UCL

UCL
SCHOOL OF
MANAGEMENT

Date	25 January 2021
From	Bert De Reyck, UCL Yael Grushka-Cockayne, University of Virginia Ioannis Fragkos, Erasmus University Rotterdam Xiaojia Guo, University of Maryland
To	Department for Transport
Topic	Peer Review Optimism Bias Research – Final Report

1. Executive Summary

1.1. This report contains a peer review of the research performed by the Oxford Global Projects team (OGP) for the Department for Transport (DfT), exploring the rates of optimism bias (OB) for use as uplifts in the economic appraisal of UK infrastructure projects. These represent the typical rates of cost overruns, time overruns and benefit shortfalls in projects, and are estimated by comparing outturn data with estimates produced at different business case approval points. The report also contains an overview of recent developments in reference class forecasting (RCF), and a method for accounting for inflation in economic appraisals.

1.2. The research conducted by the OGP team expands on previous work carried out for DfT in 2004, adding data from more than 6,500 projects, both from the UK and internationally, from a range of project types (including road, rail, bridges, tunnels and buildings). The report estimates OB rates at each main business case stage (SOBC, OBC and FBC)[1], and now also provides OB estimates for cost, time, and benefits.

1.3. The aim of this review is to provide assurance to DfT that the research performed by the OGP team is methodologically sound and the findings robust and fit for purpose of incorporating in DfT's transport appraisal guidance. The review focuses on issues surrounding the data, the analysis, the definitions and interpretation of core concepts, the treatment of inflation, the robustness of the derived forecasts across the business case phases, and how to implement the recommendations contained in the report.

1.4. On the whole, we found that the analysis performed by OGP team is rigorous, its findings well supported by the data, and its recommendations plausible. **Our review has, however, revealed some potential areas for improvement in the analysis, and highlights some issues with respect to the proposed optimism bias adjustments and their implementation, as well as the proposed treatment of inflation. Additionally, we feel that the recommendations of the report may not be suitable for smaller projects, as the analysis was performed on a dataset of mainly larger and complex projects, and it would be helpful if the authors would**

[1] Strategic Outline Business Case, Outline Business Case, Full Business Case.

provide guidance on the nature, size and duration of projects the recommendations apply to.

1.5. Note that given the confidentiality of the data, we were unable to replicate the analysis described in the report ourselves or offer additional insights from other analyses.

2. Review of the methodology

2.1. The OGP team ("the authors") uses reference class forecasting (RCF) for predicting the systematic underestimation of cost and schedule overruns and benefit shortfalls in projects. The underlying cause of these overruns can be due to optimism bias (OB) or strategic misrepresentation (SM). OB can be thought of as an unconscious bias, whereas SM is a conscious one.

2.2. The authors review recent development in reference class forecasting, and conclude that RCF is still the most appropriate methodology for analysing and predicting systematic cost and schedule overruns. Several examples of research are referenced, indicting the superior performance of RCF when compared to traditional forecasting methods including quantitative risk assessment through Monte Carlo simulation (QRA) and earned value management (EVM).

2.3. We agree that an outside-view approach (like RCF) is still the most suitable approach for predicting typical cost and schedule overruns, especially in early stages of projects where little detail is available on the risk of individual project components, work packages, and activities. We do feel, however, that the statement "a reference class comprising 20-30 past, similar projects is robust to derive meaningful insights" could be better supported by evidence from a statistical analysis of RCF predictions and actual cost outturns since the first study on OB in 2004.

2.4. For determining whether different types of projects warrant separate reference classes, The OGP team analyses the differences one at a time. An alternative approach would be using a multi-factor model, including all predictors in one model, potentially along with their interactions. This would increase the robustness of the results, and a validation of the categories proposed by the authors (see also 4.7).

2.5. When comparing different project types, the OGP team tests whether the medians of the respective RCF distributions are different, using a Wilcoxon rank-sum test. We recommend, however, that this analysis is performed on the mean, which is the estimate that should be used for calculating OB uplifts for the purpose of economic analysis (see also 6.1). In fact, since the authors also recommend using the 80th percentile of the RCF distribution (RCF80) as an evaluation criterion, and due to the skewness in the distributions, the analysis could be strengthened by testing whether the respective distributions are different, rather than just the median. For testing whether two means are different, t-tests could be used, or alternative tests that do not rely on the normality assumption. For testing whether two distributions are different, the Kolmogorov-Smirnov test could be used.

2.6. The OGP report provides a snapshot analysis of cost, schedule and benefit shortfalls. We understand, however, that the authors also carried out a longitudinal study, so it would be good if they could briefly comment on any findings regarding a possible trend in shortfalls, or the lack of such a trend.

2.7. The authors also mention risk-based estimation (RBE), a bottom-up approach in contrast to the top-down RCF approach, and write that best results in forecasting accuracy will be achieved by combining top-down and bottom-up approaches. We agree with this statement, but do not necessarily agree with the statement that "the top-down RCF approach and the bottom-up QRA approach should reach similar conclusions". In particular,

there have been recent developments in using machine learning to predict cost and time overruns at a granular level in projects (at the activity or work package level), which we believe could outperform top-down RCF. However, this approach is only feasible when detailed plans are available, which still makes RCF the primary choice for early-stage economic analyses.

2.8. The OGP team removed all inflation from the data using country-specific World Bank GDP implicit deflators. This deviates from the HMT guidance to use the ONS deflator, which was motivated to ensure comparability across international figures. We believe it would be helpful if the results using the ONS deflator were also presented. This would also be useful as a sensitivity analysis, to check whether and how much the results depend on the adoption of a specific inflation measure.

2.9. The OGP team also presents an RCF approach for forecasting inflation. For this purpose, the inflation of 116 projects in the UK dataset is computed (where both nominal and real-term cost overruns were available), resulting in an inflation RCF distribution. The results show a median inflation figure of 4%. We believe that this approach, i.e. using evidence-based inflation estimates, is preferable compared to using ONS or World Bank GDP deflators. We do believe, however, that the mean, which was not reported, should be used instead for economic analysis (see 6.1).

2.10. The authors also propose two different approaches for dealing with inflation, depending on whether the project is fully exposed to inflation, or protected from inflation through contractual clauses. When a project is protected from inflation, the authors propose using an inflation estimate based on real-term versus nominal forecasts, rather than on the GDP deflator. We agree that this could be a good way forward.

3. Review of the data

3.1. The analysis is based on a total of 7,043 projects, including 355 rail projects, 977 road projects, 117 fixed links projects, 149 building projects, 5,303 IT projects, 48[2] land & property transactions, 20 rolling stock purchases, and 74 projects monitored for operational expenditure. 423 (6%) of these projects are from the UK.

3.2. We recommend including descriptive statistics on the projects in the database and in the reference classes (on project characteristics such as size, duration, budget, complexity, etc.), as it seems that the majority are large projects. For instance, the authors define rail new builds as projects with a base cost estimate exceeding £7M, which corresponds to the 90th percentile in the Network Rail dataset, i.e. with 90% of Network Rail project budgets smaller than £7M. This indicates that the recommendations of the report are not suitable for the majority of improvement and enhancement projects. The building projects in the dataset are also more complex than typical residential buildings, so the recommendations should perhaps only be applied for complex projects. It would, therefore, be helpful if the authors would provide guidance on the nature, size and duration of projects the recommendations apply to.

3.3. The quality assurance (QA) procedures followed by the OGP seem reasonable and provide sufficient confidence in the accuracy of the data capture and cleansing.

3.4. Cost estimates are adjusted using the World Bank GDP deflator instead of the ONS GDP deflator recommended by HMT, which we believe to be a correct choice given the international nature of the dataset assembled by the OGP team.

[2] There are some inconsistencies in the reported numbers, with 48 land & property projects reported on p.8, but 88 on p.11.

4. Review of the analysis

4.1. Cost overruns, schedule overruns, opex overruns and benefit shortfall seem to be correctly calculated. However, on page 13 of the OGP report, the formula used for calculating benefit shortfalls in cases where the expected benefits were negative, is incorrect. Instead of *(actual benefits - estimated benefits)/estimated benefits -1*, this should be *(estimated benefits – actual benefits)/estimated benefits*, or *1 – actual benefits/estimated benefits*. Alternatively, one could use *(actual benefits – estimated benefits)/abs(estimated benefits), regardless of the direction of the expected benefits*. As we do not have access to the underlying data, we cannot verify whether the calculations that form part of the analysis are correct but assume that the error lies in the description of the formula in the report rather than in the actual analysis. This should, however, be verified.

4.2. The OGP team analyse cost, schedule and benefits shortfalls in both UK and international projects (dominated by US and European projects), across the different stage gates of project appraisal (SOBC, OBC, FBC), and across different project types (rail, roads, fixed links, buildings, IT, land & property, and rolling stock). A more granular analysis reveals that the project type (bridge vs. tunnel), geographical region (continent, or UK vs. world), and project size (measured by either duration or budget) do not have a significant impact on the observed shortfalls. This result is similar to the effect observed by De Reyck et al. (2015) for conventional UK rail improvement and enhancement projects, where size (measured by budget), geographical region, and asset categories were not seen to have a significant impact on required OB adjustments.

4.3. For the comparative analysis of overruns between project types, the authors advocate the use of the Wilcoxon rank-sum test (to compare the medians of the distributions) instead of t-tests, as the data does not follow a normal distribution. One should note, however, that the Wilcoxon rank-sum test does assume that the variance (spread) in the compared distributions are equal. In fact, the Wilcoxon test is designed to detect location shifts between two (independent) samples; i.e. it does not estimate the difference in medians but rather the median of the difference. If the two examined distributions are the same and vary only on shift, then this is the same as the difference of medians, but in general this is not the case. In the OGP database, the distributions do not seem the same, and the variances not equal.

4.4. We also have a concern that in this analysis, the OGP team focussed on the median of the RCF distributions, rather than the mean, which is the estimate that should be used for calculating uplifts for the purpose of economic analyses (see 6.1). Therefore, we recommend that this analysis is redone based on the RCFmean. In fact, since the authors also recommend using RCF80 as an evaluation criterion, and due to the skewness in the distributions, we also recommend to test whether each two distributions are significantly different (e.g. using a K-S test). We observed, for instance in Figure 1, that the distributions of benefit deviation (the third figure) are very different for bridges and tunnels, and although their medians look the same, the distributions (and therefore percentiles) don't seem to be.

4.5. Additionally, we recommend that the observed differences in medians (RCF50), means (RCFmean) and RCF80s are also reported (and not only in the box plots), as the fact that the differences in the medians were not found to be statistically significant at a 5% level could be due to the experiment being underpowered due to small sample sizes in some categories; therefore, large differences in observed differences in medians (although designated as not significantly different), means and P80s would still be interesting to highlight for further study.

4.6. In Section 3.2.2 of the report ("UK vs World"), the OGP team addresses the similarity between international and UK-based projects in terms of cost, benefits and schedule, for road, rail and fixed link projects. It may be useful to extend this analysis to IT projects, which appear to have a large UK-based sample. In addition, it might be useful to rerun the UK vs World analysis when all categories are pooled, which will yield a more reliable analysis of any effect. In the spirit of this suggestion, the team may consider, as a robustness check, to report the corresponding overruns of Tables 5, 6 and 7 when the roads and IT references classes are UK-based instead of International, as in these two categories, there are relatively large samples of UK projects.

4.7. It might be useful to examine the impact of certain attributes (such as location) when other factors are taken into an account, in a multivariate model. Adding controls such as time and location fixed effects could capture the impact of time-varying effects (such as technology) or regional effects on the overruns, and therefore lead to a better understanding of the underlying relationships. Alternatively, machine learning methodologies such as clustering and principal component analysis could be used to validate existing findings, or possibly underpin hidden patterns in the data.

4.8. We would also recommend checking the robustness and generalisability of the obtained findings. In particular, it might be useful to conduct an out-of-sample analysis to demonstrate the predictive accuracy of the model, using separate training and test sets. In addition, it would also be useful to conduct a bootstrapping analysis, in which one drops a random number of data points from every reference class, recalculates the corresponding overrun, and repeats for a number of times. The variability of the obtained overruns is a good indicator for the robustness of the corresponding recommendations.

5. Review of the results

5.1. Mean non-opex cost overruns range from -4% (cost savings) in land and property transactions to 44% in building projects. Opex overruns range from 1% in rail projects to 70% in road projects. Schedule overruns range from 4% in rolling stock projects to 32% in building projects. Benefit shortfalls range from 21% (i.e. overachieved benefits) for IT projects to minus 25% in rail projects. The mean cost and schedule overrun are relatively stable across the different project types (except for the cost of land and property transactions and the time for rolling stock purchases). Benefit deviations are more volatile across the different categories, with IT projects clearly anomalous, with benefits being underpromised.

5.2. Looking at UK projects only, mean cost overruns range from 9% in IT projects to 60% in building projects (but with a small sample size). Schedule overruns range from -2% in road projects (ahead of schedule) to 78% in IT projects. Benefit shortfalls range from -1% for road projects to -29% in IT projects. In terms of costs, rail and fixed links projects are the worst offenders, with IT projects showing vastly higher delays than any other project type, along with the highest benefit shortfalls (in contrast to international projects).

5.3. The OGP team also analysed the Network Rail dataset of 5,294 projects studied by De Reyck et al. (2015), and found identical results. They advise to continue using the OB adjustments of De Reyck et al. (2015) for conventional rail enhancements and replacements projects. The findings of De Reyck et al. (2015) also seem in line with the findings of the OGP team for UK rail projects. In particular, De Reyck et al. (2015) report mean OB values of 64% in GRIP stage 1, whereas the OGP team reports a mean cost overrun of 56% for the equivalent SOBC stage. Other estimates are difficult to compare as the stage gates do not perfectly align, and because De Reyck et al. (2015) recommend OB uplifts based on the QRA mean rather than the base estimates for GRIP stages 3, 4 and 5 (see also 6.2).

5.4. We have also compared the results obtained by the OGP team with the current transport appraisal guidance used by DfT. For road projects, the current guidance recommends OB cost uplifts of 44% (SOBC), 15% (OBC), and 3% (FBC) for road projects, whereas the mean cost overruns reported by the OGP team are 48%, 25% and 22%. **The results are very similar for the SOBC stage, with, however, current DfT recommendations significantly lower for OBC and especially FBC. This could be explained by the fact that the DfT recommends applying OB uplifts based on the QRA mean costs (where available), rather than the base costs, whereas the OGP team always uses the base costs. As the QRA mean is typically (substantially) higher than the base cost, this may explain the differences in the results for OBC and FBC, where QRA analyses are often available. The recommended uplift for SOBC, where QRA analyses are less likely to available, are more in line.**

5.5. We see a similar pattern for (conventional) rail projects, where the current guidance recommends OB cost uplifts of 66% (SOBC), 18% (OBC), and 4% (FBC), and the mean cost overruns reported by the OGP team are 56%, 33% and 30%. For fixed links, DfT guidance recommends 66%, 23%, and 6%, with the OGP team reporting 55%, 32% and 28%; again a similar pattern. For building projects, DfT recommends 51% and 6% for SOBC and FBC, respectively, with the OGP team reporting 70% and 44%; although deviations are larger here, the pattern is again similar. **IT projects, however, show very different results, especially at the SOBC stage, with DfT recommending 200% and the OGP team only 69%; we feel that the 200% finding of Mott McDonald may not be accurate anymore (or perhaps never was).**

5.6. A more granular analysis of cost, schedule and benefit shortfalls at different stage gates (SOBC, OBC, FBC) reveals that median shortfalls remain constant between the different stages. This result was unexpected, and is explained by the OGP team as being a result of (1) risk being resolved in the construction phase of a project, i.e. after FBC, and (2) early estimates being sticky due to anchoring and lock-in. We have no reason to suspect that this reasoning is incorrect. However, to validate these findings, we compared the results with the analysis of de Reyck et al. (2015) for conventional UK rail projects. The OGP team reports a median cost shortfall of 19% across the three different stage gates. De Reyck et al. (2015) do not report median OB values, and instead report mean OB values (see 6.1), but we have been able to calculate the median OB values using the same data. We found that the median OB values between GRIP stages 1 and 2 were significantly different at 40% vs. 60%, but no significant differences were observed between the median OB values for GRIP stages 3, 4 and 5 (0%, 0%, -1%). Note that the OB estimates between GRIP 1/2 and GRIP 3/4/5 cannot be directly compared as for GRIP 1/2 they are based on the base cost estimate, whereas for GRIP 3/4/5 they are based on the QRA mean. Therefore, the only observed difference was seen between GRIP stages 1 and 2, with a possible, but unobservable difference between GRIP stages 2 and 3. **Based on these results, we would expect to see a significant difference between the median OB values between GRIP stages 1 and 3, or between SOBC and OBC, which contradicts the results of the OGP study. This may require some further investigation.**

5.7. Note, however, that whether or not the median shortfalls are constant throughout the various stage gates or not, is not very relevant, as the mean shortfall should be used for economic analysis rather than the median (see 6.1). The mean shortfalls in the report do show an improving pattern as a project navigates through the different stage gates. This is a result of a decrease in OB in the tails of the RCF distributions when moving to the different stage gates, i.e. the worst-performing projects are not as extreme in later stages as compared to earlier stages. We believe that this is still consistent with the anchoring and lock-in theory, as it is reasonable to assume that when extreme deviations are expected, even in early stages this will probably break through the anchoring and lock-in effects, and be highlighted early, with smaller deviations being hidden until the construction phase starts. As a result, although the RCFmedian estimates are similar across

the different stages, the RCFmeans are not, and are higher at earlier stages. As the mean rather than the median should be used for calculating OB adjustments for the purpose of economic analyses (see 6.2), the OB values will be different across the different stages, with higher values in the early stages, as expected, and in line with the DfT guidance, which states that "the allowance for optimism bias should be largest at the initial stage of the life of a [..] project, e.g. Strategic Outline Business Case, to decrease in a more detailed business case, e.g. Outline Business Case, and smallest in the presence of a fully detailed business case, e.g. Full Business Case".

6. Implementation

6.1. In various places throughout the report, the OGP group focuses on the RCF50 when calculating required uplifts (contingencies). For instance, in Section 3.3, the authors "suggest picking RCF50 as the low contingency bound", and on page 60, they present an example of a rail project (at FBC stage) that attracts an uplift at P50 of 19% and 60% at P80. The mean, however, is 30%. For the individual economic assessment of a project, it is indeed useful to know a level of cost that the project is unlikely to exceed, e.g. P80. However, when setting contingencies, it is less useful to know the cost level that has a 50% chance of being exceeded. Rather, the mean is more informative, and tells us that if we include a contingency at that level, although individual projects may still exceed their contingency, a (large enough) portfolio of projects should stay within the overall envelope of cost. Therefore, we recommend using RCF mean, not RCF50, when setting contingencies. Note that the authors do refer to using (trimmed) means for economic appraisals in Section 3.3, but recommend using medians when setting contingencies.

6.2. Similarly, the authors focus on RCF50 when presenting their methodology for dealing with inflation; they do not report the mean inflation figures in the UK dataset. We recommend that the means are also reported, and that these should be used for contingency setting and economic analyses.

6.3. When using means, the authors recommend using trimmed means (e.g. 5%-95%). It would be helpful to report which projects are being trimmed in this way, and what the impact on the overrun estimates are. Historical projects with extreme results are an important part of the (tail) risks that projects entail, so trimming extreme outcomes could affect the reliability of overrun forecasts, by ignoring the possibility of extreme future outcomes. Alternatively, if trimming is deemed necessary, one could trim on extreme observed deviations, rather than on fixed percentiles, as extreme values could point to erroneous inputs, which need to be removed.

6.4. The authors recommend applying uplifts to base estimates rather than estimates that include contingencies, including the mean from a quantitative risk assessment (QRA mean, the average result from a Monte Carlo simulation). However, when QRA has been done, the QRA mean is usually a more accurate estimate of the outturn cost than the base cost estimate. Therefore, a case could be made for applying OB uplifts to the base QRA mean, when available. In principle, both approaches can work well, as long as they are performed consistently, and with the required uplifts correctly estimated based on deviations of outturn costs with either the base estimate or the QRA mean.

6.5. We have observed from Figure 5 in the report that rail schedule overruns and benefit shortfalls look quite different between the UK and the rest of the world, with UK projects performing remarkably better. Although the difference could be due to the small UK sample size (possibly making the difference not statistically significant), we recommend considering that when the sample size is sufficient, UK projects rather than worldwide projects should be used as a reference class. This advice is not driven by the fact that UK projects necessarily perform better or worse than other projects worldwide, but practices

in terms of definition of budgets at the different stage gates may be different, whereas the UK has a relatively standardized approach with well-defined approval stages.

6.6. A final observation relates to scope changes. The authors point out that cost overruns are often the result of scope change. We agree, and in fact, significant scope changes are often cited as the main cause of substantial cost overruns. We would, however, like to add a note of caution that scope change, whether small or substantial, should not be used as an excuse for cost or time overruns. Unless projects go back to the SOBC stage for approval of the expanded scope[3], potential overruns due to possible future scope changes should still be considered when setting budgets, so that everyone is aware that original budget targets and timelines might be missed because of intentional changes to the project scope in later stages. Normally this should also result in increased expected benefits, so it should not necessarily be considered as a negative, but project owners and decision makers should be aware of this possibility in advance, by setting the appropriate uplifts.

7. Detailed Comments

7.1. The report would benefit from grammar checking. In some instances, acronyms are presented before being defined, and not always used in a consistent fashion. We have attached a marked-up pdf of the report to help the authors with this.

7.2. Page 5, "The use of these methods … delays"; repetition

7.3. Page 6, RBE is mentioned but not defined until p.40

7.4. Page 8, Table 2: Add "Trimmed" to mean column. Define "Frequency" column more clearly. Consider adding units to each category of overruns.

7.5. Page 17 through 19, Figures 1, 2, 4, 5 and 6: unclear what the scale is of the values on the y-axes. In Figure 3 on page 18, both scales are not clear.

7.6. Page 21, it says "Table 5 below displays an overview of the 18 total international capital cost reference classes", but there are 21 reference classes. Similarly, the next paragraph says there are 15 schedule reference classes in table 6 but actually there are 18.

7.7. Reference to report of De Reyck et al. (2015) is missing from the reference list at the end of the document (although it is referenced in a footnote on p.11).

7.8. "SOBC" rather than "SOC" in Appendix A.

7.9. It would be useful to investigate how trimming the tails of the distribution influences the reported results. Specifically, a graph showing how the reported shortfalls change when varying the level of trimming will add useful insight on the robustness of the results.

7.10. Some numerical values in Table 2 may need some clarifications, in particular how the "Frequency" column relates to the median values. For instance, 2 out of 10 Rolling Stock projects have a schedule overrun, but the Median and RCF 80 values are both 0%.

7.11. In the provided Excel file, cost and schedule overruns are reported for all project categories, excluding Land and Property projects, where only costs are reported. It would be useful to report reference classes for obtained benefits as well. For rolling stock schedules, the OBC overruns seem not to be sorted. For OBC Road overruns the 5% and

[3] When possible, we recommend that significant scope changes are brought back to SOBC so that, if approved, new base estimates can be set, to which the normal OB uplifts can again be applied. Doing this consistently would avoid the OB uplifts having to incorporate the possibility of major scope changes, resulting in smaller and more reliable uplifts, and fewer extreme deviations between cost outturns and budgets (thereby requiring less trimming of the datasets when computing suitable uplifts).

95% values appear to be calculated using the 5% and 10% and the 90% and 95% FBC values, respectively (i.e., cells E4 and E22 are formulas). We were unsure whether this is intentional? Finally, several values across different categories appear to be identical (e.g., the Road (OBC) and Road (SOC) values for 15% are both identical, in all 13 decimal digits; this happens with several values).

8. References

De Reyck, B., Grushka-Cockayne, Y., Fragkos, I., Harrison, J., Read, D. & Bartlett, M. (2015): Optimism Bias Study

Flyvbjerg, B. (2004): Procedures for Dealing with Optimism Bias in Transport Planning

Mott MacDonald (2002): Review of Large Public Procurement in the UK