principle, decline, but none of them did. A contact person in the ministry (and its subordinate agency when relevant) was identified.

In *Step 2* the evaluation team was established, usually following a public call. Researchers participated in some evaluation teams in order to gather experience in the use of the model. The team consisted of three or four people, all with good evaluation skills and knowledge of the sector. The scope was set to approximately three person-months of work per evaluation, depending somewhat on the project's complexity and availability of data.

Table 1
Definitions of the six evaluation criteria.

Level of success	Evaluation criterion	Definition
Operational	Efficiency	This criterion concerns project implementation and outputs in terms of cost, time and quality, and how economically the project organization has converted inputs into outputs.
Tactical	Effectiveness	This concerns whether the agreed outcome has been obtained and to what extent the project has contributed to this outcome.
Strategic	Other impacts	This includes all consequences beyond the agreed outcome (i.e. side effects) that can be attributed as the result of the project, positive and negative, short term and long term, for different stakeholders.
	Relevance	A project is relevant if there is a need for what the project delivers. Project relevance is measured in relation to national political priorities, but also stakeholders' preferences. It is essential to bring conflicts of interest to light as part of the evaluation.
	Sustainability	A project is sustainable if its benefits are likely to persist throughout its lifetime. This usually requires that the total impacts (financial, environmental and social) are acceptable in the long run.
	Benefit-cost efficiency	This should be measured in terms of total willingness to pay in relation to cost, or secondarily in terms of outcome in relation to cost (i.e. cost-effectiveness).

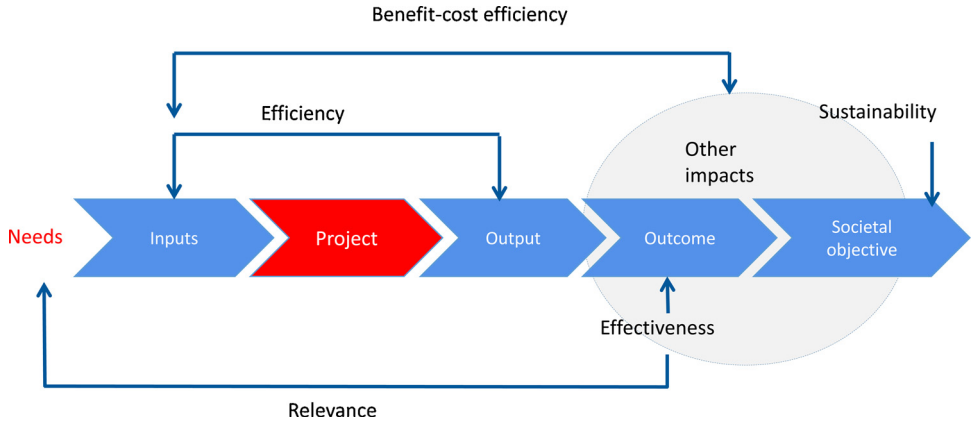

Fig. 2. The six evaluation criteria shown in relation to the logic model.

In *Step 3* the evaluation team reviewed and, if necessary, adjusted the logic model. Then it operationalized each of the six criteria by selecting more specific indicators or evaluation questions to be answered.

In *Step 4* the evaluation was carried out by collecting and analysing data, and answering the evaluation questions by combining different data sources and methods. We leaned on a number of authors who have suggested the mixing of methods to improve rigour in evaluations of complex interventions (see for example Ton, 2012; Green et al., 2017; Yin, 2013a,b), and the use of so-called rapid evaluation methods when faced with restricted budgets and timelines (World Bank, 2004; Bamberger, Rugh, Church, and Fort, 2004; Samset, 2003). As mentioned above, the use of existing literature and expert statements to assess the goodness of the programme theory was used as a supplementary approach to strengthen the validity of findings. Detecting impacts beyond the intended effects normally requires a wide, inductive and multidisciplinary approach.

In *Step 5* the evaluation team summarized its assessment for each criterion by setting a score between 1 and 6, where 1 is failure and 6 is highly successful. Score 4 should be awarded when the result for the relevant criterion is acceptable, but not an over-achievement. An overall guideline for score-setting was prepared in advance, to assist the evaluation teams.

Step 6. The final report was made public and distributed. The report and key results were stored in a database that is openly available to the public, (www.ntnu.edu/concept/evaluation-reports). The ministry and responsible agency were encouraged to follow up the results internally.

4.3. Meta-evaluation

This paper represents a meta-evaluation based on the findings and lessons learned from the 20 first evaluations. The term meta-evaluation is ambiguous. Generally, it implies that original analyses of data become the objects of a new analysis on a higher level (Glass, 1976). The much used UK Magenta Book (HM Treasury, 2011) primarily refers to meta-evaluation as a synthesis of a number of related evaluations, with the purpose of providing some estimate of the average or combined effect. This interpretation is close to what Yin (2013a) defines as cross-case synthesis. On the other hand, Scriven (2015) refers to meta-evaluation as an evaluation of one or more evaluations in order to identify their strengths, limitations and other uses, against a set of quality standards. A similar interpretation is suggested by Stufflebeam (2010) who distinguishes between three groups of standards: technical adequacy, utility and cost-effectiveness.

The present study applies the OECD's definition which includes both the above-mentioned interpretations: meta-evaluations are here defined as "*evaluations designed to aggregate findings from a series of evaluations. It can also be used to denote the evaluation of an evaluation to judge its quality and/or assess the performance of the evaluators*" (OECD, 2002, p. 26, our underlinings).

First, we present an aggregation and synthesis of the findings from the 20 separate evaluations, to establish the success of Norwegian investment projects. Second, we evaluate the evaluations themselves, including the suitability of the methodological framework.

Table 2
Key information relating to the sample projects. Sorted according to the year of evaluation.

No.	Project name	Sector	Start (yr)	End (yr)	Eval. (yr)	Evaluator
1	Customs area, Svinesund	Building	2004	2005	2012	SINTEF; Concept
2	Asker–Sandvika	Rail	2001	2005	2012	VTI
3	E18 Momarken–Sekkelsten	Road	2005	2007	2012	Concept
4	Skjold class MTB	Defence	2003	2013	2012	Scanteam; Concept
5	E6 Riksgr.–Svingenskogen	Road	2002	2004	2014	COWI
6	Svalbard Research Park	Building	2003	2005	2014	Concept
7	Lofoten fixed link	Road	2003	2007	2014	UiN; Nordlandsforskning AS
8	Eiksund fixed link	Road	2003	2008	2014	Menon
9	NAV ICT Basis	ICT	2006	2010	2014	NIBR
10	Østfold University College	Building	2003	2006	2015	SINTEF; Concept
11	E16 Kløfta–Nybakk	Road	2005	2007	2015	Urbanet
12	Rv 519 Finnfast fixed link	Road	2005	2007	2015	Menon
13	Sandnes–Stavanger	Rail	2005	2009	2015	Oslo Economics; Atkins
14	Military area. Østlandet	Defence	2002	2012	2015	Prokonsult AS
15	Perform	ICT	2008	2012	2015	Menon; Vivento
16	Halden Prison	Building	2006	2010	2016	Oslo Economics; Tyrilistiftelsen; Sweco
17	New Opera House. Oslo	Building	2005	2008	2016	HR Prosjekt
18	E6 Svingenskogen–Åsgård	Road	2005	2008	2016	Menon; Concept
19	E6 Åsgard–Halmstad	Road	2004	2005	2016	Menon; Concept
20	Gevingåsen railway tunnel	Rail	2009	2011	2017	Concept; SINTEF

The main data source for the first part was the 20 evaluation reports. We coded and summarized the assessments done by the evaluation teams, based on a set of questions prepared for the study. It included, for each evaluation criterion, the overall score as well as a range of more detailed indicators (e.g. for efficiency, it involved time, cost and quality, respectively). Different measurement scales were used for different aspects, including inter alia the number scale used for score-setting, binary variables (achieved/not achieved, etc.) and qualitative descriptions. Accordingly, aggregation of findings across projects was partly quantitative and partly qualitative. The scores awarded on each criterion were particularly useful since they allowed us to compute averages and see how they differed across sectors (although the number of projects was too small for statistical testing).

The second part of the meta-evaluation consisted of registration and assessments of various quality aspects of the evaluations themselves as well as the evaluation teams. Here we used a separate set of questions which was based on common advice from the project management and evaluation literature, and more specific recommendations concerning the OECD-DAC criteria. The questions included, inter alia, whether the criteria were interpreted and applied as intended, rigour in design, triangulation, adequate and efficient use of data sources, unbiased score-setting, use of a benchmark, and a sufficiently broad (multi-disciplinary) evaluation team. Ministries' and agencies' actual use of the findings in their planning of new projects was not included. Again, different measurement scales were used, many of them binary or qualitative. Our assessments were unavoidably subjective, but meticulously explained and documented.

Furthermore, a focus group meeting was held with 11 experienced evaluators, all of whom had participated in one or more evaluations. The group was confronted with preliminary findings and assessments, and the participants were given the opportunity to comment on and share their experiences and assessments of the model, the process and the need for guidance. The focus group meeting largely confirmed the picture presented by us, but assessments and conclusions were updated on some points.

Finally, it should be noted that the author also drew on her own experiences from the process of commissioning and following-up of the 20 evaluations, as well as participating in three of them.

5. Findings: public project success

In this section, we present and discuss the aggregated findings.

Table 3 summarizes the scoring results from the 20 evaluations. The overall picture is very positive, with average scores between 4 and 5 for all criteria, and the highest scores for efficiency and effectiveness.

Efficiency concerns the project's operational success. As shown in Table 3, 19 out of 20 projects scored 4 or better on this criterion. In total, 15 projects were completed with a final cost below the approved cost frame, 15 were completed within the time frame, and 16 were considered to meet high quality standards. The projects were also largely well organized and managed. The findings relating to cost control have been confirmed in a study that included 78 completed projects that had been through external quality assurance (Welde, 2017). In some projects, deviations from the cost frame were considerable, in both positive and negative direction. In some of these, the evaluator suggested that the cost frame was not realistic (typically because market uncertainty was underestimated). This indicates that the quality-at-entry scheme may not serve as guarantee for realistic budgets.

With regard to tactical project success, as measured by *effectiveness*, 18 projects scored 4 or better, which means that most of the projects' outcomes were in accordance with plans. Three projects received top scores, i.e. two road projects that realized very large time savings and reductions in accident levels, and one ICT project that generated major benefits in terms of time savings for the agency and improved quality of the services for the users.

The other four evaluation criteria express strategic project success. The evaluators concluded that the projects performed acceptably in these dimensions too. All 20 scored 4 or better on at least one strategic criterion.18 projects scored 4 or better on *other impacts*. Negative side effects were identified in few cases, and several projects generated positive side effects. One project, the New Opera House in Oslo, received top score for its very positive contribution to urban development, while some projects could have done more to avoid negative side effects. 17 projects scored 4 or better on *relevance*. This implies that most were considered solutions to real problems. However, some involved conflicts of interest (i.e. they were not equally relevant in all perspectives). This was the case in regards the Lofoten fixed link, a road project that connected a remote region to areas that were more urban, but left the neighbouring remote region even more isolated. 19 out of 20 projects scored 4 or better on *sustainability*. This implies that project benefits were largely expected to continue in a sustained number of years. However, the projects had to be sustainable in *all* aspects (i.e. financial, environmental and social) to be assigned a top score. For example, one project (defence) scored very low, because growing operational and maintenance costs had made it financially unsustainable. The projects scored slightly lower on *benefit-cost efficiency*. In total, 13 out of 20 projects scored 4 or better. The five most profitable projects were all road projects in urban areas.

In summary, most of the projects were considered successful in more than one aspect, and especially in operational terms. There appears to be some correlation between the scores for the various criteria. This is not surprising, since a well-thought-out and carefully planned project will normally be successful in several respects. However, there may also be conflicts, for example when some of the projects scored high on relevance and sustainability, and lower on benefit-cost efficiency. All three railway projects were well aligned with the government's strategy for sustainable transport. But with passenger numbers that were lower than estimated by the time of evaluation, and a relatively high capital cost, the value-for-money was considered low. We agree that not all projects can or should be 'profitable', but one should at least consider whether a simpler solution, still with acceptable goal achievement, would substantially improve value-for-money. Similarly, a project with high effectiveness but negative side effects should perhaps have been redesigned, such that it got a better overall score.

One interesting observation is that many of the projects that scored high in tactical and strategic terms were not aimed at specific stakeholder groups or regions, but rather followed from national political objectives. This supports earlier findings that when specific stakeholder groups manage to mobilize government funding for 'their' project, the project may turn out to be less relevant from the perspective of the wider society − a phenomenon known as perverse incentives in project

Table 3
Evaluation results (N = 20).

	Efficiency	Effectiveness	Other impacts	Relevance	Sustainability	Benefit-cost eff.
Average score	4.7	4.7	4.4	4.6	4.6	4.2
Median score	5	5	4	5	5	4
Min.–Max. score	3–6	3–6	3–6	3–6	2–6	2–6
4 or better (no. of projects)	19	18	18	17	19	13
5 or better (no. of projects)	13	12	8	13	13	9

Table 4
Evaluation results: average per sector.

	No. of projects	Efficiency	Effectiveness	Other impacts	Relevance	Sustainability	Benefit-cost eff.
Building	5	5.4	4.2	4.6	4.6	4.8	3.8
Defence	2	4.5	4.5	4.5	4.5	3.5	3.5
ICT	2	5.0	5.5	4.5	4.0	5.5	4.0
Railway	3	4.3	3.3	4.0	4.7	4.7	2.7
Road	8	4.4	5.3	4.3	4.6	4.5	5.3

selection (see Volden & Samset, 2015).

Although the number of evaluated projects is low, the results indicate some interesting sectoral differences, as shown in Table 4. In the following we will briefly comment on the three largest project categories (i.e., including at least three projects). The *building projects* performed excellently in operational terms, implying that they were delivered within time and budget and with a high quality. Some of these buildings were awarded architectural prizes. However, they scored slightly lower tactically and strategically, some of them with outcomes that had still not been realized several years after completion. Some of the projects had ambitious goals, such as to 'co-locate departments A and B and realize professional synergies', and should have devoted more attention to the fact that it takes more than a building to obtain such goals. The *railway projects* were closely aligned with government strategies for a 'green shift' in transport, and thus were considered relevant and sustainable. However, they scored low on benefit-cost efficiency and on effectiveness. As for building projects, ambitious goals usually require more than the physical infrastructure, and this should have been given more attention than was done in the sampled projects. The *road projects* scored high on most criteria, but somewhat lower on efficiency and other impacts. Road projects experienced the biggest cost overruns, but also the biggest cost underruns, implying that the Public Roads Agency should make efforts to obtain more accurate estimates. And, that more attention should be paid to side effects in the planning phase.

Overall, the findings concerning project success are very positive, which seems to conflict the public discourse and studies that demonstrate a low level of success in public projects. For instance, Flyvbjerg, Garbuio, and Lovallo, (2009) used the expression "*over budget, over time, over and over again*" and explained the widespread problem of cost overruns by delusion and deception. We think that caution should be made when referring to public projects as generally unsuccessful. Firstly, the media as well as the academic literature have mostly been concerned with cost control, which is only one aspect of project success. Secondly, as noted by Love, Smith, Simpson, Regan and Olatunji, (2015), different empirical studies of cost control come to very different conclusions, depending, inter alia, on the point of reference from which a cost overrun is measured (those that find the largest overruns typically compare with early and uncertain estimates). That said, we believe that the 20 Norwegian projects stand out as successful, which can be explained, at least to some extent, by the quality-at-entry requirements which ensure that they are thoroughly planned and reviewed before being submitted to Parliament for approval and funding. It should be noted that the remedy suggested by Flyvbjerg et al. (2009) to avoid delusion and deception, was to take an *outside view* on project planning and estimation, which is exactly what the quality-at-entry scheme does.

On the other hand we cannot eliminate the risk that some scores are positively biased. Wiig and Holm-Hansen (2014) found that *positive* side effects were mentioned more often than negative ones in evaluations of development assistance projects. They suggested that evaluators may be reluctant to criticise without hard evidence on which to base their criticisms, but willing to mention a positive issue that has the same level of uncertainty. While acknowledging the general risk of positive bias, we believe it is moderate to low in the 20 evaluations. The main reason is that the evaluations were organized by a third party, and all the teams were entirely independent of the projects.

However, another and more pertinent matter is how to ensure that the scores are well calibrated across projects. The scores were set by a different evaluation team in each case, based on the team's subjective assessments. As will be discussed in the next section, we believe there is a need for clearer guidelines for score-setting. In particular, the level of ambition inherent in the goal hierarchy must be taken into account when deciding on the score for effectiveness. We suspect that different levels of ambition related to project goals may explain some of the sectoral differences when it comes to effectiveness.

The findings of the evaluations should be useful for the purpose of accountability as well as for learning and improvement. Although each evaluation was limited in time and resources, it may identify major risks and problems that should be examined in more detail by the responsible ministry. Furthermore, the findings should provide input to the appraisal and planning processes of future investment projects funded by national government. It should be noted again that the sample of projects is not statistically significant, thus any attempts to generalize findings should be regarded as preliminary and tentative. Over time, when the database of evaluated projects is larger, it should be examined whether the patterns described above still hold.

6. Lessons learned about the evaluation framework and procedure

In the second part of the meta-evaluation, we look at the evaluations themselves, at how they were implemented, and at the framework's suitability for the purpose. The 20 evaluations have provided useful experiences and a basis for consolidating the model and practice.

6.1. General experiences of the model

The evaluators agreed that the chosen framework worked well, and that the six criteria covered the main aspects of public project success. Some noted that the strategic criteria, other impacts, relevance and sustainability, were 'eye-openers'. Knowing that pure economic evaluations are often considered by decision-makers to be too narrow (Nyborg, 1998), our evaluators agreed that the six-criteria framework should be more relevant.

The process of disaggregating the criteria into specific indicators and then aggregating the findings to provide answers to each criterion, provides a good balance between the need for standardization and flexibility. The evaluations have converged more and more into a common form, and their quality has improved over time.

However, we also see some challenges related to the interpretation of other impacts, relevance and sustainability. Some evaluators treated these strategic aspects superficially when realizing that they could not be measured in quantitative terms. Others interpreted them too narrowly, in the same way as found by Chianza (2008). Others, still, confused them with benefit-cost efficiency, downplaying, for example, environmental, social and ethical concerns. The explanation may be that many of the evaluation teams were dominated by economists. The evaluators confirmed that they were uncertain and wanted more guidance on the interpretation of these criteria.

6.2. Methodological rigour versus available resources

The evaluations consist of six criteria, each of which requires proper treatment. At the same time, the budget and time available implies clear limitations with regard to scope and methodological rigour. A timely question is therefore whether it is possible to give a professionally sound assessment of all six criteria. Experience so far suggests that the answer is yes. The 20 evaluations are largely of an acceptable quality. Admittedly, they are 'rapid' and the scores are sometimes uncertain, but this is not uncommon in evaluations, and must be accepted as long as the choice of methods and limitations are communicated. These findings support those of Samset (2003), Bamberger et al. (2004) and others who have argued that it is possible to conduct evaluations of acceptable quality under budget, time and data constraints.

Furthermore, the framework is flexible, and in individual cases, evaluators may spend more resources on a particular criterion, while treating other criteria more superficially. That has happened, for example, in cases of large deviations from the cost frame, which implied a need to look more closely into efficiency.

6.3. Methods for data collection and analysis

Different evaluation teams mostly chose the same methods for data collection and of analysis. For efficiency, they typically used data from project reports, interviews and benchmarking of cost data with similar projects. For effectiveness, they used time-series data for outcome indicators (often including comparison groups, such as similar geographic regions without the new infrastructure) and interviews with a wide range of stakeholders. For the strategic criteria, the evaluators used a combination of different sources, predominantly qualitative ones. For benefit-cost efficiency, they used all existing data and a set of assumptions and price tags. All evaluations included site visits.

It is costly to collect primary data, and evaluators must therefore prioritize carefully. Few of the evaluators for the studied 20 projects had done extensive outcome evaluations including control groups. Generally, they had chosen simple and informal methods.

The quality of the evaluations rests strongly on the ability to use a broad approach with a wide range of sources and methods. Most the evaluators did use triangulation to an acceptable extent, but some focused too much on quantitative data and the experiment as the gold standard. For example, one evaluation report devoted more space to discussing the difficulties of quantifying benefit-cost efficiency than to describing it with alternative data.

6.4. The project logic and uncritical evaluators

Reference data, in terms of descriptions of the goal hierarchy or logic model, existed in all projects, but often had weaknesses, which is not uncommon in project evaluations (cf. Samset et al., 2014). The quality-at-entry scheme requires each project to have a defined goal hierarchy. Despite this, more often than not, there were problems such as a missing causal logic or the wrong level of ambition (too high or too low). The evaluators handled this problem in quite different ways. Some re-established the logic model *as it should be*, as basis for their evaluation, while others made only minor or no adjustments. Although evaluators should not 'overrule' the formally agreed goals, we think they should interpret what the project is expected to do and then take this into account when setting scores. A project should not be awarded a score of 6 if its goal was trivial, and likewise, it should not be awarded a score of 1 if the goal was unattainable.

Few, if any, of the evaluators chose a truly theory-based approach to evaluation, which was somewhat surprising. Scholarly literature as well as past evaluations might have been helpful when deciding whether certain changes were likely to be an effect of the project (i.e. the attribution problem). Hatling, Damman, and Halvorsen (2016) concluded that a commonly used assumption in 'co-location' projects, namely that they will automatically generate synergies between residents of the building, is not well grounded in the literature. Kaplan and Garrett (2005) mentioned that a common example of theory failure is to assume that a new technology or infrastructure will make people change their habits without additional measures, such as training and financial incentives. A review of the programme theory could have revealed such a failure, and may similarly reveal redundant project components.

6.5. The evaluation teams

It was required that evaluation teams had no relation to the projects they evaluated. Furthermore, that they had expertise within evaluation, economics and project management, and some knowledge of the sector and type of project. As noted by Scriven (2015), an evaluation team must be broad and represent different perspectives and disciplines, as this is essential for comprehensive and balanced assessments. In our view, the latter was not always satisfied in the studied evaluations, as some of the teams consisted primarily of economists. Only 12 out of the 20 evaluations were performed by sufficiently broad teams. By contrast, 19 out of 20 had high or very high levels of expertise within economics.

6.6. Score-setting – the need for common guidelines

Score-setting was an essential part of the studied evaluations. Our findings indicate that the use of scores is valuable for drawing lessons across projects and sectors. However, experience suggests that efforts should be made to ensure that results are well calibrated. When scores are set by different teams, they may interpret and use the scale differently. A relevant question is whether we could have applied a more objective quantitative summary measure, where scores are obtained from the application of an algorithm that brings the same result independently from the evaluation team (see for instance Chiesa & Frattini, 2007). Unfortunately, we think the answer is no. As long as the framework is used to evaluate different projects with different types of outcomes, different stakeholders, etc., subjective judgements, regarding the choice of indicators as well as score-setting, cannot be avoided. Instead, we believe the solution to ensure calibrated results is clearer guidelines for evaluators. This should be seen in relation to the above-mentioned need for common interpretations of the strategic evaluation criteria, and the need to adjust the goals so that the levels of ambition are realistic across projects.

We also think it is important to be open about the level of uncertainty when it comes to score-setting, as are some evaluators actually, for example, by assigning 'high', 'medium' or 'low' uncertainty next to each score.

7. Conclusions

In Norway, ministries and agencies with large investment projects have become quite good at appraisal and planning. Since the year 2000, the project decision documents have gone through external quality assurance. The assumption is that this will also contribute to improved project performance. However, ex-post evaluations of government investment projects are still rare. Worsley (2014) referred to ex-post evaluations as "the weak link" in the assessment process for transport projects in OECD countries. This is perhaps not surprising. In contrast to, for example, health or educational programs, an infrastructure project cannot be implemented stepwise. Therefore, it could be argued that whereas good planning is crucial, ex-post evaluation is a waste of time and resources. However, that would be an erroneous conclusion because there is much to learn from one project to another, both within and between sectors. Given the poor reputation of public projects in high-income countries in general (Flyvbjerg et al., 2003), the potential to improve project practices is considerable. So is the potential to improve project planning, governance and the quality-at-entry scheme

itself. Evaluation should be based on the project's logic model, as recommended by several authors in the extant evaluation literature (see for instance Samset, 2003). It should ask not only about economic aspects, but take a broad and multifaceted view on project success. In their most recent economic survey of Norway, the OECD (2017), focusing primarily on transport projects, suggests that ex-post evaluation of projects are conducted more systematically, and that a broad framework is applied, to strengthen scope, accuracy and credibility. We have applied a generic evaluation framework inspired by the one recommended by the OECD-DAC for the evaluation of development assistance projects and programs.

A key finding in this study is that most of the projects were rather successful, as considered by the evaluators. This contrasts the public discourse and studies, by Flyvbjerg and others that demonstrate a low level of success in public projects. The 20 projects were highly successful in operational terms, and somewhat more varied in tactical and strategic terms. Some projects scored high on relevance and sustainability, but low on benefit-cost efficiency, and vice versa. This type of deviance needs to be communicated to project owners and various stakeholders, who might have conflicting views on the weighting of the criteria. The evaluations thus provide a basis for discussing whether a better balance between different concerns could have been possible. The possibility to compare, and learn across different sectors, is also considered useful. Some sectors are better at cost control, others at benefits realization, and still others at sustainability, etc.

Our conclusion is that the evaluation results by and large provide a realistic picture of the projects' success. Although the degree of success may seem very high, there is no reason to believe that there is a serious bias on behalf of the evaluators. It is a sample of the country's largest investment projects, which have been through a particularly comprehensive analytical and political process up-front, before they were approved individually by the country's highest authority, i.e. the Parliament.

The evaluators' experiences of the evaluation framework were largely positive. This time evaluation is not limited to aspects of project management success, which has traditionally been the main focus in the project management community. Neither is the framework limited to benefit-cost efficiency, which is normally the main focus in the transport sector. (Other sectors rarely conduct evaluations at all). Instead, the six criteria cover intended and unintended effects alike, goal-oriented and efficiency perspectives, and explicitly raise questions about the long-term viability. Also, this meta-evaluation revealed some improvement points, and the lessons learned should result in a set of requirements and guidelines for future evaluations, regarding how the teams should be put together, how the criteria should be understood, and clear, common principles for score-setting.

One lesson to be drawn is that the evaluation format used in development projects in low-income countries (LIC) is also well-suited in high-income countries (HIC). The reason is that there is no fundamental difference between investment projects in the two types of countries. All projects are implemented to have an impact, and evaluations should be useful for planners, beneficiaries, sponsors and other stakeholders alike. The main difference may be that HICs pay particular attention to projects' value-for-money as measured by the benefit-cost analysis, while in development projects social and ethical justifications may weigh heavier for donors and recipient countries. This has been taken into account in the Norwegian context by expanding the evaluation format with a separate assessment of benefit-cost efficiency.

In evaluations of development assistance projects, the trend in recent years has been to perform larger, strategic and often thematic evaluations, and not only focus on individual projects (OECD, 2016). This approach should be considered for public investment projects in high-income countries too, and we think that our project evaluations would provide useful input to such a broader topic.

In addition to the improved evaluation framework, ministries and agencies need to see the benefits of the evaluations and their learning potential. It is still too early to determine whether these 20 evaluations have led to improved practices, but this will be an interesting topic for future studies. It is well-known that it is more difficult to obtain learning and improvements when evaluations are initiated and conducted by an external party than from internal reviews. However, Reichborn-Kjennerud and Vabo (2017), who studied Norwegian ministries and agencies' ability to learn from performance audits, concluded rather positively. They found that audit reports were often used for improvements in planning and management systems, provided that the reports were found to be relevant, of good quality and sufficiently balanced.

Over time, hopefully, a large number of project evaluations will be produced corresponding to this framework. One ambition is to further improve their quality and ensure that scoring will become better calibrated over time. Since the projects in each sector have similar outcomes, allowing for rather standardized measures, the resulting evaluation database would then provide a valuable basis for robust practices and better determinants of government investment projects' success.

Funding

The work was supported by the Concept Research Programme at the Norwegian University of Science and Technology, which in turn is funded by the Norwegian Ministry of Finance.

Declaration of interest

No conflicts of interest.

Acknowledgements

The author would like to thank all the involved evaluators who contributed to the 20 evaluations as well as in focus group discussions. A special thanks to professor Knut Samset for great inspiration, interesting discussions and useful comments to an earlier draft.

References

ALNAP (2006). *Evaluating humanitarian action using the OECD-DAC criteria: An ALNAP guide for humanitarian agencies.* London: Overseas Development Institute.

Baccarini, D. (1999). The logical framework method for defining project success. *Project Management Institute, 30*(4), 25–32.

Bakewell, O., & Garbutt, A. (2005). *The use and abuse of the logical framework approach: A review of international development NGOs' experiences* [A report for Sida].

Bamberger, M., Rugh, J., Church, M., & Fort, L. (2004). Shoestring evaluation: Designing impact evaluations under budget, time and data constraints. *American Journal of Evaluation, 25*(1), 5–37.

Boardman, A., Greenberg, D., Vining, A., & Weimer, D. (2011). *Cost-benefit analysis: Concepts and practice* (4th ed.). Upper Saddle River, NJ: Pearson.

Brousselle, A., & Champagne, F. (2011). Program theory evaluation: Logic analysis. *Evaluation and Program Planning, 34*(1), 69–78.

Chen, H. T. (1990). *Theory-driven evaluations.* Thousand Oaks, CA: Sage.

Chianza, T. (2008). The OECD/DAC criteria for international development interventions: An assessment and ideas for improvement. *Journal of MultiDisciplinary Evaluation, 5*(9), 41–51.

Chiesa, V., & Frattini, F. (2007). Exploring the differences in performance measurement between research and development: Evidence from a multiple case study. *R&D Management, 37*(4), 283–301.

Connell, J. P., & Kubisch, A. C. (1998). Applying a theory of change approach to the evaluation of comprehensive community initiatives: Progress, prospects and problems. In K. Fulbright-Anderson, A. C. Kubisch, & J. P. Connell (Eds.). *New approaches to evaluating community initiatives. vol. II theory, measurement and analysis* (pp. 15–44). Washington D.C: Aspen Institute.

Eliasson, J., Börjesson, M., Odeck, J., & Welde, M. (2015). Does benefit-cost efficiency influence transport investment decisions? *Journal of Transport Economics and Policy, 49*(3), 377–396.

European Commission (2013). *EVALSED: The resource for the evaluation of socio-Economic development –Evaluation guide.* [Downloadable from: http://ec.europa.eu/regional_policy/en/information/publications/evaluations-guidance-documents/2013/evalsed-the-resource-for-the-evaluation-of-socio-economic-development-evaluation-guide. (Accessed 26 May 2017)].

European Commission (2015). *Guidelines on evaluation and fitness checks.* [http://ec.europa.eu/smart-regulation/guidelines/ug_chap6_en.htm. (Accessed 26 May 2017)].

Flyvbjerg, B., Skamris Holm, M. K., & Buhl, S. L. (2003). How common and how large are cost overruns in transport infrastructure projects? *Transport Review, 23*(1), 71–88.
Flyvbjerg, B., Garbuio, M., & Lovallo, D. (2009). Delusion and deception in large infrastructure projects: Two models for explaining and preventing executive disaster. *California Management Review, 51*(2), 170–194.
Flyvbjerg, B. (2014). What you should know about megaprojects and why: An overview. *Project Management Journal, 45*(2), 6–19.
Gasper, D. (2000). Evaluating the 'logical framework approach' towards learning-oriented development evaluation. *Public Administration and Development, 20*, 17–28.
Glass, G. (1976). Primary, secondary and meta-Analysis of research. *Educational Researcher, 5*.
Green, J., Roberts, H., Petticrew, M., Steinbach, R., Goodman, A., Jones, A., et al. (2017). Integrating quasi-experimental and inductive designs in evaluation: A case study of the impact of free bus travel on public health. *Evaluation, 21*(4), 391–406.
HM Treasury (2011). *The magenta book. guidance for evaluation*.
Hatling, M., Damman, S., & Halvorsen, T. (2016). *Samlokaliseringseffekter – hva sier litteraturen? En oversikt over resultater fra relevante studier*Trondheim: SINTEF [SINTEF Report no A27652].
Heinzerling, L., & Ackerman, F. (2002). *Pricing the priceless: Cost benefit analysis of environmental protection.* Washington, D.C: Georgetown Environmental Law and Policy Institute.
House, E. (2000). The limits of cost benefit evaluation. *Evaluation, 6*(1), 79–86.
Jacob, S., Speer, S., & Furubo, J. E. (2015). The institutionalization of evaluation matters: Updating the International Atlas of Evaluation 10 years later. *Evaluation, 21*(1), 6–31.
Kaplan, S., & Garrett, K. (2005). The use of logic models by community-based initiatives. *Evaluation and Program Planning, 28*, 167–172.
Karlsen, T., & Jentoft, N. (2013). Programteori versus samfunnsvitenskapelig teori. In A. Halvorsen, E. L. Madsen, & N. Jentoft (Eds.). *Evaluering: Tradisjoner, praksis og mangfold* (pp. 164–180). Bergen: Fagbokforlaget.
Kvalheim, E. V., Christensen, T., Samset, K., & Volden, G. H. (2015). *Har regjeringen fått et bedre beslutningsgrunnlag? Om effekten av å innføre konseptvalgutredning (KVU) og ekstern kvalitetssikring (KS1 og KS2) for store statlige investeringsprosjekter. [Concept working paper].* [https://www.ntnu.no/documents/1261860271/1262021752/4200+Regjeringens+beslutningsgrunnlag+sluttrapport.pdf/ff529615-f4c4-4b9d-a561-1d4790f7faea. (Accessed 26 May 2017)].
Love, P. E. D., Smith, J., Simpson, I., Regan, M., & Olatunji, O. (2015). Understanding the landscape of overruns in transport infrastructure projects. *Environment and Planning B: Planning and Design, 42* [490–209.
McLaughlin, J. A., & Jordan, G. B. (1999). Logic models: A tool for telling your programs performance story. *Evaluation and Program Planning, 22*(1), 65–72.
Morris, P. W. G., & Hough, G. H. (1991). *The anatomy of major projects: A study of the reality of project management.* Chichester, UK: John Wiley & Sons.
Morris, P. W. (2013). *Reconstructing project management.* Chichester, UK: John Wiley & Sons.
Norwegian Ministry of Finance (2015). Meld. St. 3 (2015-2016). Statsrekneskapen 2015.
Nyborg, K. (1998). Some Norwegian politicians' use of cost-benefit analysis. *Public Choice, 95*, 381–401.
OECD (1991). *DAC principles for evaluation of development assistance.* Paris: OECD, Development Assistance Committee.
OECD (2002). *Glossary of key terms in evaluation and results based management.* Paris: OECD, Development Assistance Committee.
OECD (2016). *Evaluation systems in development Co-operation.* Paris: OECD [2016 Review].
OECD (2017). *OECD economy surveys. Norway 2018.* Paris: OECD.
Pawson, R., & Tilley, N. (1997). An introduction to scientific realist evaluation. In E. Chelimsky, & W. Shadish (Eds.). *Evaluation for the 21 st century: A handbook* (pp. 405–418). Thousand Oaks, CA: Sage.
Picciotto, R. (2013). The logic of development effectiveness: Is it time for the broader evaluation community to take notice? *Evaluation, 19*(2), 155–170.
Project Management Institute (2013). *A Guide to the project management body of knowledge (PMBOK® Guide)* (5th ed.). PMI Inc.
Rambøll, & Agenda Kaupang (2016). *Bruk av evaluering i statlig styring* [Report R9392. https://evalueringsportalen.no/evaluering/bruk-av-evaluering-i-statlig-styring/Rapport%20DF%C3%98%20Bruk%20av%20evaluering%20i%20statlig%20styring.pdf/@@inline. (Accessed 26 May 2017)].
Reichborn-Kjennerud, K., & Vabo, S. I. (2017). Performance audit as a contributor to change and improvement in public administration. *Evaluation, 23*(1), 6–23.
Rogers, P., Petrosino, A., Huebner, T., & Hacsi, T. (2000). Program theory evaluation: Practice: promise and problems. *New Directions for Evaluation, 87*, 5–13.
Rosenberg, L., Posner, L. D., & Hanley, E. J. (1970). *Project evaluation and the project appraisal reporting system* [Final report submitted to the USAID, by Fry consultants incorporated].
Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks: Sage.
Samset, K., & Christensen, T. (2017). Ex ante project evaluation and the complexity of early decision-making. *Public Organization Review, 17*(1), 1–17.
Samset, K., & Volden, G. H. (2015). Front-end definition of projects: Ten paradoxes and some reflections regarding project management and project governance. *International Journal of Project Management, 34*(2), 297–313.
Samset, K., Andersen, B., & Austeng, K. (2014). To which extent do projects explore the opportunity space? A study of conceptual appraisals and the choice of conceptual solutions. *International Journal of Managing Projects in Business, 7*(3), 473–492.
Samset, K. (2003). *Project evaluation: Making projects succeed.* Trondheim: Tapir Akademisk Forlag.
Sartorius, R. H. (1991). The logical framework approach to project design and management. *Evaluation Practice, 12*(2), 139–147.
Scriven, M. (2015). *Key evaluation checklist (KEC).* [http://michaelscriven.info/papersandpublications.html].
Stufflebeam, D. (2010). Meta-evaluation. *Journal of Multi-Disciplinary Evaluation, 7*(15), 99–158.
Sullivan, H., & Stewart, M. (2006). Who owns the theory of change. *Evaluation, 12*(2), 179–199.
Ton, G. (2012). The mixing of methods. A three-step process for improving rigour in impact evaluations. *Evaluation, 18*(1), 5–25.
van Wee, B. (2007). Large infrastructure projects: A review of the quality of demand forecasts and cost estimations. *Environment and Planning B: Planning and Design, 34*, 611–625.
Volden, G. H., & Samset, K. (2015). Perverse incentives in the front-end: Public funding and counterproductive projects. *Paper at IRNOP conference*.
Volden, G. H., & Samset, K. (2017). Quality assurance in megaproject management: The Norwegian way. In B. Flyvbjerg (Ed.). *The oxford handbook of megaproject management* (pp. 406–427). Oxford: Oxford University Press.
Weiss, C. (1997). Theory-based evaluation: Past present, and future. *New Direction for Evaluation, 76*, 41–55.
Welde, M. (2017). *Kostnadskontroll i store statlige investeringer underlagt ordningen med ekstern kvalitetssikring*Trondheim: Ex Ante Academic Publisher [Concept Report No 51].
Wiig, H., & Holm-Hansen, J. (2014). *Unintended effects in evaluations of norwegian aid: A desk study*Oslo: Norwegian Agency for Development Cooperation [NORAD] [Norad Report 2/2014].
Williams, T., & Samset, K. (2010). Issues in front-end decision making on projects. *Project Management Journal, 41*, 38–49.
World Bank (2004). *Monitoring and evaluation: Some tools, methods and approaches.* Washington D.C: World Bank.
Worsley, T. (2014). *Ex post assessment of transport investments and policy interventions: Prerequisites for ex-post assessments and methodological challenges. ITS roundtable summary and conclusions.* [International transport Foundation Discussion Paper 2014–19. https://www.econstor.eu/bitstream/10419/109153/1/818314141.pdf. (Accesssed 26 May 2017).
Yin, R. (2013a). *Case study research Design and methods.* Sage.
Yin, R. (2013b). Validity and generalization in future case study evaluations. *Evaluation, 19*(3), 321–332.

Gro Holst Volden holds the position of research director of the Concept Research Programme on Front-end Management of Major Investment Projects, at the Norwegian University of Science and Technology. Her research is within project governance, public decision processes, and appraisal and evaluation of major public investments. She in an economist from the Norwegian School of Economics, and has a prior career as a senior advisor in the consulting industry and in the government agency for financial management. Volden is currently president of the Norwegian Evaluation Society.