Evaluation of Probabilistic Project Cost Estimates

Magne Jørgensen, Morten Welde, and Torleif Halkjelsvik

***Abstract*—Evaluation of cost estimates should be fair and give incentives for accuracy. These goals, we argue, are challenged by a lack of precision in what is meant by a cost estimate and the use of evaluation measures that do not reward the most accurate cost estimates. To improve the situation, we suggest the use of probabilistic cost estimates and propose guidelines on how to evaluate such estimates. The guidelines emphasize the importance of a match between the type of cost estimate provided by the estimators and the chosen cost evaluation measure, and the need for an evaluation of both the calibration and the informativeness of the estimates. The feasibility of the guidelines is exemplified in an analysis of a set of 69 large Norwegian governmental projects. The evaluation indicated that the projects had quite accurate and unbiased P50 estimates and that the prediction intervals were reasonably well-calibrated. It also showed that the cost prediction intervals were noninformative with respect to differences in cost uncertainty and, consequently, not useful to identify projects with higher cost uncertainty. The results demonstrate the usefulness of applying the proposed cost estimation evaluation guidelines.**

***Index Terms*—Cost distribution, cost estimation bias, cost estimation error, cost overrun, cost prediction intervals, probabilistic cost estimation.**

I. Introduction

PROJECT cost estimates are essential input to, amongst others, project plans, budgets, and bids [1]. Reports on cost estimation performance claim that there is an unfortunate tendency toward project cost overruns, see for example [2]–[7]. The strength of this tendency toward cost overruns varies from report to report, which is not surprising given differences in samples and contexts. In addition, the research that shows the strongest tendency towards cost overrun, for example the research reported in [2], [3], has been criticized for methodological flaws, see for example [8]. Differences in reported accuracy and bias may also be attributed to differences in the point of time of producing the cost estimates and to differences in stabilities in the project scopes [9].

The above complexities and problems related to aggregating and evaluating cost estimation performance are clearly important and need to be dealt with, but there are also other challenges in the evaluation of cost estimates. Two such challenges, which motivate the work described in this article, are: if we are not clear about what is meant by a "cost estimate," it will be difficult to properly evaluate how good it is; and evaluation measures should give the best possible score to the best possible cost estimates, i.e., they should reward unbiased cost estimates.

The first challenge, which is related to the current lack of precision in what is meant by a cost estimate, has been pointed out in, for example, [10], where the authors state that: "It may be considered surprising that neither the Project Management Body of Knowledge (2013) nor the Association of Project Management (2016) provide a definition for "cost overruns" or "cost over-budget," presumably assuming that its meaning is straightforward and its calculation clear." As far as we have identified, few studies on cost estimation accuracy and bias explicitly state the precise meaning of the analyzed cost estimates. Instead, they use general formulations such as "estimated costs are defined as budgeted, or forecasted, construction costs at the time of decision to build" [2, p. 281]; "Cost estimating could be defined as the process where an estimator arrives at an expenditure of resources necessary to complete a project in accordance with plans and specifications" [4, p. 141]; or "a compilation of all the probable costs of the elements of a project or effort included within an agreed upon scope" [11, p. 35]. Other research studies include no description of the intended meaning of the cost estimates at all [12]. The situation seems to be similar to what Gneiting describes as [13, p. 748]: "the common practice of requesting 'some' point forecast, and then evaluating the forecasters by using 'some' (set of) scoring function(s), is not a meaningful endeavor." In this article, we argue that the use of a probabilistic framework, i.e., the use of cost estimates referring to defined points on cost outcome distributions, enables precision in what is meant by cost estimate.

The second challenge is about finding meaningful and fair ways of evaluating cost estimates, may be described as the need for a match between the type of estimates provided by the projects and the evaluation measures used to evaluate them. By a match, we mean that the cost evaluation measure should give the best expected score for the best possible cost estimates. It should not reward misrepresentations and biased cost estimates. An example of a possible lack of match is when the overrun (bias) of estimates of the most likely cost of projects is evaluated by use of the mean relative error.[1] We explain the meaning and importance of a match between the evaluation measure and the

Manuscript received November 19, 2020; revised January 25, 2021 and March 10, 2021; accepted March 15, 2021. Review of this manuscript was arranged by Department Editor Y. H. Kwak. *(Corresponding author: Magne Jørgensen.)*

Magne Jørgensen is with the Simula Metropolitan Center for Digital Engineering, 0167 Oslo, Norway, and also with the Oslo Metropolitan University, 0167 Oslo, Norway (e-mail: magnej@simula.no).

Morten Welde is with the Norwegian University of Science and Technology, 7491 Trondheim, Norway (e-mail: morten.welde@ntnu.no).

Torleif Halkjelsvik is with the Simula Metropolitan Center for Digital Engineering, 0167 Oslo, Norway (e-mail: torleif@simula.no).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TEM.2021.3067050.

Digital Object Identifier 10.1109/TEM.2021.3067050

[1]The reason for this is that, assuming a non-symmetric cost outcome distribution, even when giving perfect estimates of the most likely costs (the mode values of the cost distributions), we should expect a non-optimal score of cost overrun if the evaluation measure is the mean relative error.

type of estimate in more detail in Section II, along with a set of evaluation measures and matching types of cost estimates. This need for a match between measures and types of estimate (proper scoring rules) is the same as pointed out in, for example, [13], [14].

The main goal of this article is to address the above two challenges by providing guidelines that support better evaluation of cost estimates. Specifically, we advocate for the adoption of probabilistic estimation, and we describe how the evaluation measures can be matched to a given type of cost estimate. The use of probabilistic estimates and matching evaluation measures allows for fair evaluations that incentivize giving unbiased estimates.

The rest of this article is organized as follows. Section II starts by introducing how to use a probabilistic cost estimation framework. It then proposes guidelines for better evaluation of cost estimates. The section also includes information about the scope of the guidelines and when it is meaningful to use them. Section III exemplifies the use of the guidelines in the cost estimates of 69 large governmental projects, all of them using probabilistic cost estimates and meeting the conditions for meaningful use of the guidelines. Section IV discusses the use of the guidelines in relation to the dataset, including discussion on the implications for evaluation of cost estimates and some limitations of the proposed guidelines. Finally, Section V concludes this article.

II. Guidelines for the Evaluation of Cost Estimates

This section starts with the presentation of a probabilistic framework for cost estimates (see Section II-A). A probabilistic framework, as argued in for example [15], is essential to enable precise interpretation and communication of the intended meaning of estimates. The section then provides guidance on matching estimation evaluation measures to types of cost estimates (see Section II-B), and how to evaluate both the calibration and informativeness of cost prediction intervals and distributions (see Section II-C). Note that there are several challenges in the evaluation of cost estimates that are not covered by the proposed guidelines, see for example [16] for information about other challenges. Other challenges include those related to project scope changes and the presence of cost estimates given at different stages of the project lifecycle. The guidelines proposed in this section assume that actual costs refer to completed projects that are comparable with estimated projects. The guidelines also assume that the evaluation measures are used on a set of cost estimates of a similar type and that they are derived from a similar stage of a project's lifecycle.

A. Probabilistic Framework for Cost Estimates

Consider a construction project where the cost outcome is influenced by stochastic factors, such as labor availability, equipment utilization, the weather, and mistakes. As these factors are not fully under the control of the project management, one may consider the final cost as a realization from a distribution of potential cost outcomes. The actual cost of a hypothetical project may, for example, be considered to be a number sampled from a distribution similar to that depicted in Fig. 1.[2] Here, there is a 50 percent chance of sampling a final cost above EUR 27 million and a 90% chance of sampling a cost between EUR 12 million and EUR 62 million. The shaded area shows the density probability distribution, and the *S*-curve displays the cumulative probability distribution. The values displayed on the *x*-axis are the mode, the median, the mean, and the 85th percentile (the P85 value) of the cumulative probability distribution.

A cost estimate may in principle refer to any part of a cost outcome distribution. Consequently, stating that we have estimated the cost as, for example, EUR 20 million, provides limited information. We cannot know whether the intended meaning of the cost estimate is the most likely (mode), the median, the mean, or some other point of the outcome distribution. Clearly, when the intended meaning of cost estimates is unknown it is hard to evaluate their accuracy and bias meaningfully.

Probabilistic estimates may be given as *PX* estimates, where *X* is the percentile of the outcome distribution. We may, for example, attempt to estimate the P50 (the median of the cost outcome distribution), which can be used as input to a project plan, or we could estimate the P85 (85% chance of no overrun) for the purpose of obtaining a portfolio level budget. *PX* estimates can be considered as both single-point cost estimates and one-sided cost prediction intervals, i.e., the intervals from 0 to *PX*.

Probabilistic estimates may also be given as two-sided cost prediction intervals (PI_X), where the *X* value indicates the likelihood of including the actual value in the interval. For example, a nearly perfect PI_{90} for the project in Fig. 1 would be the interval from EUR 11.9 million to EUR 61.7 million. Finally, probabilistic estimates may be given as the complete cost outcome distributions, such as the complete distribution in Fig. 1.

The above terminology is the same as that used in other management research [19], [20] and in statistics [15].

B. Evaluation of Single-Point Probabilistic Cost Estimates

A fair evaluation of estimation error should be based on a match between the type of cost estimate evaluated and the error measure. By a match, for the purpose of the present article, we mean that the loss function implied by the error measure should be minimized by the intended type of estimate [21], [22]. That is, the evaluation measure should give the best possible score for cost estimates when the estimated and true positions in the cost distributions are the same.[3] An alternative formulation of

[2]The hypothetical cost distribution in Fig. 1 follows a log-normal distribution with parameters $\mu = 3.3$ (mean), and $\sigma = 0.5$ (standard deviation). A log-normal distribution may be a good fit for many real-life cost outcome distributions and has the property that the mean cost is higher than the median cost, which is higher than the most likely (mode) cost. Two other properties of log-normal distributions that make them useful for cost distributions are that thy have only positive values and that they are right-skewed. A right-skewed cost distribution is consistent with that project cost can be much higher than estimated, but are seldom much lower. While, for example, a 200% cost overrun is possible, a 200% cost underrun is impossible. More on the properties and the use of log-normal distributions in cost estimation and other disciplines can be found in [17] and [18].

[3]Formally, this match may be formulated such that the optimal single-point probabilistic cost estimate $(\widehat{\mathrm{est}})$ from the estimated cost distribution F minimizes

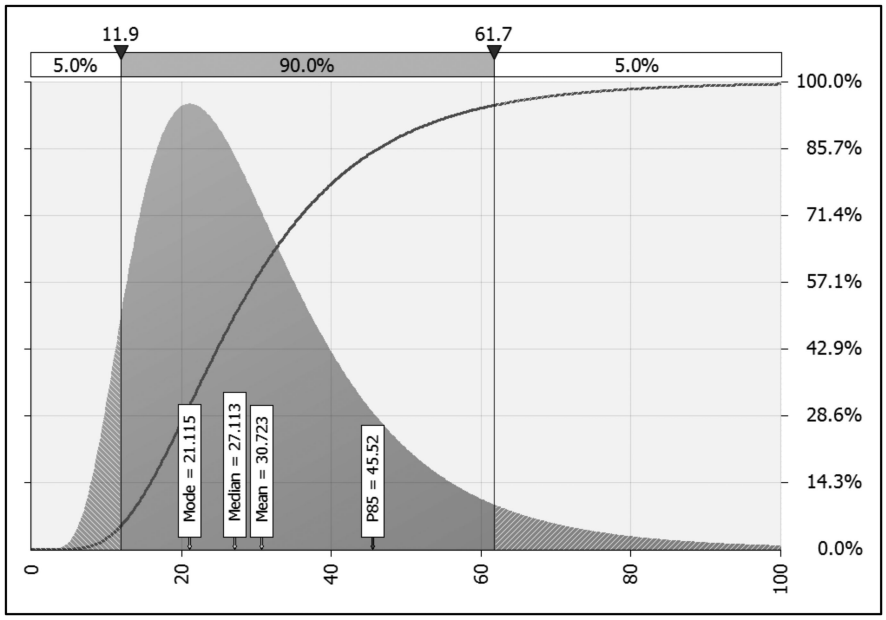

Fig. 1. Density and cumulative probability distribution of the cost of a hypothetical construction project.

the match criterion is that the best possible expected evaluation score should be achieved when we have the best possible cost estimates. The best possible cost estimates are here those that perfectly represent the intended position in the cost outcome distributions, e.g., estimates where the estimated P50 (median) costs equal the actual P50 cost outcomes or the estimated mean costs equal the actual mean cost. In other words, when we request P50 cost estimates and the given estimates equal the true P50 costs in the underlying cost uncertainty distributions, this should give the optimal expected score for the appropriate evaluation measure. An example of lack of match is when the evaluation of the bias of P50 cost estimates applies the mean error ($\frac{1}{N}\sum_{i=1}^{N}(\mathrm{act}_i - \mathrm{est}_i)$) as measure. In this case, the expected mean error is *not* minimized by perfect estimates of the P50 costs, but instead by perfect estimates of the mean costs. To achieve a match between the type of estimates and the evaluation measure in this example, we either have to request estimates of the mean cost, or change the evaluation measure to one that is minimized by the P50 cost. If this is not done, we would not reward the provision of cost estimates that truly reflect their intended probabilistic interpretation.

Table I gives a selection of cost estimation error measures and matching type of cost estimates. The selection includes measures in common use in the evaluation of cost estimation accuracy and measures that match common single-point cost estimates such as the mean and the P50 (median). Mathematical proof of the match between the error measures and the type of estimates can be found in the references provided in the second column of the table.

One observation of particular interest is that the expected value of the mean absolute relative error (absolute difference between estimated and actual divided by the estimated costs) is not minimized by any common type of cost estimate. As given in Table I, the minimizing type of cost estimate for this error measure is far from intuitive. That is to say, the expected estimation error is not minimized when providing perfect estimates of the most likely cost, the median cost, or the mean cost, or, as far as we know, any other type of cost estimate used in research or practice. The use of the mean absolute relative error may for this reason result in unfortunate cost estimation incentives and is an example of a nonmatching evaluation measure.[4]

Table I gives measures of error (or accuracy), which calculate error irrespective of whether the error represents overrun or underrun, i.e., the unsigned error. For measures of bias (signed error or magnitude of overrun/underrun), the scores can be either positive or negative, and the best score is zero. A selection of measures of bias with matching single-point estimates is given in Table II. The presented bias measures are either in common use or potentially useful as measures that match common types of cost estimate. Unlike the error measures, we were unable to find prior research studies with proof of the match. Our proofs are given in Appendix 1.

Table II implies, amongst others, that it may be unfair to evaluate P50 cost estimates using the mean relative error. Assume, for example, that we have a set of projects with a cost distribution like the one displayed in Fig. 1, and that all projects used as their estimates the true P50 estimate of the cost. Sampling from the

the expected ($\mathbb{E}$) loss (L), which is represented by the error measure, for the distribution of actual cost *G*, i.e., $\widehat{\mathrm{est}} = \arg\min_{\mathrm{est}} \mathbb{E}_F L(\mathrm{est}, G)$.

[4] Assuming the log-normal distribution in Fig. 1, which has the parameters $\mu = 3.3$ (mean), and $\sigma = 0.5$ (standard deviation), the cost estimate that minimizes the expected value of the mean absolute relative error can be shown to equal $e^{\mu+\sigma^2} = e^{3.3+0.5^2} = 34.8$ million , which is the 69.1 percentile of the cost distribution (for proof, see [23]). A cost estimate of EUR 34.8 million is higher than both the mode (21 million), the median (27 million) and the mean (31 million) cost. The mean absolute relative error would consequently reward very high cost estimates (higher than the mean cost), and be optimized by a type of cost estimate that is difficult to interpret.

TABLE I
ERROR MEASURES WITH MATCHING SINGLE-POINT ESTIMATES

Error measure	**Single-point estimate that, when optimal, minimizes the expected value of the measured error**
Mean absolute error: $\frac{1}{N}\sum_{i=1}^{N}\lvert act_i - est_i\rvert$	The estimates of the P50 (median) cost [13].
Mean square error: $\frac{1}{N}\sum_{i=1}^{N}(act_i - est_i)^2$	The estimate of the mean cost [13].
Mean absolute relative error: $\frac{1}{N}\sum_{i=1}^{N}\frac{\lvert act_i - est_i\rvert}{est_i}$	The median values of stochastic variables with density function symmetric to $act_i \cdot f_i(act_i)$, where f_i is the cost density function of project *i* [13].
Mean absolute log-error: $\frac{1}{N}\sum_{i=1}^{N}\lvert\ln(act_i) - \ln(est_i)\rvert$	The estimates of the P50 (median) cost [24].

TABLE II
BIAS MEASURES WITH MATCHING SINGLE-POINT ESTIMATES

Bias measure	**Single-point estimate that, when optimal, has an expected value of zero for the measured bias**
Mean error: $\frac{1}{N}\sum_{i=1}^{N}(act_i - est_i)$	The estimate of the mean cost
Median error: $Median\ of (act_i - est_i), i = 1 \dots N$	The estimate of the P50 (median) cost
Mean relative error: $\frac{1}{N}\sum_{i=1}^{N}\frac{(act_i - est_i)}{est_i}$	The estimates of mean cost
Median relative error: $Median\ of\ \frac{(\mathrm{act} - est)}{est}$, i=1…N	The estimate of the P50 (median) cost
Median logarithmic error: $Median\ of\ \ln(act_i) - \ln(est_i)$, for i=1…N	The estimate of the P50 (median) cost
Median log-error: $Median\ of\ \frac{(act_i - est_i)}{\min(act_i, est_i)}, i = 1 \dots N$	The estimate of the P50 (median) cost

cost uncertainty distribution in Fig. 1(we took 10.000 samples), we find that even with perfect P50 estimates, i.e., the cost estimates representing the median of the actual cost distribution, there would be a bias toward cost overrun of around 13% when using the mean relative error as the evaluation measure. Notice also that a measure based on the median relative error would give an expected bias of zero and establish a proper evaluation of the performance of P50 cost estimates.

When evaluating the bias or accuracy of single point, probabilistic cost estimates, the only information needed is the actual

cost and the estimated cost, where the estimated cost need to include information of its intended location in the cost distribution, i.e., its position in the estimated cost distribution. This means that it is meaningful to evaluate the estimation accuracy and bias of individual probabilistic cost estimate. This is not the case for the evaluation of calibration and informativeness of cost prediction intervals and cost probability distribution. Here, we need a set of projects for meaningful evaluation. Exactly how large set of projects we need depends on the purpose of the evaluation and how confident one need to be in the evaluation results. Less than 10–20 projects would, we believe, in most cases lead to low robustness of the evaluation of calibration and informativeness of prediction intervals and probability distributions.

C. Evaluating Prediction Intervals and Probability Distributions

When probabilistic cost estimates are given as prediction intervals or as full distributions, they are typically evaluated in terms of calibration [25], [26]. If, for example, a set of 90% cost prediction intervals (PI_{90}) includes about 90% of the actual cost values, the prediction intervals are considered to be well calibrated, suggesting good cost estimation performance. While the degree of calibration is useful information, we will argue that this is not sufficient for a complete evaluation of the performance of cost prediction intervals or probability distributions.

To enable a more complete evaluation, we propose the inclusion of measures of informativeness. Informativeness, evaluated together with calibration, may in this context be seen as related to the "sharpness-subject-to-calibration" criterion, proposed by Gneiting *et al*. [27]. That is, the ideal cost uncertainty estimate is the one that is perfectly calibrated and with as narrow (sharp) an estimated cost uncertainty distribution or prediction interval as possible. Informativeness, as we use it here, is a more general concept, including measures of informativeness other than sharpness. It includes, for example, measures of how well the variation in estimated cost uncertainty reflects the variation of actual cost uncertainty.

1) Why is Informativeness Important?: Assume that three estimators (*A*, *B*, and *C*) have been asked to provide the P50 estimates for a set of projects. For illustration purposes, assume that these P50 estimates are perfect, i.e., they represent the true median values of the underlying cost uncertainty distributions. Additionally, the estimators are asked to provide P85 estimates of the cost of these projects, which is required for budgeting purposes. The estimators calculate P85 estimates by adding cost contingency to the P50 estimate. Below, we show how the three estimators apply different strategies that all succeed in providing perfectly calibrated P85 estimates, but with substantial differences in informativeness and need for contingency.

1) *Perfect Estimator:* This estimator correctly identifies the individual risks (the complete, true underlying uncertainty distributions) for each of the projects, i.e., this estimator provides perfect P85 estimates based on contingencies tailored for each individual project. Consequently, there is an 85% probability that the actual cost of each project is within the respective P85 estimates, and, in the long run, the hit rate across multiple projects will be 85%. The difference between projects contingencies represents here the actual difference in cost uncertainty between the projects.
2) *Focused, Overall Uncertainty-Oriented Estimator:*[5] This estimator is unable to assess the differences between projects in their degree of uncertainty, but knows the overall cost uncertainty and uses this to create the P85 estimates. For example, the estimator has access to historical data and notes that 85% of previously completed projects spent less than 130% of their P50 estimates of the cost. To calculate the P85 estimates, the estimator multiplies all P50 estimates by a factor of 1.3 to find the individual P85 estimates. This strategy will also result in a hit rate of 85%. Treating the uncertainty of all projects the same, however, means that the informativeness of the P85 estimates is low. The estimates are, for example, not informative regarding differences in cost uncertainty among the projects. The median cost contingency required to achieve an 85% hit rate will be higher than (or the same as) that of the perfect estimator.[6]
3) *Unfocused, Overall Uncertainty-Oriented Estimator:* Like the overall uncertainty-oriented estimator, this estimator has knowledge about the overall cost uncertainty and is unable to identify the differences in cost uncertainty between the projects. Unlike the focused estimator (B), this estimator randomly varies the multiplication factor around the factor representing the overall uncertainty. This strategy may also lead to perfectly calibrated P85 estimates. It does, however, also lead to P85 estimates that are even less informative (misleading variation in added contingency) and a need for higher cost contingencies.

All the above uncertainty estimation strategies may lead to perfect calibration, yet they produce cost estimates that vary substantially in informativeness. Furthermore, they differ in terms of the magnitude of the contingencies needed to obtain good calibration and illustrates the value of using informativeness as an additional criterion in the evaluation of prediction intervals.

2) Measures of Calibration and Informativeness: Table III gives a selection of measures of calibration, and informativeness of cost prediction intervals and probability distributions. The table includes frequently used measures, such as hit rate, to evaluate the calibration of prediction intervals [28], [29], along with measures that are not in such common use. This is in particular the case for measures enabling evaluation of the informativeness of cost prediction intervals and distributions [30], [31].

The measures of hit rate, PIT, and PIT histogram are measures of calibration. The measure of relative width and the correlation

[5]This is sometimes referred to as a climatologic forecaster, see for example [27].

[6]The perfect estimator will be able to correctly identify the cost uncertainty of all projects, which implies that the median cost contingency is given by the project with the median highest (middle) cost uncertainty. This cost contingency will typically be lower than that of the project with the 85% highest cost uncertainty (the 85 percentile), which is the cost contingency used by the estimators of type B or C.

TABLE III
MEASURES OF CALIBRATION AND INFORMATIVENESS OF PREDICTION INTERVALS AND DISTRIBUTIONS

Prediction intervals measures
Hit Rate$_X$: $\frac{1}{N}\sum_{i=1}^{N} h_i$, $h_i = \begin{cases} 1, act_i \in PI_{i,X} \\ 0, \text{otherwise} \end{cases}$ $PI_{i,X}$ is the prediction interval intended to include the actual cost with X% probability for project i. For a set of well calibrated PI_X intervals, the hit rate will in the long run be close to X%. For example, the hit rate of a set of well calibrated one-sided intervals of the type P85 = [0; P85 estimate] will contain around 85% of the actual costs, and a set of well calibrated two-sided (central) intervals of the type PI_{90} = [P5; P95] will contain around 90% of the actual costs.
Mean relative width: $\frac{1}{N}\sum_{i=1}^{N} \frac{(P90_i - P10_i)}{P50_i}$ When comparing different cost prediction intervals for the same projects, narrower widths indicate better estimation performance, given similar calibration levels.
Correlation (r) between the relative width of uncertainty intervals and the estimation error of point estimates.
While a correlation of zero would indicate no ability to separate low and high cost uncertainty, the maximum possible correlation is determined by the shape of the cost uncertainty distribution and the overall distribution of projects' levels of uncertainty.
Cost distribution measures
Probability integral transform (PIT): $F_i(act_i)$, where F_i is the (cumulative) estimated cost distribution for project i. PIT is consequently the value (between 0 and 1) we get when actual cost is used as the input to the estimated cost distribution F.
PIT histogram: A histogram of PIT values of a set of actual cost. The histogram will show a uniform distribution, given sufficient observations, when all estimated cost distributions are perfectly calibrated. A ∪-shaped PIT histogram suggests that the estimated cost distributions have tended to be too narrow, while a ∩-shaped PIT histogram suggests that the distributions have tended to be too wide.
Median continuous ranked probability score (CRPS): Median of $\int_{\mathbb{R}} (F_i(z) - \mathbb{I}(act_i \leq z))^2 dz$, for i=1..N, where the identity function $\mathbb{I}(act_i \leq z)$ is 1 when $act_i \leq z$, otherwise 0. The value (the integral) is minimized when the estimated distribution F equals the outcome distribution G. This measure will give better scores for sharper (narrower and more informative) cost distributions for the same level of calibration.

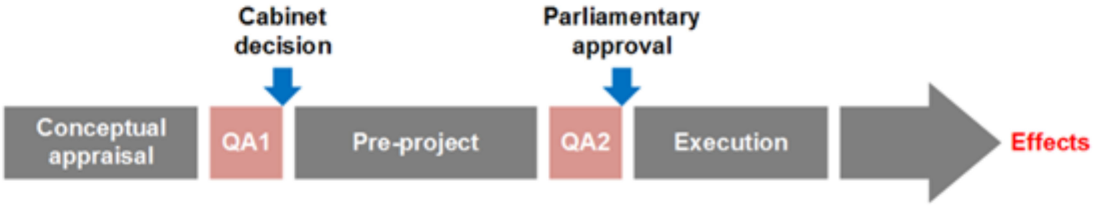

Fig. 2. Norwegian QA scheme for governmental projects.

between the relative width and the estimation error indicate the informativeness. The CRPS combines calibration and informativeness, measured as the sharpness of the cost uncertainty distribution. The CRPS measure and the PIT-based measures have, to our knowledge, not been used to evaluate probabilistic project cost estimates before, but they are in common use in other domains, such as in the evaluation of weather forecast models' predictions [32].

III. Evaluation of Cost Estimates in Norwegian Governmental Projects

This section aims to exemplify the feasibility and usefulness of the proposed evaluation guidelines by applying them on a sample of large Norwegian governmental projects. The projects used probabilistic cost estimates, the cost estimates were given at about the same stage of the project lifecycle, all cost estimates were given with the same intended probabilistic meaning, and, to our knowledge, there were no large scope changes that could explain large cost estimation errors. These projects consequently meet the conditions for meaningful aggregation and evaluation of cost estimation performance, and use of the proposed guidelines.

The analyzed projects and the process leading to the production of probabilistic cost estimates are described in Section III-A, followed by a report on the results from our evaluation of the single-point probabilistic cost estimates in Section III-B, and the results from our evaluation of the cost prediction intervals and distributions in Section III-C.

A. Projects and Their Cost Estimates

The analyzed projects had all been subject to a comprehensive governmental quality assurance (QA) scheme, which was compulsory for projects with an expected cost exceeding about EUR 75 million. All projects produced probabilistic cost estimates as input to budgets and project plans. Through the national research program Concept,[7] we obtained access to quality assured data about the probabilistic cost estimates and the actual cost of 69 of these projects, which consisted of 35 road construction projects, eight railway construction projects, fourteen building construction projects, seven information system development projects, and five procurement projects.

The QA scheme, requiring and supporting the production of probabilistic cost estimates, was introduced in the year 2000 after years of large cost overruns and projects failing to deliver the intended effects. The QA scheme is a gateway model (see Fig. 2).

The starting point of a project is a conceptual appraisal, where the responsible department assesses the cost and benefits of different alternatives and selects one of them. If the cabinet decides to move forward, the details of the project may be further developed for the chosen alternative. Before a project can be submitted to the Parliament for approval and achieve funding, the project must undergo a second round of external QA2. QA2 has an emphasis on project management and cost estimation issues, and produces, or at least checks the quality of, the probabilistic cost estimates.

The QA work in QA2 is extensive, with a mean duration of 6.5 months [33]. Central to the cost estimation in QA2 is the production of the P50 and P85 estimates of the cost, but the full cost uncertainty distributions are also estimated. The P50 estimates of cost are used as input to the plans and budgets at the project level, while the P85 estimates produced are used as input to budgets at the portfolio level. We collected information about the full cost distributions, including information about the P10, P50, P85, P90 and mean cost estimates.

When interpreting the results from our analyses of the cost estimates, the following should be considered.

1) The analyzed cost estimates are typically derived close to projects' start as part of QA2, and are likely to be less uncertain than the earliest cost estimates for the projects. The cost estimates are, however, derived before requesting and accepting bids for the project work and/or selecting providers. This means that even when the cost estimates are based on thorough estimation processes, the remaining cost uncertainty may be substantial.
2) The formal decision on whether to start the project occurs after QA2. According to [34], there are reasons to believe that the implemented QA scheme and its connected stage-gate decision model for project development have been useful to stop weak project proposals and reduce the risk of cost overruns. Bukkestein [35] reports that 7% of projects have been stopped after QA2. This removal of poorly planned and estimated projects may have increase the expected accuracy of the remaining cost estimates compared to other project contexts.
3) The method used to produce the probabilistic cost estimates was typically based on requesting expert estimates of the minimum (typically the P10 estimate), the most likely (the mode), and the maximum (typically the P90 estimate) cost of all cost elements, together with identifying and assessing cost risk elements. The estimates of cost elements and risks were typically aggregated to find

[7][Online]. Available: www.ntnu.no/concept

TABLE IV
ESTIMATION ERROR AND BIAS OF THE P50 ESTIMATES

Measure of estimation error and bias	**Results**
Mean absolute error: $\frac{1}{N}\sum_{i=1}^{N}\lvert act_i - est_i\rvert$	EUR 22.1 million
Mean log-error: $\frac{1}{N}\sum_{i=1}^{N}\lvert \ln(act_i) - \ln(est_i)\rvert$	15%
Median error (bias): Median of $(act_i - est_i)$, i=1...N	EUR 3.8 million
Median relative error (bias): $Median\ of\ \frac{(\mathrm{act}-est)}{est}$, i=1...N	4%
Median log-error (bias): $Median\ of\ \ln(act_i) - \ln(est_i)$, for i=1...N	4%

the cost estimate of the total project using Monte Carlo simulations. While the estimation and aggregation process may be systematic and based on robust statistical theory, the realism of the cost estimates will rely on the quality of the input from the estimation experts' judgments.

4) The average duration between estimation of the cost and completion of the project was seven years. To enable a fair comparison of the estimated and the actual cost, using domain-specific cost indexes, we index regulated the cost estimates and actual costs to reflect the completion year of each project. The index regulation of the estimated and actual cost led to a mean increase in cost of 27%, i.e., typically around a 3% index-regulated cost increase per elapsed year. Without this index-regulation the measured level of cost overrun would have been larger, but misleading.

B. Estimation Error and Bias of Single-Point Estimates

As described earlier, the P50 estimate was used as input to the budget at the project level, while the P85 estimate was used as input to the budget at the portfolio level. We chose not to evaluate the P85 estimate as a single-point cost estimate, but will later evaluate it as a one-sided prediction interval. The reason for this is that meaningful (matching) evaluation of a P85 single-point cost estimate would require the use of an error measure where one cost unit overrun is close to 5.7 (= 85%/15%) times worse than one cost unit underrun.[8] To assume this type of loss function was deemed not meaningful in our context.

Table IV gives the projects' estimation error and bias using measures that match the P50 type of cost estimate.

As given in Table IV, the mean log-error is 15%. Whether we should consider this as indicating a good estimation performance requires knowledge about the underlying cost uncertainty. To illustrate that a mean log-error of 15% can reflect very accurate cost estimates, consider the cost outcome distribution in Fig. 3. Fig. 3 exemplifies a cost uncertainty distribution where perfect P50 estimates of costs lead to an expectation of 15% estimation error. For the cost distribution in Fig. 3, it is 67 percent probable (one standard deviation) that the actual cost is between 80% and 120% of the perfect P50 estimate, and 90% likely that the actual cost is between 70% and 140% of the perfect P50 estimate. Cost distributions with this level of uncertainty may not be unusual for the types of large project included in our dataset, which means that an estimation error of 15% may in fact indicate very high estimation accuracy. Clearly, it is unlikely that the P50 estimates of the projects in our dataset are perfect. The above analysis, nevertheless, suggests that much of the estimation error is caused by an underlying cost uncertainty, and cannot be used to claim low estimation performance. Without knowing the underlying cost uncertainty, it is hard to evaluate exactly how good the estimation performance is in terms of accuracy.

A more traditional way of assessing the cost estimation performance is to compare it with that found in other studies. Using this approach, we may argue that the cost estimation error for the analyzed projects is low, since it is better than that reported in several other studies on project costs or effort estimates. For example, the review paper [37], reports an aggregated median estimation error of 25% and the analysis of Dutch infrastructure projects in the work of Cantarelli *et al*. [38] suggests a mean estimation error of 29%. However, it may be difficult to meaningfully compare estimation performances across different studies, because the nature of the estimates may differ substantially (different types of cost estimates, early versus late estimates, different degrees of cost uncertainty in different industries, different possibilities to adapt the deliveries to fit the budgeted cost, etc.). In addition, as pointed out earlier, most earlier studies do not describe what is meant by cost estimates, for example whether the cost estimates represent the most likely cost, the P50 cost, the P85 cost, or a mix of different types of estimates.

Perfect P50 estimates of cost should lead to zero bias when using a matching evaluation measure. Table IV gives that the cost estimation bias of the analyzed projects is close to this optimum,

[8]The only accuracy evaluation measure matching a single-point P85 estimate, see [36], is $\frac{1}{N}\sum_{i=1}^{N}\begin{cases}0.85\cdot|\mathrm{act}_i-\mathrm{est}_i|, & \text{if } \mathrm{est}_i \le \mathrm{act}_i\\ 0.15\cdot|\mathrm{act}_i-\mathrm{est}_i|, & \text{otherwise}\end{cases}$

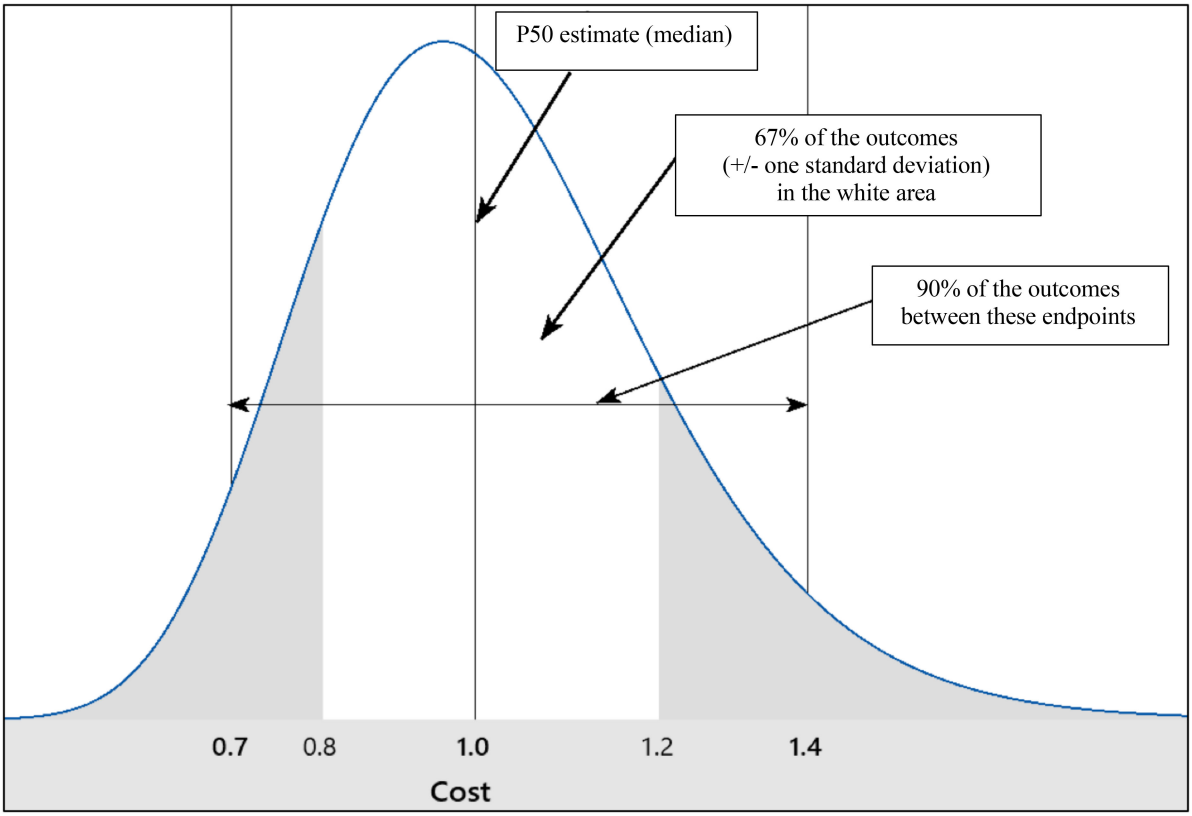

Fig 3. Example of a cost uncertainty distribution with perfect P50 estimates leading to an estimation error of 15%.

with both median relative error and median log-error measured as 4%. The bias is not statistically significantly different from zero ($p = 0.52$, Wilcoxon signed-rank test). A median overrun of 4%is substantially better than is typically reported for other projects of similar type. For example, the mean cost overrun reported in [3] was 45% for railway construction projects and 20 percent for road construction (however, see the critique of this analysis in [8]), and the median overrun reported in the review paper [37] was 21%.

The distribution of the cost estimation bias of the projects is displayed in Fig. 4. Fig. 4 shows a longer tail toward cost underruns than overruns, and that there were no very large overruns. The largest overrun had a log-error of 31% (relative error of 37%). The mean absolute log-error of projects with cost underruns was 18%, which was larger than the mean absolute log-error of projects with cost overruns, calculated as 12%. This result deviates from what is found in other studies, where cost overruns tend to be larger than cost underruns, see for example [39]–[41]. The current result may indicate that the QA process has been successful in improving cost realism and/or stopping projects with over-optimistic cost estimates.

A mean bias close to zero is, however, not necessarily an indicator of good estimation performance. In our case, as displayed in Fig. 5 , the earliest projects (starting in the period 2001–2003) had a mean underrun of 16%, while the later projects (starting in the period 2010–2012) had a mean overrun of 8%. It was mainly in the middle periods (2004–2006 and 2007–2009) that the cost estimation bias was low. The bias toward cost overruns in the last period compensated for the bias toward cost underruns in the earlier projects and resulted in low overall cost estimation bias, which demonstrates that measures of cost estimation bias should be analyzed and interpreted carefully.

C. Calibration and Informativeness of Prediction Intervals and Distributions

Table V gives measures of the calibration and informativeness of the cost estimates for the projects. The PIT and CRPS values require information about the full cost distributions, which were derived by fitting log-normal distributions to each project's estimates of the P50, P90, and the mean cost, using the distribution fitting functionality in the risk analysis tool @Risk.

The best calibrated *PX* interval in Table V was the P50 interval, which included 42%of projects' actual costs. The difference between the hit rate of 42%and the norm (50%) was not statistically significant ($p = 0.23$). The P50 estimates of cost were used for projects' plans and budgets, and it was particularly important that these estimates were well calibrated. However, the relatively good calibration hid the fact that the P50 cost estimates, together with the P10, P85, and P90 cost estimates, had been too high for the first projects (starting in the period 2001–2003) and too low for the later projects (starting in the period 2010–2012, see Fig. 6). Had the *PX* estimates been better calibrated for the last period, the total calibration would actually look worse. As with the measurement of estimation bias (where our median-based analysis of the P50 estimate correspond closely with the above

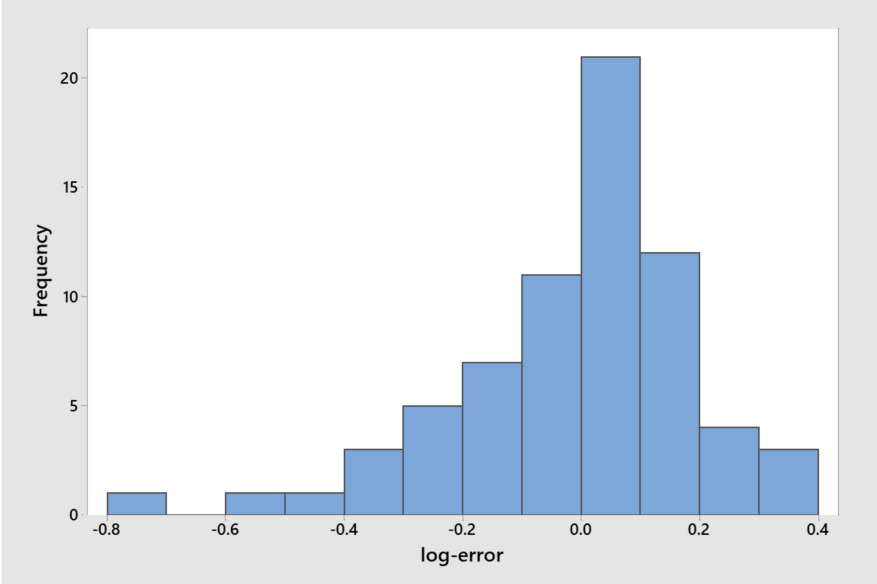

Fig. 4. Histogram of the estimation error (log-error).

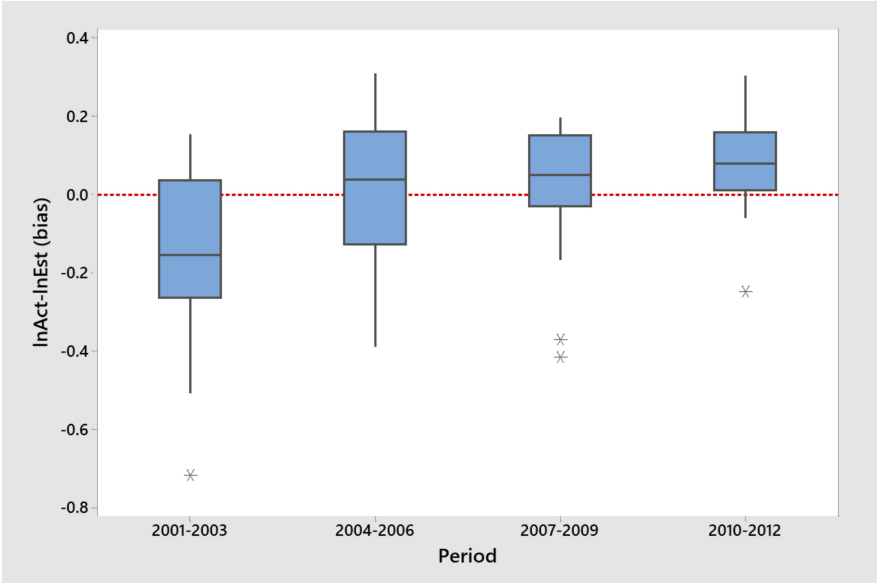

Fig. 5. Boxplot of cost estimation bias of different time periods.

analysis of the P50 interval), good performance on calibration may hide underlying biases.

The PIT histogram in Fig. 7 shows the calibration of the full distribution. A perfectly calibrated estimated cost distribution would give a uniform histogram. The shape of the histogram in Fig. 7 suggests that there were too many actual costs lower than the lower *PX*-values and too many higher than the higher *PX*-values, i.e., the estimated cost distributions tend to be too narrow to reflect the actual cost uncertainty.

Table V gives a correlation between the relative width (as an indicator of the estimated cost uncertainty) and the estimation error (as an indicator of the actual cost uncertainty) of -0.02. This suggests a very low, or perhaps no, ability to separate projects with low and high cost uncertainty, i.e., the prediction intervals contained little, if any, information about differences in the cost uncertainty of the projects. Fig. 8, which includes a smoothed line based on a locally weighted scatterplot smoothing function (degree of smoothing 0.5, number of steps 4) further illustrates this lack of correlation.

It is difficult interpret the scores in Table V on relative width and the CRPS in terms of estimation performance without comparable values from other datasets. These scores may however be useful for comparing the performance of the actual estimates with alternative estimation strategies. This is what we do below.

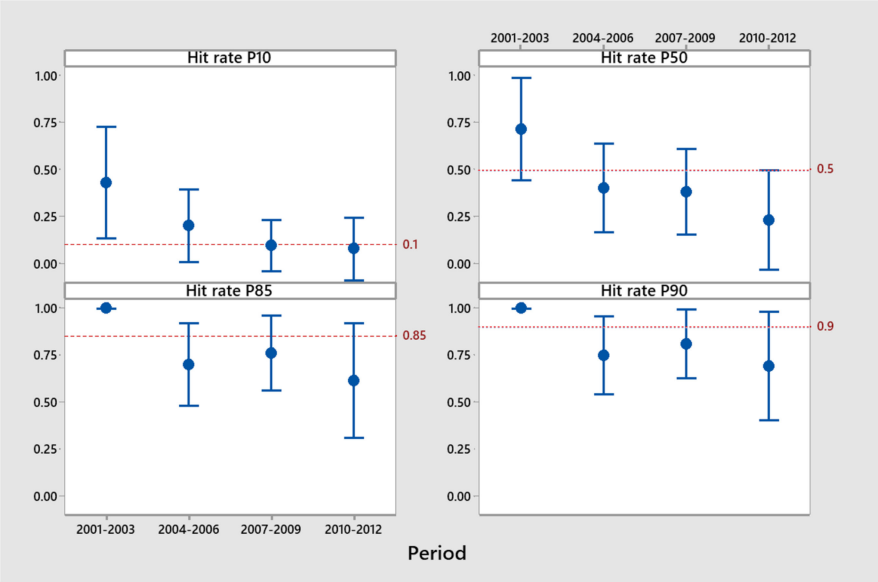

Fig. 6. Hit rates over time.

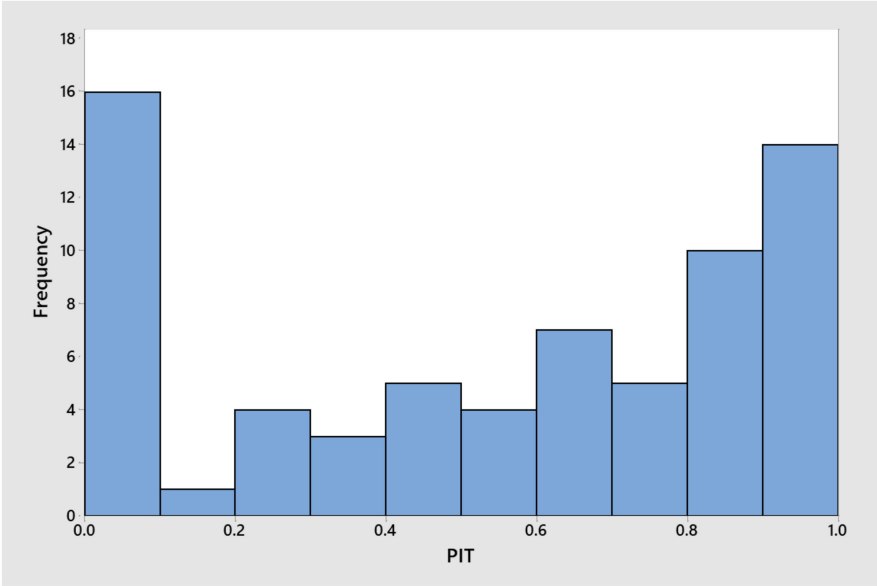

Fig. 7. Histogram of PIT values.

1) Comparison With (Hypothetical) Estimates Provided by a Focused, Overall Uncertainty-Oriented Estimator: As pointed out in Section II-C-1, a focused overall uncertainty-oriented estimator gives perfectly calibrated prediction intervals without being able to separate low and high cost uncertainty projects. If the results from the use of that strategy are similar to our cost uncertainty estimation results, this suggests that the current (expensive) cost uncertainty estimation process could be replaced by a much simpler process. In addition, it might suggest that the estimators provided estimates using such a non-informative strategy.

We implemented the focused overall uncertainty-oriented strategy for the analyzed projects by assuming that the (hypothetical) estimators had information about the overall (aggregated) cost uncertainty of the projects, and used this to provide new *PX* estimates. This may have been the case if, for example, the estimator knew based on her/his experience that 85%of prior projects had an actual cost of less than X times the estimated most likely cost, and they multiplied all estimates of the most likely cost with a factor of X to find the P85 estimate. We derived the overall cost uncertainty distribution by fitting different types of distribution to the actual cost estimation error data. The best fit was for a normal distribution with $\mu = 0.05$ and $\sigma = 0.20$, where multiplying the estimate of the most likely cost by 1.20 would give the P85 value (the 85th percentile).

TABLE V
CALIBRATION AND INFORMATIVENESS

Measures	Result
Calibration	
Hit rates	P10 interval: 19% (p=0.04)
	P50 interval: 42% (p=0.23)
	P85 interval: 75% (p=0.04)
The p-value indicates the probability of observing a difference (or larger differences) from the normative percentage (e.g., 10% for P10), given that the true difference is zero (binomial test of one-sample proportion).	P90 interval: 80% (p=0.01)
PIT values	See Figure 7 for a histogram of the PIT values
Informativeness	
Mean relative width of the 80% prediction intervals, calculated as: (P90-P10)/P50	0.29
Spearman's rank-correlation between relative width for 80% prediction intervals and log-error.	-0.02 (p=0.85)
Calibration and informativeness	
Median CRPS	109

We applied this focused, overall uncertainty-oriented strategy using the estimated most likely cost and assuming the fitted cost uncertainty distribution for all projects. This gave better calibration results, but worse informativeness results, than was the case for the actual cost uncertainty estimates.

1) *Calibration:* The P10 estimates from our hypothetical estimator had a hit rate of 13% instead of 19%, the P50 estimates had a hit rate of 48% instead of 42%, the P85 estimates had a hit rate of 90% instead of 75%, and the P90 estimates had a hit rate of 93% instead of 80%. These data demonstrate a clear improvement in calibration by use of the focused, overall uncertainty-oriented strategy.
2) *Informativeness:* The 80% prediction intervals had a relative width of 0.49 for all projects, and there was consequently zero correlation between the estimated and actual cost uncertainty. While a correlation of zero is similar to that of the analyzed projects, a mean relative width of the 80% prediction interval of 0.49 is substantially higher than the mean relative width of 0.29 of the analyzed projects (Wilcoxon signed-rank test of median differences of relative width gives $p<0.001$). This lower informativeness is, as expected, connected with higher P85 estimates. While the required contingency, when going from the most likely cost to the P85 estimate, was 26% when using the overall uncertainty-oriented, the actual mean contingency was only 13%.

The CRPS score for the focused, overall uncertainty-oriented strategy is better than that of the actual projects, with a median value of 88 compared to the previously found value of 109. This difference is statistically significant (Wilcoxon signed-rank test of median differences of the CRPS scores gives $p<0.001$)

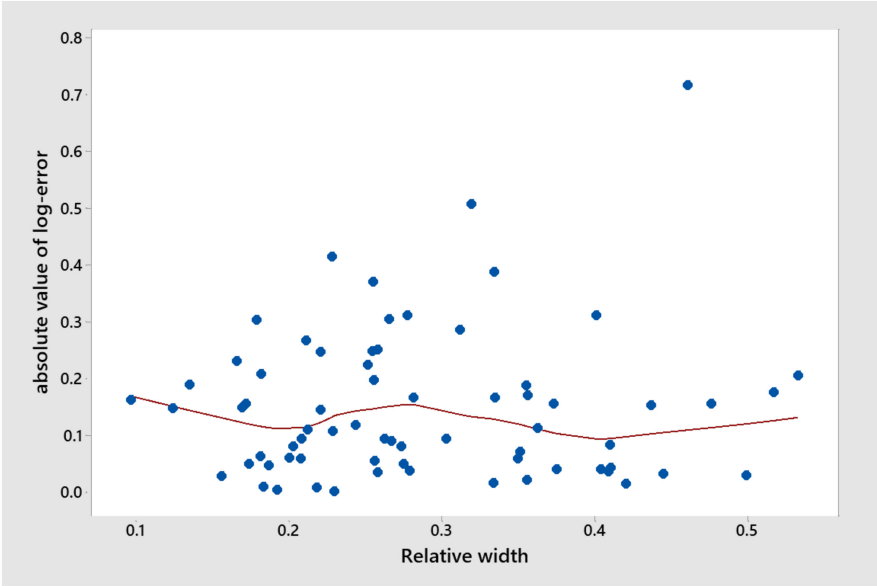

Fig. 8. Association between the estimated (as indicated by the relative width of the 80% prediction interval) and the actual cost uncertainty (as indicated by the absolute value of the log-error).

but hard to interpret, as the CRPS score reflects a weighted combination of the estimates' calibration and informativeness.

In total, the use of a simple focused, overall uncertainty-oriented strategy led to near-perfect calibration at the cost of wider cost distributions and a need for higher cost contingencies. This means that even though the estimates of cost uncertainty in the analyzed set of projects have clear improvement potential, the organizations may not benefit from replacing the current estimation process with the simpler, focused, overall uncertainty-oriented strategy.

IV. Implications and Limitations

We have argued that the use of probabilistic cost estimates supports the need for more precise interpretation and communication of what is meant by a cost estimate. The use of such estimates also enables us to analyze the match between the type of estimate and the chosen cost estimation evaluation measure, and to analyze characteristics of the cost prediction intervals and cost distributions. This section discusses implications of the proposed guidelines on how to evaluate probabilistic cost estimates (see Section IV-A) and the limitations of the proposed guidelines (see Section IV-B).

A. Implications

The guidelines require the use of probabilistic cost estimates. There is much literature on how to establish probabilistic frameworks for predictions, see for example [20], [42]. Without the use of such cost estimates, it is hard to see how cost estimates can be given a precise meaning and how one can assess the meaningfulness and fairness of cost estimation evaluation measures. Thus, a key message of the present article is to avoid the common practice of obtaining "some" cost estimates with unknown interpretations and evaluating them with "some" accuracy or bias measure without knowing to what extent the evaluation gives a fair and unbiased evaluation. An advantage of using a probabilistic cost estimation framework is that it may remind the evaluator on that even the best cost estimates will not have an expected estimation error of zero. The lowest achievable estimation error increases as a function of the level of uncertainty in the projects. An implication of this is that, whenever possible, an evaluation of cost estimation error performance should take into account the level of underlying cost uncertainty.

The proposed criterion of a match between the type of cost estimate and the evaluation measure has several implications for the use of evaluation measures in industry and research. The main implication is that without a match between estimates and evaluation measures, we will not know to what extent measured cost overruns and errors are results of true estimation bias and error, or just inappropriate use of evaluation measures. Not only will this be unfortunate for the interpretation of the measurements, but it may also establish unfortunate incentives where the estimators are rewarded for underestimating or overestimating the cost.

An example of lack of match, with potential unfortunate incentives, is the use of the error measure *mean absolute relative error* ($\frac{1}{N}\sum_{i=1}^{N}\frac{|\mathrm{act}_i-\mathrm{est}_i|}{\mathrm{est}_i}$). This error measure is not minimized by any commonly used type of cost estimate. If the cost uncertainty, for example, is log-normally distributed,[9] the type of estimate that minimizes the error is higher than both the mean and the median cost. Estimators will consequently typically be rewarded for over-estimation of the mean and the median cost. If the cost uncertainty is high and strongly non-symmetric, this evaluation may reward strongly biased cost estimates. Correspondingly, when dividing by the actual instead of the estimated

[9]See previous discussion on the good fit of log-normal distributions, and also [18].

cost ($\frac{1}{N}\sum_{i=1}^{N}\frac{|\mathrm{act}_i-\mathrm{est}_i|}{\mathrm{act}_i}$), which also is an error measure in common use, the type of estimate that minimizes the error is lower than both the mean and the median cost.[10] In this case the evaluation measure rewards under-estimation of the cost. The use of any of the two versions of the mean absolute relative error is not advisable, because they do not match any common or well-defined types of point estimates. We recommend instead the use of the mean absolute log-error ($\frac{1}{N}\sum_{i=1}^{N}|\ln(\mathrm{act}_i)-\ln(\mathrm{est}_i)|$), which has a match with estimates of the median cost (P50 estimates).

There are similar implications for the evaluation of cost estimation bias. We have, for example, that the bias measure *mean relative error* ($\frac{1}{N}\sum_{i=1}^{N}\frac{(\mathrm{act}_i-\mathrm{est}_i)}{\mathrm{est}_i}$) should only be used when the intended meaning of the estimate is the mean (expected) cost. Note that this match between evaluation measure and type of estimate is no longer the case if we divide by the actual instead of the estimated cost, which is common in some contexts, see for example [43].[11] This measure, i.e., the measure ($\frac{1}{N}\sum_{i=1}^{N}\frac{(\mathrm{act}_i-\mathrm{est}_i)}{\mathrm{act}_i}$), is consequently not advisable to use as measure of cost overruns. When the intended meaning of the estimates is the median cost, i.e., when we use P50 cost estimates, we will have a match for the bias measure *median relative error*. In this case, the use of mean relative error will give unfortunate incentives and misleading evaluations.

All types of *PX* cost estimates, e.g., P10 and P90 cost estimates, may in principle be evaluated as single-point cost estimates. Unfortunately, the *only* matching error measure for single-point cost estimates on the *PX* format, see [36], is: $\frac{1}{N}\sum_{i=1}^{N}\begin{cases}\alpha\cdot|\mathrm{act}_i-\mathrm{est}_i|, & if\ \mathrm{est}_i\le\mathrm{act}_i\\(1-\alpha)\cdot|\mathrm{act}_i-\mathrm{est}_i|, & otherwise\end{cases}$, where α is the estimated quantile (*X*-value of the *PX*-estimate). For example, in the case of P85 estimates

$$\frac{1}{N}\sum_{i=1}^{N}\begin{cases}0.85\cdot|\mathrm{act}_i-\mathrm{est}_i|, & if\ \mathrm{est}_i\le\mathrm{act}_i\\0.15\cdot|\mathrm{act}_i-\mathrm{est}_i|, & otherwise.\end{cases}$$

This error measure is only meaningful when we believe that overrunning the *PX* estimate should give an error penalty α /(1- α) times higher than that of underrunning it. For example, to use this measure for P85 cost estimates, we should agree on that overrunning the P85 cost estimate with one cost unit is 5.67 (0.85/0.15) times worse than underrunning it with one unit. This is not necessarily the case and many types of *PX* estimates should for this reason only be evaluated with respect to calibration and informativeness. As there, in general, are no other matching error measure for *PX* cost estimates, the commonly used cost estimation error measures should not be used for these cost estimates. The exception here is, of course, the use of the P50 cost estimates, which, as described above, has other matching error measures.

[10]The minimizing estimate for a log-normal distribution equals in this case $e^{\mu+\sigma^2}$, where μ is the mean and σ is the standard deviation of the distribution.

[11]In this case, no common understandings of cost estimates would give a match. Perfect estimates of the mean cost with actual cost as the denominator, for example, would give the expectation of a mean relative error (cost overrun) of approx. $\frac{\mathrm{Var(act)}}{\mu^2}$, see [44].

Our study of the Norwegian projects resulted in additional implications for proper evaluation of cost estimates. In particular, we found that a good score on mean or median cost estimation bias needed in-depth analysis to be used as an indicator of good estimation performance. In our project dataset, the low average cost overrun was caused by the presence of substantial cost underrun of the early projects, compensated by substantial cost overruns of the later projects. In this case, most estimates were biased, but there was a change in the direction of the bias over time. Had the cost estimates of the later projects been unbiased, the total bias would have been worse. We recommend, for this reason, the use of trend analyses to get better insight into how low (or high) cost overruns have been achieved.

Our guidelines also imply potential for improvements in the evaluation of cost prediction intervals and distributions. Our most important point is that the evaluations should be extended to include informativeness. Traditionally, the evaluation of cost prediction intervals has focused on calibration, i.e., on comparing the stated confidence level with the actual inclusion rate (hit rate). The analysis of the data in the case study demonstrates that good calibration is not sufficient to claim high performance on cost prediction intervals. In addition to calibration, the prediction intervals and distributions should be evaluated in terms of informativeness, such as the correlation between indicators of estimated and the actual cost uncertainty. An important reason for the need to evaluate informativeness is that there are simple uncertainty assessment strategies that achieve close to perfectly calibrated cost prediction intervals but are non-informative about differences in underlying cost uncertainty.

An alternative to reporting both calibration and informativeness is to use a measure that combines calibration and informativeness, such as the CRPS measure. As this combination-based evaluation measures imply predefined weighting of the importance of each factor, we find that reporting both calibration and informativeness separately is likely to be the better option, leaving it to the user to weigh their relative importance.

B. Limitations

There are several topics that are outside the scope of the guidelines presented in this article. Amongst others, the guidelines do not address how to produce probabilistic cost estimates, they do not cover all aspects of what is needed to properly evaluate probabilistic cost estimates, and they do not address how to analyze the reasons for low and high cost estimation performance. These topics are important to succeed with better cost estimation and there is a need for more research on all of them.

We chose not to include much discussions on the relationship between the choice of evaluation measure and loss functions. If an organization evaluates cost estimates in terms of the accuracy measure mean absolute error ($\frac{1}{N}\sum_{i=1}^{N}|\mathrm{act}_i-\mathrm{est}_i|$), it chooses a loss function that penalizes cost overruns just as much as cost underruns. This may not reflect the true losses of the organization, which may prefer cost underruns to overruns. In spite of the intuitive appeal of selecting cost evaluation measures that reflect the true loss function of an organization, work on this

has not been very promising. True loss functions turn out to be difficult to formulate. Furthermore, they vary over time, they vary between and within organizations, and the corresponding evaluation measures may be difficult to understand, see for example [45]. It may, nevertheless, be useful to reflect on the implied loss situation when selecting a cost estimation evaluation measure.

Probabilistic cost estimates enable precise communication of cost estimates and provide a framework for matching types of estimate and evaluation measures, but may also be difficult to understand and use. As an illustration, the CEO of a large Norwegian government agency reported that all of their projects in the last five years had an actual cost below their portfolio level budget, which was based on their P85 estimates of the cost. While a 100% hit rate of P85 estimates is a strong indication of overestimation, the CEO's conclusion was that this showed that his organization had been very good at managing its projects [46]. To understand and use probabilistic frameworks properly may require some training, see [47] for an overview of some of the challenges and possible solutions to improved understanding and use of probabilities. While the complexity related to the understanding and use of probabilities is a real challenge, we believe that there is no proper way around the use and evaluation of it, unless we want to continue using cost estimates with unknown meaning and select evaluation measures that do not reward the most realistic cost estimates.

V. Conclusion

To carry out meaningful cost estimation evaluations, we need to know what was meant by the cost estimates and to had a proper evaluation framework. In this article, we argue that the use of probabilistic cost estimates enables precision in communicating what was meant by the estimated cost, and supports the implementation of fair cost estimation evaluation measures. We propose a match criterion for measures of cost estimation accuracy and bias. This match criterion implies that measures of estimation error and bias should give the optimal scores to the best possible cost estimates. Best possible cost estimates are here understood as estimates that are perfect representations of the intended position in the underlying cost distributions. It was, for example, the case when the estimated mean costs of a set of projects equals the actual mean costs of the projects' underlying cost distributions. We also propose that cost prediction intervals and distributions should not only be evaluated with respect to calibration, but also include an evaluation of informativeness. This point is particularly important because evaluation of calibration alone fails to distinguish between estimators with and without ability to separate projects with high and low cost uncertainty.

The feasibility of the proposed evaluation guidelines was examined through an analysis of probabilistic cost estimates from 69 large Norwegian governmental projects. The analysis supported the feasibility and usefulness of the guidelines. We conclude that the proposed guidelines, applied on probabilistic cost estimates, are promising in providing guidance for meaningful, fair, and useful evaluation of cost estimation performance.

Appendix

1) $[\frac{1}{N}\sum_{i=1}^{N}(\mathrm{act}_i - \mu_i)] = \frac{1}{N}\cdot[E(\mathrm{act}_1 - \mu_1) + \ldots + E(\mathrm{act}_N - \mu_N)] = \frac{1}{N}\cdot[E(act_1) - E(\mu_1) + \ldots + E(\mathrm{act}_N) - E(\mu_N)] = \frac{1}{N}\cdot[E(\mu_1) - E(\mu_1) + \ldots + E(\mu_N) - E(\mu_N)] = 0$, where $E(X)$ is the expected value of X and μ_i is the mean value of cost distribution i. $E(X+Y) = E(X)+E(Y)$ due to linearity of the mean values for independent distributions.
2) When using the median as the estimate, there is an equal probability of actual costs being higher or lower than the estimate. We then have the expectation that half of the actual values will be higher and half will be lower than the median, and that the median difference will be zero. This does not change when dividing the difference by estimated cost or log-transforming the values.
3) $E[\frac{1}{N}\sum_{i=1}^{N}\frac{(\mathrm{act}_i-\mu_i)}{\mu_i}] = \frac{1}{N}\cdot[E(\frac{\mathrm{act}_1-\mu_1}{\mu_1}) + \ldots + E(\frac{\mathrm{act}_N-\mu_N}{\mu_N})] = \frac{1}{N}\cdot[E(\mathrm{act}_1 - \mu_1)\cdot E(\frac{1}{\mu_1}) + \ldots + E(\mathrm{act}_N - \mu_N)\cdot E(\frac{1}{\mu_N})] = \frac{1}{N}\cdot[(E(\mu_1) - E(\mu_1))\cdot E(\frac{1}{\mu_1}) + \ldots + (E(\mu_N) - E(\mu_N))\cdot E(\frac{1}{\mu_N})] = 0.$

References

[1] L. A. Smith and T. Mandakovic, "Estimating: The input into good project planning," *IEEE Trans. Eng. Manage.*, vol. EM-32, no. 4, pp. 181–185, Nov. 1985.

[2] B. Flyvbjerg, M. S. Holm, and S. Buhl, "Underestimating costs in public works projects: Error or lie?," *J. Amer. Planning Assoc.,* vol. 68, no. 3, pp. 279–295, 2002.

[3] B. Flyvbjerg, M. K. Skamris Holm, and S. L. Buhl, "What causes cost overrun in transport infrastructure projects?," *Transport Rev.,* vol. 24, no. 1, pp. 3–18, 2004.

[4] A. Aljohani, D. Ahiaga-Dagbui, and D. Moore, "Construction projects cost overrun: What does the literature tell us?," *Int. J. Innov. Manage. Technol.,* vol. 8, no. 2, p. 137, 2017.

[5] P. E. Love, J. Zhou, D. J. Edwards, Z. Irani, and C.-P. Sing, "Off the rails: The cost performance of infrastructure rail projects," *Transp. Res. Part A Policy Pract.,* vol. 99, pp. 14–29, 2017.

[6] J. Odeck, "Cost overruns in road construction—What are their sizes and determinants?," *Transport Policy,* vol. 11, no. 1, pp. 43–53, 2004.

[7] J. Odeck, "Variation in cost overruns of transportation projects: An econometric meta-regression analysis of studies reported in the literature," *Transportation,* vol. 46, no. 4, pp. 1345–1368, 2019.

[8] P. E. Love and D. D. Ahiaga-Dagbui, "Debunking fake news in a post-truth era: The plausible untruths of cost underestimation in transport infrastructure projects," *Transp. Res. Part A, Policy Pract.*, vol. 113, pp. 357–368, 2018.

[9] B. Andersen, K. Samset, and M. Welde, "Low estimates–high stakes: Underestimation of costs at the front-end of projects," *Int. J. Manag. Projects Bus.,* vol. 9, pp. 171–193, 2016.

[10] D. C. Invernizzi, G. Locatelli, and N. J. Brookes, "Cost overruns–helping to define what they really mean," in *Proc. Inst. Civil Engineers-Civil Eng.*, 2017, vol. 171, pp. 85–90.

[11] "Cost engineering terminology." *AACE Int., Recommended practice 10S-90: Cost engineering terminology*. Accessed: Nov. 13, 2020. [Online]. Available: https://web.aacei.org/docs/default-source/rps/10s-90.pdf?sfvrsn=58

[12] C. Sauer, A. Gemino, and B. H. Reich, "The impact of size and volatility on IT project performance," *Commun. ACM,* vol. 50, no. 11, pp. 79–84, 2007.

[13] T. Gneiting, "Making and evaluating point forecasts," *J. Amer. Statist. Assoc.,* vol. 106, no. 494, pp. 746–762, Jun 2011.

[14] K. Mitchell and C. Ferro, "Proper scoring rules for interval probabilistic forecasts," *Quart. J. Roy. Meteorolog. Soc.,* vol. 143, no. 704, pp. 1597–1607, 2017.

[15] T. Gneiting and M. Katzfuss, "Probabilistic forecasting," *Annu. Rev. Statist. Appl.,* vol. 1, pp. 125–151, 2014.

[16] S. Grimstad and M. Jørgensen, "A framework for the analysis of software cost estimation accuracy," in *Proc. ACM/IEEE Int. Symp. Empirical Softw. Eng.*, 2006, pp. 58–65.
[17] E. Limpert, W. A. Stahel, and M. Abbt, "Log-normal distributions across the sciences: Keys and clues: On the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—Normal or log-normal: That is the question," *BioScience,* vol. 51, no. 5, pp. 341–352, 2001.
[18] T. Halkjelsvik and M. Jørgensen, "Predictions and the uncertainty of the future," in *Time Predictions*. Berlin, Germany: Springer, 2018, pp. 13–33.
[19] G. A. Barraza, W. E. Back, and F. Mata, "Probabilistic forecasting of project performance using stochastic S curves," *J. Construction Eng. Manage.,* vol. 130, no. 1, pp. 25–32, 2004.
[20] J. E. Diekmann, "Probabilistic estimating: Mathematics and applications," *J. Construction Eng. Manage.,* vol. 109, no. 3, pp. 297–308, 1983.
[21] C. W. Granger and M. H. Pesaran, "Economic and statistical measures of forecast accuracy," *J. Forecasting,* vol. 19, no. 7, pp. 537–560, 2000.
[22] A. Carvalho, "An overview of applications of proper scoring rules," *Decis. Anal.,* vol. 13, no. 4, pp. 223–242, 2016.
[23] V. R. R. Jose, "Percentage and relative error measures in forecast evaluation," *Oper. Res.,* vol. 65, no. 1, pp. 200–211, 2017.
[24] L. Törnqvist, P. Vartia, and Y. O. Vartia, "How should relative changes be measured?," *Amer. Statist.,* vol. 39, no. 1, pp. 43–46, 1985.
[25] M. Welde and J. Odeck, "Cost escalations in the front-end of projects–empirical evidence from norwegian road projects," *Transport Rev.,* vol. 37, no. 5, pp. 612–630, 2017.
[26] C. R. McKenzie, M. J. Liersch, and I. Yaniv, "Overconfidence in interval estimates: What does expertise buy you?," *Organ. Behav. Human Decis. Processes*, vol. 107, no. 2, pp. 179–191, 2008.
[27] T. Gneiting, F. Balabdaoui, and A. E. Raftery, "Probabilistic forecasts, calibration and sharpness," *J. Roy. Statist. Soc., Ser. B (Statist. Methodol.),* vol. 69, no. 2, pp. 243–268, 2007.
[28] G. Van de Venter and D. Michayluk, "An insight into overconfidence in the forecasting abilities of financial advisors," *Australian J. Manage.,* vol. 32, no. 3, pp. 545–557, 2008.
[29] J. S. Armstrong and K. Green, "The forecasting dictionary," in *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Berlin, Germany: Springer, 2001, pp. 761–819.
[30] M. Jorgensen, "Realism in assessment of effort estimation uncertainty: It matters how you ask," *IEEE Trans. Softw. Eng.*, vol. 30, no. 4, pp. 209–217, Apr. 2004.
[31] M. Jørgensen, "Evaluating probabilistic software development effort estimates: Maximizing informativeness subject to calibration," *Inf. Softw. Technol.,* vol. 115, pp. 93–96, 2019.
[32] P. Friederichs and T. L. Thorarinsdottir, "Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction," *Environmetrics,* vol. 23, no. 7, pp. 579–594, 2012.
[33] "Kartlegging av tid og kostnader ved KS-ordningen ('Analysis of the time and cost of the quality assurance scheme')," EY, London, U.K. 2016. [Online]. Available: https://www.regjeringen.no/contentassets/c4f336fcd28746b0807d645cda98b0ad/24112016_ks-ordning.pdf
[34] G. H. Volden, "Up-Front governance of major public investment projects," Ph.D. dissertation, Norwegian Univ. Sci. Technol., Trondheim, Norway, 2019.
[35] I. Bukkestein, "Kartlegging av status for prosjekter som har vært gjennom KS2," Trondheim, Norway: Norwegian Univ. Sci. Technol., 2020.
[36] T. Gneiting, "Quantiles as optimal point forecasts," *Int. J. Forecasting*, vol. 27, no. 2, pp. 197–207, 2011.
[37] T. Halkjelsvik and M. Jørgensen, "From origami to software development: A review of studies on judgment-based predictions of performance time," *Psychol. Bull.,* vol. 138, no. 2, pp. 238–271, 2012.
[38] C. C. Cantarelli, B. van Wee, E. J. Molin, and B. Flyvbjerg, "Different cost performance: Different determinants?: The case of cost overruns in Dutch transport infrastructure projects," *Transport Policy,* vol. 22, pp. 88–95, 2012.
[39] A. Budzier and B. Flyvbjerg, "Double whammy-how ICT projects are fooled by randomness and screwed by political intent," 2013.
[40] J. L. Eveleens and C. Verhoef, "Quantifying IT forecast quality," *Sci. Comput. Prog..,* vol. 74, no. 11/12, pp. 934–988, 2009.
[41] J. Bertisen and G. A. Davis, "Bias and error in mine project capital cost estimation," *Eng. Economist,* vol. 53, no. 2, pp. 118–139, 2008.
[42] T. Halkjelsvik and M. Jørgensen, *Time Predictions: Understanding and Avoiding Unrealism in Project Planning and Everyday Life*. Berlin, Germany: Springer, 2018.
[43] V. Mahnic, "A capstone course on agile software development using scrum," *IEEE Trans. Educ.,* vol. 55, no. 1, pp. 99–106, Feb. 2011.
[44] A. Stuart, S. Arnold, J. K. Ord, A. O'Hagan, and J. Forster, *Kendall's Advanced Theory of Statistics*. Hoboken, NJ, USA: Wiley, 1994.
[45] J. S. Armstrong and R. Fildes, "Correspondence on the selection of error measures for comparisons among forecasting methods," *J. Forecasting*, vol. 14, no. 1, pp. 67–71, 1995.
[46] "God styring i statsbygg," Byggeindustrien, Tidaholm Sweden, [Online]. Available: http://www.bygg.no/article/1387402
[47] C. Batanero, E. J. Chernoff, J. Engel, H. S. Lee, and E. Sánchez, *Research on Teaching and Learning Probability*. Berlin, Germany: Springer, 2016.