Journal of Benefit-Cost Analysis (2024), 1–16
doi:10.1017/bca.2024.29

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Forecast Errors and Welfare Conclusions Based on the Flyvbjerg Database

Timo Välilä

The Bartlett School of Sustainable Construction, University College London, London, United Kingdom of Great Britain and Northern Ireland
Email: t.valila@ucl.ac.uk

Keywords: major investment projects; forecast errors; benefit–cost analysis

JEL codes: D61; H43; H54

Abstract

The so-called Flyvbjerg database is the largest source of data on the performance of major investment projects. It has generated influential analyses of the magnitude of and reasons for cost overruns and demand shortfalls in major projects. Those analyses have demonstrated, among other things, the systematic presence of large forecast errors in both construction costs and in user demand in the first year of operation. They have also linked those results to the social welfare consequences of the underlying projects, suggesting that the large and systematic forecast errors are indicative of welfare destruction. Given how influential those analyses have been, this paper examines the link between the database, empirical analyses thereof, and social benefit–cost analysis (BCA). To that end, both the measurement of variables in the database and the estimation of forecast errors are contrasted against BCA. The conditions for the estimated forecast errors to approximate those obtained from a BCA are spelled out, and the scope for drawing welfare conclusions based on those estimates is discussed. Furthermore, numerical simulations are presented to explore whether the estimated forecast errors do indeed imply likely welfare destruction. The simulations suggest that as large as the forecast errors are, welfare destruction is no foregone conclusion.

1. Introduction

The so-called Flyvbjerg database, thus labeled in Flyvbjerg and Gardner (2023, Appendix A), is the most extensive, albeit not freely accessible, source in existence of information about the performance of individual major projects. It has grown over time into a meta-database comprising data from both primary and secondary sources (Flyvbjerg & Bester, 2021, Appendix) with global coverage, now containing data on over 16,000 projects of 25 distinct "types" (Flyvbjerg & Gardner, 2023, Appendix A).

The database has given rise to a large number of analyses of the magnitude of errors in forecasting project costs and demand, as well as reasons for the emergence and systematic character of those errors. Those analyses have been based on different generations of

samples from the database. Flyvbjerg et al. (2002); Flyvbjerg et al. (2003); as well as Flyvbjerg (2009) all analyze 258 transport infrastructure projects. Ansar et al. (2016) analyze a sample of 95 road (74) and rail (21) infrastructure projects in China. Flyvbjerg (2016) as well as Flyvbjerg and Bester (2021) analyze a sample of 2,062 projects in eight infrastructure sectors and report detailed results for a subsample of 1,603 projects. Flyvbjerg and Gardner (2023, Appendix A), refer to the entire database of 16,000 projects.

A central objective for the development of the database has been to create as extensive a sample of comparable project-level information as possible, so as to enable systematic quantitative analyses of their performance. To that end, the data collection has been focused on variables that are most commonly available, notably the cost of constructing new major assets and the demand for those assets (Flyvbjerg & Bester, 2021). Both variables are measured first at the point of decision-making about the project, representing the forecasts on which the decision is based, and subsequently at the completion of asset construction and at the end of the first year of operation, respectively. Based on these observations, forecast errors measured as ratios of actual outcomes to estimates at decision-making have then been calculated and reported for the different project types (Flyvbjerg, 2016; Flyvbjerg & Bester, 2021).[1]

It is obvious that the data available in the database do not allow for the carrying out of social benefit–cost analyses (BCA) of the projects, nor has that ever been a stated objective of the articles drawing on it. Flyvbjerg and Bester (2021) acknowledge that the results of their analysis are "…as much about impact prediction as about cost–benefit analysis…" (p. 398). What exactly that impact refers to is unclear, however. Elsewhere in the same article, the authors make explicit references to the economic efficiency and social welfare consequences of the projects analyzed, including: "We document the extent of cost overruns and benefit shortfalls, and forecasting bias in public investments. We further assess whether such inaccuracies seriously distort effective resource allocation, which is found to be the case" (p. 395). It is not clear whether the reference to "effective resource allocation" is intentional or meant to refer to efficient resource allocation. Further, based on observed forecast errors in construction costs and first-year demand, the authors conclude: "Such systematic and significant bias in cost–benefit analysis is likely to lead to resource misallocation, including initiating investments that ultimately turn out to have negative net benefits and should never have been started" (p. 401).

Also, other articles based on the database draw conclusions about the economic efficiency and social welfare consequences of the projects therein. Analyzing forecast errors in construction costs, Flyvbjerg et al. (2002, pp. 290–291) conclude that their underestimation is "likely to lead to the misallocation of scarce resources." Flyvbjerg et al. (2003, p. 86) suggest that the underestimation of construction costs is "to the detriment of social and economic welfare." Flyvbjerg (2009, p. 353) concludes that the projects that "look best on paper" are implemented but become "the worst, or unfit test, projects in reality, in the sense that they are the very projects that will encounter most problems during construction and operations in terms of the largest cost overruns, benefit shortfalls, and risks of nonviability." Flyvbjerg (2016, p. 182) mentions the simultaneous presence of cost overruns and benefit shortfalls, which are "bad for viability" of the "average project."

[1] In earlier work, Flyvbjerg et al. (2002) as well as Flyvbjerg (2005) define forecasting errors as the difference between the actual outcome and the estimate, relative to the estimate. The measure used subsequently by Flyvbjerg (2016) as well as Flyvbjerg and Bester (2021) is just a simple transformation of that earlier definition.

https://doi.org/10.1017/bca.2024.29 Published online by Cambridge University Press

The merits of the Flyvbjerg database as a source of research into the magnitude, systematic nature, and, importantly, determinants of forecasting errors in major projects are beyond question. Ever since Flyvbjerg et al. (2002), Flyvberg and co-authors have emphasized the systematic nature of forecasting errors observed in major projects, distinguishing them from purely random errors with the expected value of zero that one would observe in any unbiased forecasting exercise. This has led them to study the types and sources of biases in forecasting and also to recommend remedies against such biases, notably in Flyvbjerg and Bester (2021).

Given the available data on individual projects in the database, the rather far-reaching conclusions quoted above concerning those projects' consequences for social welfare are much less beyond question. Therefore, the objective of this article is to explore the links between the Flyvbjerg database, analyses of forecast errors based thereon, and BCA. This is done to assess to what extent the data in the Flyvbjerg database do indeed allow analyses that can replicate BCA at least partly and, consequently, to what extent the results of such analyses can be interpreted as conveying useful information about the social welfare consequences of the projects in the database.

It should be emphasized that the justification for and importance of a closer assessment of the link between the Flyvbjerg database and social welfare goes beyond merely challenging the use of specific terms or the exact wording of conclusions in articles based on the database. The research outputs and conclusions drawn on the database have been highly visible, impactful, and influential also for public policies. It is therefore important to understand clearly whether and to what extent the database does indeed permit the drawing of general and far-reaching conclusions about the societal desirability of major projects, rather than conclusions about just forecasting errors.

The next section contrasts the measurement of project costs and benefits in the database with BCA, and it discusses the scope for drawing conclusions about the quality and societal desirability of the projects in the database from a conceptual perspective. Given the empirical evidence presented in articles based on the database, Section 3 undertakes to analyze to what extent that evidence does indeed confirm beyond any reasonable doubt the conclusions concerning the likely negative social welfare consequences of the projects in the database, or whether it is possible that the evidence could be compatible with positive social welfare consequences as well. The results from sections 2 and 3 are combined and discussed in the concluding section 4.

2. Flyvbjerg database and benefit–cost analysis

2.1 Measurement of project costs and benefits

The measurement of the costs and benefits of the projects included in the Flyvbjerg database is described in Flyvbjerg et al. (2002); Flyvbjerg et al. (2003); Flyvbjerg (2005); Næss et al. (2006); Flyvbjerg (2009); Ansar et al. (2016); Flyvbjerg (2016); as well as Flyvbjerg and Bester (2021). This section summarizes the description of the measurement principles and contrasts them with BCA.

Project costs in the Flyvbjerg database refer to construction costs, including both a forecast at the decision-making stage and the actual outcome at the completion of construction. Ansar et al. (2016), considering a sample of 95 road and rail infrastructure projects in China, present the most detailed breakdown of construction costs among the articles

https://doi.org/10.1017/bca.2024.29 Published online by Cambridge University Press

referring to the database. They include the costs of "… right-of-way acquisition and resettlement; design engineering and project management services; construction of all civil works; as well as equipment purchases excluding rolling stock" (p. 368). Excluded are also debt payments, taxes, as well as the cost of any ex-postenvironmental remedial work.

It is unclear whether the data underlying the other articles consider these same cost components, as they just refer to construction or capital costs without any further detail. Flyvbjerg and Bester (2021) recognize that while the measurement of life cycle costs would be preferable, the data to that end are not available. Flyvbjerg (2009) mentions explicitly the exclusion of financing, operation, and maintenance costs from the subsample of 258 transport infrastructure projects analyzed. None of the articles refers to nonmarket flows associated with the projects, such as deadweight losses due to taxation required to finance and fund the projects, or externalities during operation. Therefore, and given the relative difficulty of collecting such data, it is likely that nonmarket flows are also excluded from the database.

The valuation of construction costs is not explicitly explained in the articles listed above, but the sources of data described in them suggest that the costs are likely valued at market or financial accounting prices. The cost data are expressed in real terms and one currency; however, none of the articles contains any details about the exact deflators used or exchange rates applied. Flyvbjerg and Bester (2021) discuss the issue of shadow pricing, albeit of benefits, indicating that they do not apply it.

In sum, the costs in the Flyvbjerg database refer to construction costs, measured in financial rather than in economic terms. The main adjustments needed to arrive at an economic measure of costs would include the addition of life cycle costs other than during construction; the consideration of nonmarket flows; as well as the use of shadow prices whenever there is a wedge between them and market prices.

Turning to project benefits in the Flyvbjerg database, Flyvbjerg and Bester (2021) explain that the benefits are measured in the unit used by project planners. While no further information about such units is provided, articles analyzing the benefits of projects in the transport sector (Flyvbjerg, 2005; Næss et al., 2006; Flyvbjerg, 2009; Ansar et al., 2016) refer to traffic volume (number of vehicles or passengers). These physical measures of traffic volume (demand) are referred to as benefits.

For reasons of data availability, project benefits (e.g., traffic demand) are measured only at decision-making (expected quantity) and during the first year of operation (actual outcome). Flyvbjerg and Bester (2021) recognize that it would be preferable to analyze benefits during the project's entire life cycle. However, such data are rarely available, and to construct a sufficiently large sample of projects for statistical analysis, the measurement of benefits is limited to those two observations.

Furthermore, Flyvbjerg (2005) as well as Flyvbjerg and Bester (2021) argue that projects with demand and benefit shortfalls in the first year of operation tend to have shortfalls later on as well, and while the ramp-up of demand is often assumed, it often fails to materialize, or materializes only partly. Flyvbjerg (2005) suggests that forecast errors of ramp-up are likely to have a relatively small impact on the present value of total benefits anyway; forecast errors of total demand are more important. For all these reasons, the focus on the expected demand at decision-making and its actual outcome in the first year of operation is considered justified, if second-best.

Given that only data on first-year demand forecast errors are available, all conclusions concerning benefits must be derived from those forecast errors. In a way, this makes the

https://doi.org/10.1017/bca.2024.29 Published online by Cambridge University Press

analysis and its results rather trivial. However, it is still worthwhile to articulate the approach formally, so as to spell out the implicit assumptions underlying it and the limitations to the conclusions that can be drawn.

Consider first the use of data on demand, measured in physical units, to draw conclusions about benefits to users in the project's primary market. This step is based on the implicit assumption that all user benefits are derived from and proportional to the quantity demanded. That is, one assumes a relationship like:

$$B_t = f(Q_t) = w_1 * Q_t \tag{1}$$

where B_t stands for the project's benefits to users in year t of operation; Q_t stands for the quantity of demand in year t; and the parameter w_1 linking users' physical demand to their benefits is constant but unknown.

The main objective of the analyses referred to above is the quantification of forecast errors. To that end, they use the available data on demand to infer the ratio of actual benefits in the first year of operation to their expected value at decision-making:

$$\frac{Q_1}{E[Q_1]} = \frac{w_1 * Q_1}{w_1 * E[Q_1]} = \frac{B_1}{E[B_1]} \tag{2}$$

where $E[\bullet]$ denotes expectation at the time of decision-making. In other words, it is implicitly assumed that the forecast error in terms of user benefits equals the forecast error in terms of physical demand. The latter remains the only quantity measured from data; the implicit assumptions spelled out in (1) and (2) just explain how the data on physical demand is used to draw conclusions about unknown benefits.

From the perspective of social welfare analysis, the key observation from (2) is that as long as the parameter w_1 remains unknown, it is only possible to draw conclusions about forecast errors, and that any conclusions about the value of benefits remain out of reach.

Consider next the extension of this approach to life cycle (LC) user benefits. Based on (1), they can be expressed as:

$$B_{\mathrm{LC}} = \sum_{t=1}^{T} B_t = w_1 * \sum_{t=1}^{T} Q_t \tag{3}$$

Again, the parameter w_1 remains unknown, and the only data that are available concern expected and actual first-year demand. Under these circumstances, the only possible statement concerning LC user benefits is that its forecast error is assumed to equal the forecast error of first-year demand:

$$\frac{B_{\mathrm{LC}}}{E[B_{\mathrm{LC}}]} = \frac{\sum_{t=1}^{T} B_t}{\sum_{t=1}^{T} E[B_t]} = \frac{w_1 * \sum_{t=1}^{T} Q_t}{w_1 * \sum_{t=1}^{T} E[Q_t]} = \frac{Q_1}{E[Q_1]} \Leftarrow Q_t = Q_1 \forall t \tag{4}$$

The assumption that the level of demand in each subsequent year equals demand in year one is a sufficient condition for the equality of forecast errors of LC benefits and first-year demand. Under that condition, (4) holds in present value terms, as the discount factor applied to the numerator and the denominator alike cancels out.

Note, though, that the constant level of demand is not a necessary condition in (4): the forecast error of LC benefits could also equal the forecast error of first-year demand if

https://doi.org/10.1017/bca.2024.29 Published online by Cambridge University Press

demand and benefits grew at a constant rate equal to the discount rate. That is, the forecast errors of LC benefits equal those of first-year demand also when the present value of benefits is constant for all future points in time. In that case, the growth rate and discount rate cancel out in both the numerator and the denominator. However, Flyvbjerg (2005) as well as Flyvbjerg and Bester (2021) ignore that possibility on the grounds that first-year demand and benefit shortfalls are likely to persist.

Wider economic benefits of projects are not considered in the Flybjerg database. Flyvbjerg and Bester (2021) argue that they must be "roughly" proportional to demand, so any overestimation of demand would lead to an overestimation of wider benefits. To explore the implications of this argument, let us postulate the following relationship between wider benefits and demand:

$$WB_t = f(Q_t) = w_2 * Q_t \tag{5}$$

Where WB stands for wider benefits and w_2 is an unknown parameter. Again, the forecast error of wider benefits in year one – or any subsequent year – must be assumed to equal the forecast error of first-year demand:

$$\frac{WB_1}{E[WB_1]} = \frac{w_2 * Q_1}{w_2 * E[Q_1]} = \frac{Q_1}{E[Q_1]} = \frac{B_1}{E[B_1]} \tag{6}$$

The extension of (6) to the LC wider benefits is straightforward in view of (4). All in all, the unknown parameters w_1 and w_2 become irrelevant. That irrelevance, in turn, means as pointed out earlier that any conclusions about the level of benefits – be they user or wider benefits – remain out of reach.

2.2 Decision criteria and scope of possible conclusions

As explained above, the Flyvbjerg database contains data on construction cost estimates and outcomes, measured in financial terms, as well as on first-year demand forecast errors, which are assumed to translate into forecast errors in first-year benefits and, under the assumption that demand stays constant at its level in year one, into forecast errors in LC benefits.

Flyvbjerg (2016) as well as Flyvbjerg and Bester (2021) present the construction cost and user benefit forecast errors for 1,603 projects in eight separate project types. The latter concludes: "… a typical ex-ante benefit–cost ratio produced by conventional methods is overestimated by between approximately 50 and 200%, depending on investment type" (Flyvbjerg & Bester, 2021, p. 402).

That conclusion is based on the forecast error in LC benefits, approximated from the data on the forecast error in first-year demand as shown in (1)–(4) above, together with the forecast error in construction costs (denoted CC):

$$\frac{B_{LC}/CC}{E[B_{LC}]/E[CC]} = \frac{B_{LC}/E[B_{LC}]}{CC/E[CC]} = \frac{Q_1/E[Q_1]}{CC/E[CC]} \tag{7}$$

The left-hand side of (7) measures the forecast error in the benefit–cost ratio (BCR), referred to in the quote from Flyvbjerg and Bester. That same quote refers notably also to "conventional methods" of producing the BCR, which can only mean BCA. In other words, Flyvbjerg and Bester interpret (7) calculated based on data in the Flyvbjerg database as

https://doi.org/10.1017/bca.2024.29 Published online by Cambridge University Press

measuring forecast errors in BCRs, as if the latter were based on ex-ante and ex-post BCRs obtained from ex-ante and ex-post BCA.

The previous section discussed the measurement of project costs and benefits in the Flyvbjerg database and contrasted them with social BCA. Against that background, for the left-hand side of (7) to approximate the forecast error in a BCR, as obtained by a BCA, the following conditions identified in the previous section must hold:

A. Costs of operating and maintaining the project are zero;
B. There are no nonmarket effects associated with the project;
C. Market prices can be used in lieu of shadow prices;
D. All user benefits are derived from and proportional to the volume of demand;
E. The forecast error in user benefits equals the forecast error in physical demand;
F. Demand stays unchanged from its level in the first year of operation until the end of the project's economic life (or the present value of benefits is constant for all future points in time);
G. There are no wider benefits from the project.

In other words, the data in the Flyvbjerg database could be used to draw conclusions about forecast errors in BCRs if the cost of operating and maintaining the project equals its wider benefits in present value terms (conditions A and G above); the investment project is marginal in size and implemented without government intervention in a perfectly competitive market where private and social costs and benefits coincide (conditions B, C, and G); and project benefits to users are derived from and proportional to the volume of demand which is constant (or whose growth translates into constant present value of benefits in any future period; conditions D–G).

Conversely, if any of these conditions do not hold, a wedge is driven between the left-hand side of (7) as estimated from data in the Flyvbjerg database and as estimated by BCA.

It is important to recognize that the discussion in this section concerns the forecast error in the BCR, as shown in (7). As already mentioned in passing when discussing equations (2) and (4), the consideration of ratios when the parameters w_1 and w_2 remain unknown implies that it is not possible to draw any conclusions about the absolute level of project benefits. As long as the level of benefits remains unknown, it cannot be compared to the level of project costs and, consequently, it is not possible to calculate any customary BCA decision criteria for projects (Net Present Value, Economic Internal Rate of Return, or BCR). Consequently, it is also not possible to draw any firm conclusions based on the available data about the social welfare consequences of projects in the Flyvbjerg database.

Flyvbjerg et al. (2002); Flyvbjerg et al. (2003); Flyvbjerg (2009); Flyvbjerg (2016); as well as Flyvbjerg and Bester (2021) link their analyses of samples of data from the Flyvbjerg database to the societal desirability of the projects analyzed. Their conclusions refer to likely resource misallocation; detrimental social welfare consequences of ignoring cost overruns; and risk of nonviability. While none of the articles claims outright to demonstrate that the projects analyzed reduce social welfare, they consider the results of the analysis as being indicative of economic inefficiency.

3. Simulations of welfare consequences

The purpose of this section is to present some simple simulations that adopt some key parameters from the Flyvbjerg database and introduce some necessary assumptions so as

https://doi.org/10.1017/bca.2024.29 Published online by Cambridge University Press

to analyze the relationship between forecast errors of a magnitude observed in the database and possible consequences for social welfare. To be clear, the purpose of the simulations is not, and cannot be, to conduct any fully-fledged BCA of projects or types of projects in the database. Instead, the purpose is merely to explore whether the observed forecast errors can be associated with improvement in social welfare at all under some arguably reasonable assumptions, or whether they do indeed support the hypothesis that projects with forecast errors of the observed magnitudes are likely to reduce social welfare.

The simulations below aim to estimate a limited measure of projects' Net Present Value (NPV) as well as their BCR. There are, however, also other approaches to BCA under incomplete information, aimed at assessing the societal desirability of investment projects based on available data on their main benefits and costs. For example, de Rus (2011) analyses investments in high-speed rail infrastructure by "inverting" the BCA. Instead of estimating the NPV of a typical medium-distance high-speed rail investment project, he considers the minimum threshold for passenger demand in the first year of operation that would yield a positive NPV under certain assumptions. An inverted approach could in principle be applied in the present context as well: data on the level of construction costs and first-year demand, together with some supplementary assumptions, would allow one to analyze for example how the BCR varies with different values of w_1 in equation (1) and, specifically, what its threshold value is for the BCR to exceed unity. While the data on the level of construction costs and first-year demand needed for such an analysis exist in the Flyvbjerg database, they have not been reported in any of the publications. Consequently, both the simulations below and other approaches to the BCA of projects in the database, for example, in the spirit of de Rus (2011), require further assumptions to convert the reported forecast errors to levels of demand and benefits.

The simulations draw on Flyvbjerg (2016) as well as Flyvbjerg and Bester (2021), as they analyze the largest sample of 1,603 projects from the database, representing different sectors and project types. Specifically, the key parameters adopted for the simulations concern the ratio of actual to forecast construction costs and the ratio of actual to forecast benefits (as discussed, proxied by quantity demanded) in the first year of operation, both measured at sample average. Flyvbjerg (2016) as well as Flyvbjerg and Bester (2021) report weighted and unweighted sample averages. The sources do not reveal which variable has been used for the weighting, but the weighted averages are used below, simply as they indicate worse performance than the unweighted averages. In addition to sample averages, the simulations also consider the single worst-performing project type in terms of the forecast error for BCR, which is Bus Rapid Transport (BRT) projects. It is, however, to be noted that BRT is something of an outlier in terms of its small sample size of only 6 projects.

Table 1 shows the construction cost overruns and benefit shortfalls for the weighted sample averages and the worst project type, as reported by Flyvbjerg (2016) as well as Flyvbjerg and Bester (2021). It also details the additional assumptions that are needed to link the observed forecast errors to the level of project benefits during its economic life cycle and, hence, to possible consequences for social welfare.

With regard to the assumptions in Table 1, the first one concerns the derivation of the level of benefits. To that end, an assumption is made about the ratio of the expected first-year benefits to the expected construction costs. Given that the latter are known, this assumption is sufficient to establish the level of expected first-year benefits and, moreover, the level of

https://doi.org/10.1017/bca.2024.29 Published online by Cambridge University Press

Table 1. *Parameters and assumptions employed in the simulations*

	Sample average	Worst project type
Parameters		
Construction cost overruns	1.43	1.41
Benefit shortfalls	0.83	0.42
Assumptions		
The expected ratio of benefits in the first year of operation to construction costs	0.1–0.5	0.1–0.5
Economic life	25 years	25 years
Discount rate	4%	4%
	10%	10%
Demand ramp-up	0%	0%
	2% until $E[B_1]$ reached, then 0%	4% until $E[B_1]$ reached, then 0%

Source: Parameters from Flyvbjerg (2016, Table 1) and Flyvbjerg and Bester (2021, Table 1).

actual first-year benefits, as the shortfall in benefits is known, too.[2] Values in the interval between 10 and 50% are considered for the ratio. These endpoints are, of course, to some extent arbitrary, intended to reflect the significant variability between sectors and individual projects. In reality, major infrastructure projects with long economic lives may well have ratios below 10%, while it is also conceivable that, for example, the rehabilitation of an asset in high demand with relatively low investment cost can have a very short payback period. In any event, the simulations are based on linear relationships between the discounted costs and benefits on the one hand and their NPV on the other hand, so NPV outcomes for values outside the range can be seen by simple extrapolation.

It is furthermore assumed that the economic (operational) life of a typical project is 25 years long. The discount rate is assumed to be either 4 or 10%. These specific rates are chosen based on a recent review of Social Discount Rates by Groom et al. (2022). They review the values of both the Social Rate of Time Preference and Social Opportunity Cost of Capital used by international organizations as well as individual countries in developed and developing countries alike. The lowest rates in developed countries are between 1 and 5% and the highest rates in less developed countries are between 6 and 12%. The 4 and 10% rates employed are toward the upper end of both intervals, thus representing conservative but not extreme values.

[2] To see the sufficiency of that assumption, consider a numerical example based on Table 1. Flyvbjerg et al. report a ratio of actual to forecast construction costs equalling 1.43 at sample average. Say that forecast construction costs were 100 and the actual outcome 143. They also report a ratio of actual to forecast first-year benefits equalling 0.83. As we do not have any information about the level of forecast or actual first-year benefits, we must introduce some assumption to proceed. In principle, any assumption will do that links the level of forecast or actual construction costs with the ratio of actual to forecast first-year benefits. It is sufficient (but not necessary) to make an assumption that links the forecast construction costs to forecast first-year benefits. Say that we assume that link to equal 0.1; in that case, we can calculate the level of forecast first-year benefits (which equals 100 * 0.1 = 10). The level of actual first-year benefits can now be calculated as 10 * 0.83 = 8.3. The additional assumption has thus allowed us to derive both forecast and actual construction costs as well as first year benefits in level terms, allowing us to proceed to a calculation of our limited NPV measure.

https://doi.org/10.1017/bca.2024.29 Published online by Cambridge University Press

Finally, as regards the ramp-up of demand, the assumption of no ramp-up is adopted, alongside some sensitivity analysis assuming some ramp-up. For the sample average, a 2% annual ramp-up in demand is assumed from year two of operation until the forecast level of demand is reached (in year 11), after which demand is assumed unchanged. For the worst-performing project type, a 4% annual ramp-up is assumed until the forecast level of demand is reached (in year 24), after which demand is assumed unchanged. Again, these assumptions are somewhat arbitrary and only intended to illustrate the impact on the NPV of assuming some ramp-up in demand.

It must be emphasized that these assumptions are intended merely to link the observed forecast errors to possible consequences for social welfare. They are far from sufficient to draw any firm conclusions about the underlying projects' impact on welfare. Indeed, the simulations remain subject to the limitations identified in section 2.2 above, so the most that they can achieve is to examine whether we can conclude that biased forecasting of the magnitude observed in the Flyvbjerg database is sufficient evidence of a likely reduction in welfare.

Given these parameters and assumptions, it is possible to simulate a limited measure of the NPV of the sample average project as well as the average project of the worst-performing type. The results for the sample average are shown in Figure 1, and those for the worst-performing project type are shown in Figure 2.

Thus, based on all the parameters, assumptions, and caveats listed above, one cannot exclude the possibility that the average project in the sample analyzed by Flyvbjerg (2016) as well as Flyvbjerg and Bester (2021) could be welfare-improving when the social discount rate is 4% or lower (solid lines in Figure 1) and the ratio of expected first-year benefits to construction costs above 10%. Under these circumstances, there is no need to assume any ramp-up of demand. If the discount rate is 10% (dashed lines), the ratio of expected first-year benefits to construction costs must be above 20% for a positive value of the limited NPV

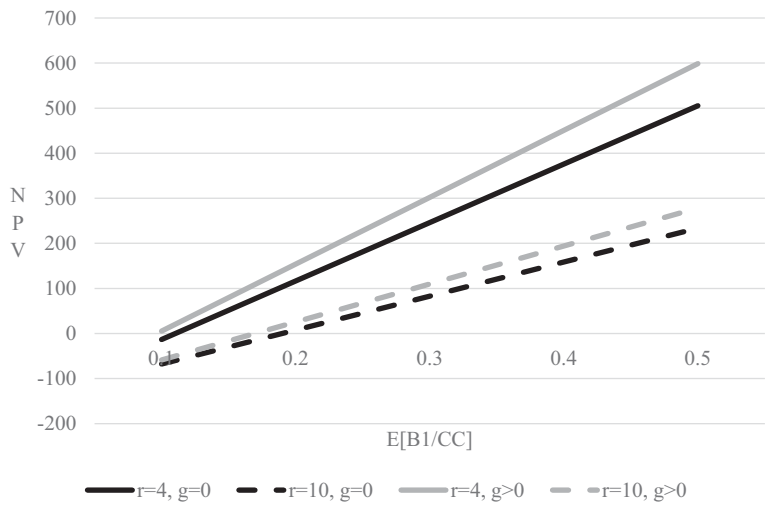

Figure 1. *Simulation results for sample average.*
Note: *NPV denotes Net Present Value; E[B_1/CC] denotes the expected ratio of benefits in the first year of operation to construction costs;* r *denotes the discount rate (in %); and* g *denotes the annual ramp-up in demand (in %).*

https://doi.org/10.1017/bca.2024.29 Published online by Cambridge University Press

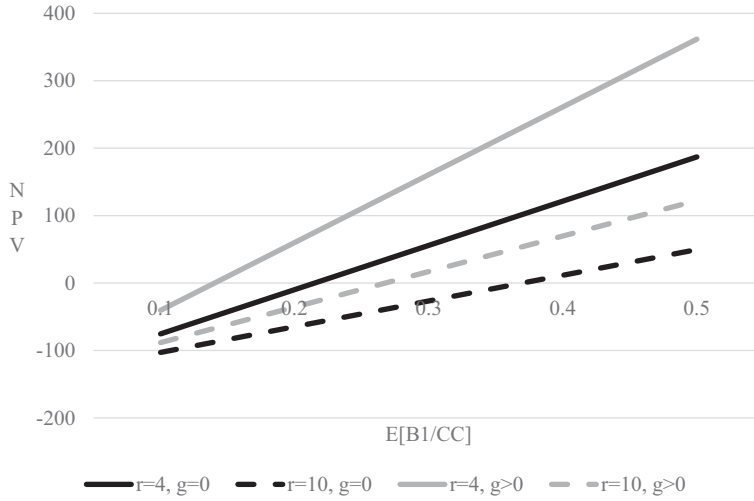

Figure 2. *Simulation results for the worst project type.*
Note: *NPV denotes Net Present Value;* E*[*B *1/CC] denotes the expected ratio of benefits in the first year of operation to construction costs;* r *denotes the discount rate (in %); and* g *denotes the annual ramp-up in demand (in %).*

measure; again, the assumption concerning demand ramp-up does not change the conclusion materially. Conversely, for a project with a ratio of expected first-year benefits to construction costs below 10 or 20%, the NPV may be negative when the social discount rate is above 4 or 10%, respectively.

For the average project in the worst-performing project type in Flyvbjerg (2016) as well as Flyvbjerg and Bester (2021), the threshold value for positive NPV of the ratio of expected first-year benefits to construction costs is above 10% when the social discount rate is 4% or lower and when some demand ramp-up is assumed. The threshold value for positive NPV is above 20% when the social discount rate is 4% or lower and when no demand ramp-up is assumed. When the social discount rate is 10%, the same thresholds are approximately 0.3 (some demand ramp-up) and 0.4 (no demand ramp-up).

In addition to NPV, Tables 2 and 3 also consider BCR as a decision criterion. For comparison with Flyvbjerg (2016) as well as Flyvbjerg and Bester (2021), and in view of the results in the previous section, Table 2 reports BCRs for the first year of operation, while Table 3 reports BCRs for the 25-year life cycle.

Tables 2 and 3 help quantify the difference between the BCR forecast errors and the level of BCRs. The forecast errors in Table 2 are calculated based on the forecast errors for first-year benefits and construction costs as reported in Table 1, discounting the first-year benefits. The forecast errors are large, as are those implied by the data in Flyvbjerg (2016, Table 1) as well as Flyvbjerg and Bester (2021, Table 1). However, the level of the life cycle BCR is in many cases above one, even when no ramp-up of demand is assumed.

To recall, equation (7) in the previous section implies that the BCR forecast errors for year one and for the entire life cycle are identical under the assumption of no ramp-up in demand. Thus, even when no ramp-up is assumed, and the forecast error of a BCR based on data on first-year benefits can be considered as forecast errors of life cycle BCR, one cannot

https://doi.org/10.1017/bca.2024.29 Published online by Cambridge University Press

Table 2. *Benefit–cost ratios for year 1*

	Minimum	Maximum	Forecast error = Actual/forecast
Forecast	0.100	0.500	
Sample average			
Actual based on parameters in Table 1	0.058	0.290	0.58
$r = 4$	0.056	0.279	0.56
$r = 10$	0.053	0.264	0.53
Worst project type			
Actual based on parameters in Table 1	0.030	0.149	0.30
$r = 4$	0.029	0.143	0.29
$r = 10$	0.027	0.135	0.27

Note: r denotes the discount rate (in %).

Table 3. *Life cycle benefit–cost ratios*

	Minimum	Maximum
Sample average		
$r = 4$, $g = 0$	0.907	4.534
$r = 4$, $g = 2$	1.037	5.187
$r = 10$, $g = 0$	0.527	2.634
$r = 10$, $g = 2$	0.590	2.948
Worst project type		
$r = 4$, $g = 0$	0.465	2.327
$r = 4$, $g = 4$	0.713	3.565
$r = 10$, $g = 0$	0.270	1.352
$r = 10$, $g = 4$	0.374	1.868

Note: r denotes the discount rate (in %), g denotes the annual ramp-up in demand (in %).

conclude that the large forecast errors are necessarily associated with welfare reduction by the typical project.

Furthermore, Monte Carlo simulations are performed considering the expected ratio of first-year benefits to construction costs as well as the discount rate as random variables.[3] Both are assumed to be uniformly distributed, with the expected ratio of first-year benefits to construction costs varying in the interval [0.1, 0.5] and the discount rate in [4%, 10%]. From each distribution, 1,000 draws are made, based on which the limited NPV measure is calculated. Figure 3 below shows the distribution of the 1,000 NPV values for the sample average (see Table 1), and Figure 4 shows the distribution of NPV for the worst project type.

For the sample average, 12.5% of the 1,000 simulated NPV values are below zero, while 87.5% are positive. As a plausibility check for this result, consider Figure 1, where the area representing possible NPV values from the Monte Carlo simulation is spun by the two black

[3] The Monte Carlo simulations were performed with the help of the Data Analyst of ChatGPT.

https://doi.org/10.1017/bca.2024.29 Published online by Cambridge University Press

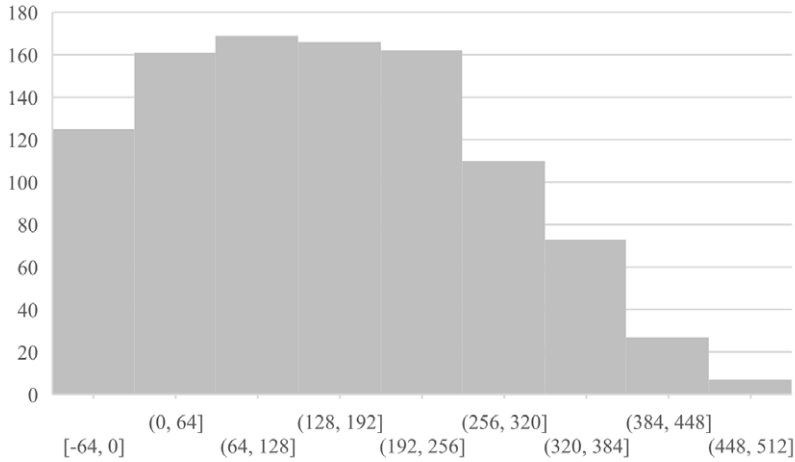

Figure 3. *Frequency distribution of NPV from Monte Carlo simulations for sample average.*

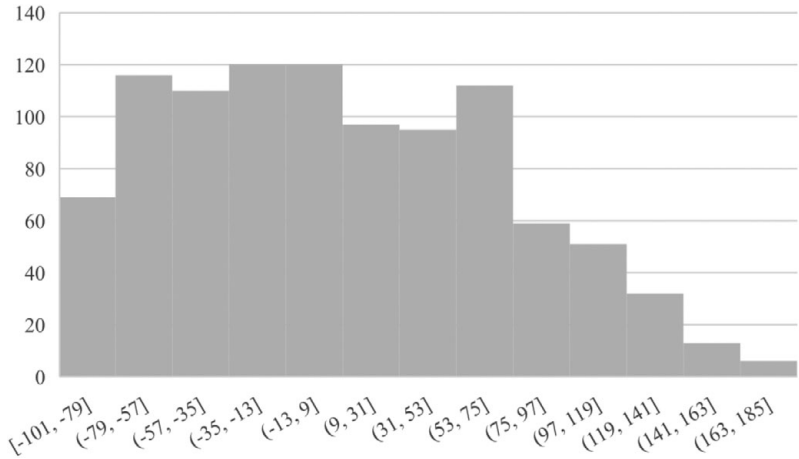

Figure 4. *Frequency distribution of NPV from Monte Carlo simulation for worst project type.*

lines as well as vertical lines (not shown in Figure 1) at 0.1 and at 0.5. The share of that area underneath the horizontal axis represents the expected share of negative NPV values from a Monte Carlo simulation. Thus, based on Figure 1, the result of the Monte Carlo simulation shown in Figure 3 appears quite plausible.

For the worst-performing project type, 49% of the 1,000 simulated NPV values are below zero and 51% are positive. Using Figure 2 as a plausibility check for this result suggests that, if anything, the share of negative NPV values from the Monte Carlo simulation may be on the high side. In other words, simulations with significantly higher numbers of draws from the random distributions might result in a slightly lower share of negative NPV values.

https://doi.org/10.1017/bca.2024.29 Published online by Cambridge University Press

All in all, these Monte Carlo simulations confirm the main insights from Figures 1 and 2 as well as from Table 3. First, using the parameters from Flyvbjerg (2016, Table 1) as well as Flyvbjerg and Bester (2021, Table 1) and imposing some minimal assumptions, it is possible to show that the significant forecast errors observed in the Flyvbjerg database can be compatible with the enhancement of social welfare. To be clear, those forecast errors can also be indicative of welfare reduction, but the latter should not be considered as a foregone conclusion.

Second, whether the average project or the average project of the worst-performing type is likely to enhance or reduce welfare depends, quite intuitively, on its characteristics, its use during its life cycle, and its economic environment. In the simulations above, these effects were reduced to considering the magnitude of expected early-stage annual benefits relative to costs; possible ramp-up of demand during operation; and the social discount rate, respectively. In a BCA, many more types of costs and benefits would be included, valued at shadow prices as necessary. They might well, of course, lead to welfare conclusions that are very different from those suggested by the simulations.

4. Conclusions

The Flyvbjerg database is the most comprehensive, albeit not openly accessible, source of data on forecast errors of major investment projects. Analyses drawing on those data have demonstrated, among other things, the systematic presence of large forecast errors in both construction costs and user demand in the first year of operation. They have also linked those results to the social welfare consequences of the underlying projects, suggesting that the large forecast errors are indicative of welfare reduction by typical projects in the database.

The analysis in this paper addresses the links between the Flyvbjerg database and BCA. Specifically, it explores how project costs and benefits are measured in the database and contrasts that measurement to BCA. Against that background, the forecast errors in BCRs based on the database are compared with forecast errors based on BCA. That comparison suggests that there are only very limited circumstances under which the reported forecast errors would coincide with forecast errors obtained through BCA. In brief, that is only the case for marginal projects with life cycle demand constant at its level in year one (or the present value of benefits constant for all future periods) and with wider benefits equaling operation and maintenance costs in present value terms and, moreover, in competitive and undistorted markets. In other cases, such as nonmarginal projects, nonconstant demand, and distorted markets, the forecast errors in projects in the database are not an accurate approximation of BCA forecast errors.

As the hard data in the database only allow the calculation of forecast errors and not the level of benefits, it is also not possible to use those data to draw any firm conclusions about the social welfare consequences of the projects. However, some of the articles based on the database do indeed make statements about likely negative welfare consequences of typical projects, given the rather large magnitude of the observed forecast errors.

The simulations performed suggest that, given the observed forecast errors, the reduction of social welfare is not a foregone conclusion. The minimal assumptions employed in the simulations were only meant to provide the simplest possible link from the observed forecast errors to social welfare consequences, with a view to assessing whether the forecast errors do

https://doi.org/10.1017/bca.2024.29 Published online by Cambridge University Press

indeed seem to imply likely welfare destruction. The results suggest that even the large forecast errors observed on the basis of the Flyvbjerg database can under some arguably reasonable assumptions be compatible with either welfare improvement or reduction.

An important assumption underlying the simulations concerns the ratio of expected benefits to construction costs. A wide range of this ratio was considered, and values at its lower end were in many cases found to result in the limited NPV measure turning negative. In other words, projects that are expected to be expensive with little use are susceptible to destroying social welfare. This is a trivial conclusion as such, but it is significant in highlighting the fact that the welfare conclusions by Flyvbjerg and his co-authors quoted in the introduction imply that most projects, or typical projects, are both expensive and little used. However, if one only has hard data on forecast errors in costs and demand, one can conclude at most that projects are more expensive and less used than expected. Whether that conclusion translates into a conclusion about negative social welfare consequences depends, among many other things, on the actual levels of construction costs and early-stage benefits.

It is impossible to say how the results of a fully-fledged BCA might differ from those obtained through the simple simulations. But the simulations serve to demonstrate that the data underlying the estimated forecast errors in Flyvbjerg (2016) as well as Flyvbjerg and Bester (2021) – even for the average project of the worst-performing type – can be compatible with both positive and negative welfare consequences, and that systematic ex-post BCAs of the projects would be necessary to draw any conclusions in that respect. Ex-post BCAs of large samples of major projects are rare, as reviewed recently by Wang and Levinson (2023), but the few available examples, while not representative, caution against concluding that major projects typically destroy social welfare.

Apart from reporting forecast errors based on the database, Flyvbjerg and Bester (2021) also draw some conclusions about BCA as a method for assessing the societal desirability of investment projects or of public policy interventions more broadly. The forecast errors are seen as evidence of BCA to be "broken", and some suggestions are made as to how BCA should be reformed to serve more effective resource allocation.

The results presented in this paper caution against too far-reaching conclusions as far as BCA as a method is concerned. To repeat, the observed forecasting errors are not comparable to forecasting errors that would be obtained using proper BCA. Cost estimates and demand forecasts are prepared for all projects, for design and budgeting purposes, regardless of whether a BCA is performed or not, and the observed forecasting errors should not be considered as an indication that the projects in the database destroy social welfare. So, in sum, the forecasting errors do not in and of themselves suffice as evidence that BCA is inaccurate, nor that BCA helps select poor projects.

That said, BCA, and project planning more broadly, is without doubt subject to the kinds of errors, cognitive biases, and strategic misrepresentation that Flyvbjerg and co-authors have so extensively analyzed. It is just that the forecasting errors reported based on the database are neither a useful nor an accurate measure of those failings; therefore, those forecasting errors should not be used as a basis for drawing conclusions about the social welfare consequences of projects in the database, nor about BCA as a method of assessing such consequences.

Acknowledgments. I thank two anonymous referees as well as Dr José Doramas Jorge Calderón and Dr Ofelia Betancor-Cruz for their detailed comments. All remaining errors are mine.

https://doi.org/10.1017/bca.2024.29 Published online by Cambridge University Press

Competing interest. Timo Välilä is employed at the European Investment Bank (EIB). The opinions expressed are personal and may not necessarily reflect the EIB group position.

References

Ansar, Atif, Bent Flyvbjerg, Alexander Budzier, and Daniel Lunn. 2016. "Does infrastructure investment lead to economic growth or economic fragility? Evidence from China." *Oxford Review of Economic Policy*, *32*(3): 360–390. https://doi.org/10.1093/oxrep/grw022.

De Rus, Ginés. 2011. "The BCA of HSR: Should the government invest in high-speed rail infrastructure?" *Journal of Benefit-Cost Analysis*, *2*(1): Article 2. https://doi.org/10.2202/2152-2812.1058.

Flyvbjerg, Bent. 2005. "Measuring inaccuracy in travel demand forecasting: Methodological considerations regarding ramp up and sampling." *Transportation Research Part A*, *39*: 522–530. https://doi.org/10.1016/j.tra.2005.02.003.

Flyvbjerg, Bent. 2009. "Survival of the unfittest: why the worst infrastructure gets built—and what we can do about it." *Oxford Review of Economic Policy*, *25*(3): 344–367. https://doi.org/10.1093/oxrep/grp024.

Flyvbjerg, Bent. 2016. "The fallacy of beneficial ignorance: A test of Hirschman's hiding hand." *World Development*, *84*: 176–189. https://doi.org/10.1016/j.worlddev.2016.03.012.

Flyvbjerg, Bent, and Dirk W. Bester. 2021. "The cost-benefit fallacy: Why cost-benefit analysis is broken and how to fix it." *Journal of Benefit-Cost Analysis*, *12*(3): 395–419. https://doi.org/10.1017/bca.2021.9.

Flyvbjerg, Bent, and Dan Gardner. 2023. *How Big Things Get Done*. New York: Currency.

Flyvbjerg, Bent, Mette Skamris Holm, and Soren Buhl. 2002. "Underestimating costs in public works projects: Error or lie?" *Journal of the American Planning Association*, *68*(3): 279–295. https://doi.org/10.1080/01944360208976273.

Flyvbjerg, Bent, Mette K. Skamris Holm, and Søren L. Buhl. 2003. "How common and how large are cost overruns in transport infrastructure projects?" *Transport Reviews*, *23*(1): 71–88. https://doi.org/10.1080/01441640309904.

Groom, Ben, Moritz A. Drupp, Mark C. Freeman, and Frikk Nesje. 2022. "The future, now: A review of social discounting." *Annual Review of Resource Economics*, *14*: 467–491. https://doi.org/10.1146/annurev-resource-111920-020721.

Næss, Petter, Bent Flyvbjerg, and Søren Buhl. 2006. "Do road planners produce more 'honest numbers' than rail planners? An analysis of accuracy in road-traffic forecasts in cities versus peripheral regions." *Transport Reviews*, *26*(5): 537–555. https://doi.org/10.1080/01441640500532005.

Wang, Yadi, and David Levinson. 2023. "The accuracy of benefit-cost analysis for transport projects supported by the Asian Development Bank." *Asian Transport Studies*, *9*: 100104. https://doi.org/10.1016/j.eastsj.2023.100104.

Cite this article: Välilä, Timo. 2024. "Forecast Errors and Welfare Conclusions Based on the Flyvbjerg Database." *Journal of Benefit-Cost Analysis*, doi:10.1017/bca.2024.29