Technical critique of HAL's CAP3195 Supporting Document 1: Megaproject Outcome Drivers – Benchmarking Fact Base

Prepared by

Oxford Global Projects (OGP)

Version: Final (Non-Confidential)

Version date: 11th June 26

Authors: Dr Giuseppe Sassano, Dr Dirk Poker, Dr Samuel Franzen, Dr Alexander Budzier

Table of Contents

1 Executive Summary

1.1 Introduction

The CAA's CAP3195 working paper considered whether the existing regulatory model for Heathrow remains appropriate and whether alternative models should be considered. HAL's response to the CAA's CAP3195 working paper deployed Supporting Document 1: Megaproject Outcome Drivers - Benchmarking Fact Base ("the Megaproject Outcome Drivers document") as a key evidential foundation for a specific regulatory argument: that Heathrow expansion should be treated as a privately financed megaproject requiring single-owner integrated delivery, constrained stakeholder and airline rights, design maturity before baseline lock-in, dedicated or accelerated regulatory oversight, and RAB-based financeability. HAL argues that regulatory models that are not consistent with these conclusions or exhibit what HAL describes as megaproject failure modes should be discounted.

Although the regulatory process has moved to CAP3251, which consults on a shortlist of models for further development, the Megaproject Outcome Drivers document is still relevant. Its conclusions are likely to shape HAL's CAP3251 response, and the weight that can be placed on those conclusions depends directly on the quality of the underlying evidence.

This report assesses the quality of the evidence and conclusions presented in the Megaproject Outcome Drivers document using three research questions. The first is methodological: whether the document meets the standards of transparency, balance and proportionality required to justify treating it as decision-grade evidence for shaping or narrowing regulatory model options. The second is evidential: whether HAL's specific claims are accurately grounded in the evidence it cites, and whether the regulatory conclusions drawn follow proportionately from that evidence. The third concerns risk quantification: what the empirical record on airport megaproject cost and schedule overruns implies for Heathrow expansion at its current stage of maturity — a question the Megaproject Outcome Drivers document raises but does not resolve.

1.2 Methodological quality review

This review assesses the Megaproject Outcome Drivers document against seven practical safeguards for judging whether an evidence synthesis is transparent, balanced and proportionate. Those safeguards cover framing, evidence identification and selection, appraisal of evidence quality, treatment of mixed or contrary evidence, transferability, proportionality, and the separation of evidence, interpretation and recommendation.

The Megaproject Outcome Drivers document adopts a framing that emphasises HAL's delivery criteria while giving insufficient weight to the CAA's wider statutory duties. It relies on an illustrative rather than definitive evidence base without explaining how sources were identified, selected or excluded, and does not consistently appraise the strength or limitations of different forms of evidence. Case-study material that is mixed, neutral or context-specific is sometimes treated as directionally supportive, while lessons from other sectors, ownership structures and project contexts are transferred to

Heathrow without a sufficiently explicit causal bridge. The methodological problem is therefore not simply that some evidence is contestable, but that the document does not provide the transparency needed to judge whether its conclusions reflect the balance of the available evidence.

The appropriate implication is one of regulatory weight. The Megaproject Outcome Drivers document does not present sufficient methodological quality for a decision-maker to determine whether it represents a neutral synthesis of the available evidence or a more selective narrative. It should therefore be treated as identifying issues for examination, not as decision-grade evidence for narrowing or rejecting options for regulatory reform.

1.3 Headline conclusion testing

This section tests the principal conclusions in the Megaproject Outcome Drivers document that appear most relevant to HAL's position on the CAP3195 regulatory models. For each conclusion, we identify the claim being made, the evidence HAL relies upon, how the claim is used to support HAL's preferred regulatory design choices, and whether the evidence fairly supports the conclusion drawn.

Our findings are summarised in Table 1-1. It shows where HAL's evidence is useful, where the inference becomes too strong, and what conclusion the material can more safely support. In all cases, the evidence supports a delivery principle but not the stronger regulatory inference. The 33% ownership metric supports integration discipline, but not necessarily ownership concentration. The stakeholder evidence supports bounded rights, but not the marginalisation of airline input. The FEED evidence supports front-end maturity, but not a FEED-only baseline trigger. The acceleration precedents support better sequencing, not weaker scrutiny. The financeability evidence supports predictable revenues and risk allocation, but not automatic absorption into HAL's existing RAB. In each case, HAL has identified important delivery principles. What the evidence does not establish is that these principles can only be met through HAL's preferred regulatory model.

Table 1-1 – Summary findings from headline conclusion testing

HAL finding tested	Evidence and examples assessed	HAL's institutional inference	Scope and limits of the evidence	Evidence-based interpretation
Single-owner integrated delivery *"Single 'owner' with integrated operating company is associated with better delivery and more reliable performance by bringing clear roles, responsibilities, and accountabilities to parties and empowering decision making."*	Gil and Fu (2022): approximately 33% higher cost per additional owner, based on three UK megaprojects. Ibrahim et al. (2020) and Lovallo et al. (2023): Integrated Project Delivery associated with approximately 15 percentage points lower cost overruns than Design-Bid-Build, where overruns average around 20%. Heathrow T2, JFK and BER are used as supporting case evidence.	HAL uses the reported evidence to argue that single-owner integration reduces cost, interface risk and accountability gaps compared with more distributed structures.	The source base points to the need for integration, but the mechanisms are not all ownership-based. Gil and Fu (2022) is an exploratory governance study; IPD is a collaborative multi-party model; and JFK and BER turn primarily on system planning, interface control and technical authority.	The evidence supports clear integration authority, defined roles and enforceable system standards. It does not show that development rights, financing exposure, construction control and operational authority must all sit inside one corporate entity.
Constrained stakeholder and airline rights *"Appropriately calibrated stakeholder engagement, including tightly defined veto rights, support timely delivery while protecting consumer interests."*	Gil and Fu (2022): each additional stakeholder admitted to megaproject decision rights is associated with approximately 0.7–1.2% higher total cost, using the same three megaprojects dataset. Shenhar and Holzmann (2017), BER, HS2, Munich T2 and Heathrow T2 are used to test decision rights, alignment and airline involvement.	HAL treats airline oversight and approval rights as potential sources of delay, scope reopening and baseline instability, using this as a justification for tighter control over stakeholder rights.	The stakeholder metric captures governance complexity, but it does not distinguish wasteful friction from legitimate requirement definition, operational alignment or value negotiation. The case evidence is also mixed: BER concerns technical authority, HS2 concerns escalation and financial-governance rights, Munich T2 formalised airline participation, and Heathrow T2 shows the risk created by unclear sign-off thresholds.	The evidence supports staged and bounded stakeholder rights, not the marginalisation of airline input. Operational requirements should be tested early enough to influence the design of a workable asset, while approval thresholds need to be sufficiently clear to maintain control of scope, affordability and programme sequencing.
Design maturity before baseline lock-in *"Sufficient time to achieve design maturity before delivery targets are locked in, as mature design before*	El Asmar and Gibson (2018): cost growth declines by approximately 0.5 percentage points per one-point increase in FEED accuracy, based on 33 industrial projects. The wider evidence base includes Flyvbjerg and	HAL uses front-end maturity evidence to support fixing cost, scope and schedule once engineering design maturity is reached.	HAL's quantified support is narrower than the regulatory conclusion it draws. El Asmar and Gibson (2018) concerns industrial FEED accuracy, not regulated megaprojects, and the model explains only 18.1% of cost-change variation. The infrastructure precedents, moreover, show	A maturity gate before target lock-in is justified, but it should be a governance-readiness test rather than an engineering threshold. For Heathrow, baseline lock-in should depend on design definition, planning and environmental

HAL finding tested	Evidence and examples assessed	HAL's institutional inference	Scope and limits of the evidence	Evidence-based interpretation
building is associated with improved cost and schedule reliability and fewer late changes."	Gardner (2023), Madrid T4, HS2, SPEED and BER.		that commitments can fail because consent, land, environmental, user-interface, governance and systems-readiness issues remain unsettled even after engineering definition has advanced.	conditions, user interfaces, regulatory assumptions, estimate range, confidence level and residual risks being clearly specified and challengeable.
Dedicated or accelerated regulatory oversight *"Dedicated or fast-tracked regulatory oversight for nationally significant infrastructure can reduce delay and improve cross-stakeholder coordination."*	Flyvbjerg, Holm and Buhl (2004): each additional year between decision to build and completion is associated with approximately 4.6% higher project cost. World Bank (2024): permit and clearance approvals account for approximately 13% of procurement delays in contracts above US$100m. ASTI, HS2 and Tideway are used as acceleration, governance and assurance precedents.	HAL uses delay-cost and acceleration precedents to support faster, more dedicated regulatory treatment for Heathrow expansion.	The sources diagnose different delay mechanisms. The 4.6% figure concerns implementation duration after the decision to build. The World Bank evidence does not mainly point to independent approvals: the 13% permit and clearance figure sits alongside stronger findings on low-quality feasibility, weak procurement capacity and scope modification. ASTI accelerated funding and assessment sequencing while retaining planning interfaces, full project assessment, output obligations and consultation. Tideway shows independent assurance embedded into delivery.	The relevant lesson is not simply to make regulation faster. Approval speed, funding speed and delivery speed should be separated, with each linked to the risk it is meant to control. Heathrow may benefit from earlier sequencing and clearer escalation, but only alongside stronger feasibility discipline, credible baseline assurance, explicit consenting-risk treatment and independent delivery scrutiny.
RAB-based financeability *"RAB-based remuneration, combined with streamlined quality governance and assurance, supports financeability and timely delivery for large, bespoke infrastructure."*	Thierie and De Moor (2019): availability-based payment structures are associated with approximately 28 bps lower loan spreads than demand-based structures, based on around 700 global infrastructure project-finance debt deals. Tideway, nuclear RAB, NAO (2017), Msulwa (2022), Budzier (2026), BEIS (2020) and DESNZ (2025) are also used in this section to test whether the financing evidence supports HAL's asset-boundary conclusion.	HAL treats RAB-based remuneration as central to preserving financeability and delivery confidence for a large bespoke asset.	The financeability evidence supports revenue certainty and risk sharing, but it does not resolve the institutional perimeter. The 28 bps finding concerns availability-based project-finance revenues, not Heathrow price-control recovery. Tideway points to project-specific licensing, ring-fencing, competitive procurement and government support. Nuclear RAB supports construction-period revenue and risk sharing for strategic energy assets, but Sizewell C has not yet produced construction outturn evidence.	The evidence supports structured cost recovery, predictable revenue and explicit risk allocation. It does not show that Heathrow expansion must be absorbed into HAL's existing corporate RAB, nor that financial separation is incompatible with delivery integration.

1.4 Megaproject overrun risk and uncertainty

Section 5 tests the risk implications of HAL's own megaproject framing. If Heathrow expansion is to be treated as a large, complex megaproject rather than ordinary airport capital expenditure, the CAA should consider not only the risks HAL associates with alternative regulatory models, but also the cost and schedule uncertainty inherent in the programme itself.

The cost and schedule overrun experienced by past, completed airport development programmes (the reference-class) shows that airport expansion and new-build programmes frequently exceed their cost and schedule estimates, even when measured from relatively mature Full Business Case (FBC) equivalent baselines. Cost outcomes are particularly exposed to upper-tail risk. The analysis also shows that scope is the dominant correlate of cost overrun within the airport reference class, and Heathrow's proposed expansion is materially broader and more interface-rich than the average project in the pool, extending beyond runway and terminal works into major road, utility, land, operational and surface-access interfaces.

The maturity of the current Heathrow proposals matters. HAL's current estimates are not FBC-equivalent: planning consent, detailed scope definition, procurement, delivery strategy, interfaces and risk allocation remain under development. The mature-stage reference-class results should therefore be treated as a lower-bound view of risk, not as the appropriate central benchmark for HAL's current estimates. The more relevant comparison is with earlier-stage Outline Business Case (OBC) or Strategic Outline Case (SOC) distributions, where the range of plausible cost and schedule outcomes is wider.

The Heathrow-specific evidence does not justify treating the third runway as unusually insulated from that risk. Recent in-flight Heathrow schemes show material estimate movement as projects mature, and the delivered Heathrow record does not establish a reliable basis for assuming the expansion programme will sit outside the airport megaproject distribution. The implication is that the third runway plans should be assessed against reference-class evidence calibrated to current maturity, scale and interface complexity.

The regulatory implication is that HAL's current estimates should be treated as maturity-dependent and uncertain, rather than as a fixed baseline for decision-making. Any reliance on those estimates should be supported by transparent evidence of maturity and disclosure of residual risks, scrutiny of estimate ranges, and explicit consideration of the risks of cost and schedule overrun.

1.5 Conclusions

The Megaproject Outcome Drivers document raises relevant delivery issues, but it does not provide a robust evidential basis for the CAA to rely on HAL's preferred regulatory conclusions. The methodological review shows that the document lacks the transparency, balance and evidential weighting needed for decision-grade evidence. The headline testing shows that HAL's evidence often supports appropriate delivery principles but does not support the stronger institutional conclusions HAL attaches to them. The megaproject risk analysis shows that Heathrow expansion should be assessed

against airport reference-class cost and schedule risk, adjusted for current maturity, scale and interface complexity.

2 Introduction

This report sets out Oxford Global Projects' (OGP) technical critique of HAL's CAP3195 Supporting Document 1: Megaproject Outcome Drivers - Benchmarking Fact Base.

2.1 Context and purpose

The CAA's CAP3195 working paper considered whether the existing regulatory model for Heathrow remains appropriate and whether alternative models should be considered. HAL's response to that consultation sought to shape the process by arguing that expansion should be treated as a privately financed megaproject requiring particular delivery conditions. The Megaproject Outcome Drivers document proposes evidential foundations for that argument to draw conclusions that generally favour integrated owner/operator accountability, tightly defined stakeholder rights, streamlined governance and RAB-based remuneration. HAL uses those conclusions to argue that certain regulatory models are consistent with successful megaproject delivery and others are not.

The regulatory process has since moved to CAP3251, which consults on a shortlist of models for further development. However, the Megaproject Outcome Drivers document is still relevant because it is likely to inform HAL's CAP3251 response. The document therefore has a live and material influence on the regulatory process. That influence needs to be proportionate to the quality of the underlying evidence. This report assesses whether it is.

2.2 Research questions

The report addresses three questions. The first is methodological: is the Megaproject Outcome Drivers document prepared and reported to a standard that would allow the CAA to treat it as decision-grade evidence for shaping or rejecting regulatory model options? The second is evidential: are the specific claims HAL makes accurately grounded in the evidence presented, and do the regulatory conclusions drawn follow proportionately from that evidence? The third concerns risk quantification: what does the empirical record on airport megaproject cost and schedule overruns imply for the risk and estimate uncertainty facing Heathrow expansion at its current stage of maturity? HAL presents some reference-class evidence on megaproject overruns, drawing in part on OGP's own data, but does not translate it into a programme-specific risk assessment calibrated to Heathrow's current position. This report does so.

2.3 Scope of the review

This report is submitted as a supporting document to Heathrow Reimagined's CAP3251 submission. The report does not develop a preferred regulatory model for Heathrow, nor seek to fill every evidential gap in HAL's analysis. The purpose of this report is to test whether the evidential foundations of HAL's position on regulatory model design are robust enough to support the conclusions it draws, and to identify where those conclusions should be accepted, qualified or rejected as insufficiently supported.

3 Methodological quality review

3.1 Basis for the methodological review

HAL uses the Megaproject Outcome Drivers document to support propositions about the delivery conditions it considers necessary for expansion — propositions that bear directly on the treatment of regulatory models in the CAP3251 process. The question is not whether the document raises relevant issues, but whether it meets the basic standards of transparency, balance and reliability required to be considered decision-grade evidence for shaping, narrowing or rejecting regulatory options.

This review assesses the Megaproject Outcome Drivers document against seven practical safeguards — set out in Annex 8.1.1 and applied in Annex 8.1.2 — that allow a decision-maker to judge whether an evidence synthesis is transparent, balanced and proportionate to the conclusions drawn: clarity and neutrality of framing; transparency of evidence identification and selection; appraisal of evidence quality and limitations; balance and treatment of mixed or contrary evidence; relevance and transferability; proportionality of conclusions to evidence strength; and separation of evidence, interpretation and recommendation.

3.2 Clarity and neutrality of framing

The Civil Aviation Act (2012) gives the CAA a general (“primary”) duty, to carry out its functions in a manner which it considers will further the interests of users of air transport services regarding the range, availability, continuity, cost and quality of airport operation services (AOS). The CAA must also carry out its functions, where appropriate, in a manner that will promote competition in the provision of AOS. The CAA's own Consumer Strategy (CAP2593, 2023) translates those requirements into an explicit goal for its economic regulation function: "Protecting consumers from substantial market power, ensuring value for money and establishing incentives for efficiency with an increasing focus on environmental sustainability." HAL's framework in the Megaproject Outcome Drivers document addresses financeability, timeliness and quality end-result. That is a selection from a wider set of statutory objectives, not a neutral analytical starting point.

The selection has directional consequences. Cost scrutiny, competitive pressure and independent assurance are primary mechanisms through which a regulator protects consumers in a capital programme. In the Megaproject Outcome Drivers document, each is assessed through its effect on HAL's criteria rather than through its protective function: stakeholder challenge appears as a source of delay, independent scrutiny as a governance burden, competitive procurement as an interface risk. A more balanced evidence base would test both the delivery risks of distributed accountability and the consumer protection risks of concentrating all development authority in a single promoter. The Megaproject Outcome Drivers document does not do so. This leaves cost to consumers, value for money, competitive discipline, independent assurance and environmental mitigation outside the analytical frame entirely.

3.3 Transparency of evidence identification and selection

The Megaproject Outcome Drivers document describes its evidence base — megaproject literature, airport case studies, and UK infrastructure precedents — as illustrative and not exhaustive. HAL then uses that illustrative evidence base to identify what it presents as the key influencing drivers of megaproject success across structure, governance and incentives. Those are strong causal claims. Identifying which factors are the key drivers of complex programme outcomes requires comprehensive, systematic and balanced review of the available evidence — not an illustrative selection that, by HAL's own description, makes no claim to be representative. The document provides no search strategy, no inclusion or exclusion criteria, no excluded evidence log, and no account of how case studies were chosen. A decision-maker reading it cannot determine whether the conclusions reflect the balance of available evidence or the selection choices of a promoter with a direct interest in the outcome.

3.4 Appraisal of evidence quality and limitations

HAL relies on different forms of evidence: academic literature, small-sample quantitative findings, practitioner analysis, case studies and policy precedents. These sources do not carry the same evidential weight. A regression finding from a small sample, a single airport case study, a government review of a rail programme and a practitioner report may all be relevant, but they do not support conclusions with the same degree of confidence. A decision-grade evidence base should identify the limitations of each type of evidence and explain how much weight should be placed on each finding. The Megaproject Outcome Drivers document does not provide that appraisal. This matters because several conclusions in the Megaproject Outcome Drivers document depend on relatively strong causal claims. But they are not established simply by citing examples of projects that experienced governance or coordination problems.

3.5 Balance and treatment of mixed or contrary evidence

In several places, the report draws relatively firm conclusions even where the underlying case-study evidence is mixed, neutral or potentially contradictory. This is particularly important where HAL's summary matrix classifies a substantial proportion of case-study observations as neutral or as 'possible challenges' rather than supportive, yet the accompanying narrative presents each finding as directionally established. Neutral evidence does not corroborate a proposition; it indicates that the available case material provides no reliable basis for a conclusion in either direction.

That is not a minor presentation issue. If neutral or contradictory evidence is converted into apparent support through narrative synthesis, the resulting conclusion becomes materially less reliable. For example, the document relies on single-owner or integrated delivery as evidence of clearer accountability and better performance, but some of the cases discussed under that framework also involved material cost growth, delay or governance complications. Madrid Terminal 4 is described as a single-owner public sponsorship model, yet the same case records a two-to-three-year delay and around 35% cost escalation. Therefore, a single-owner model is no guarantee of cost or schedule performance. HAL does not draw that lesson from its own case material.

3.6 Relevance and transferability

Transferability needs to be demonstrated rather than assumed. Several precedents in the Megaproject Outcome Drivers document involve publicly funded programmes, direct ministerial control, or materially different risk allocation structures, and the document frequently abstracts general lessons from them without explaining why those lessons apply to a privately financed, economically regulated airport expansion.

The transferability problem is highlighted where HAL relies on its own recent delivery record. The Megaproject Outcome Drivers document presents Terminal 5 and Terminal 2 as evidence that an integrated operator model supports reliable capex performance, citing both programmes as delivered on time and within budget. HAL is simultaneously the party presenting this evidence and the subject of it. Self-reported track record carries little independent evidentiary weight without external verification, especially when only a self-selected partial track record is presented. But even setting that aside, the transferability of T5 and T2 performance to the Heathrow Expansion is questionable.

T5 proceeded to construction after a long planning inquiry that had resolved the principal consenting risks before work began on site. The expansion programme is at no equivalent stage of planning or design maturity, is many times the scale of T5 in real price terms, involves significantly greater scope (including moving the M25 and other key infrastructure and displacing residential and commercial property), substantially more novel in its risk profile, and subject to consenting and regulatory uncertainty that T5 had largely resolved before a single contract was let. T2 is an even weaker comparator. The programme replaced an ageing terminal building on an existing footprint — a capital renewal exercise within a settled regulatory framework, rather than a new-build expansion involving a new runway, genuinely novel infrastructure and a contested governance structure.

3.7 Proportionality of conclusions to evidence strength

HAL argues that fragmentation, unclear decision rights, weak design maturity or financing uncertainty can create delivery risks. Those risks should be considered. But the evidential threshold is higher where a conclusion is used to influence which regulatory models remain on the shortlist or how they are designed.

Several of HAL's conclusions do not reach the required threshold. Evidence that some megaprojects experienced coordination problems does not by itself establish that integrated ownership is necessary. Evidence that delayed approvals can increase costs does not necessarily establish that stakeholder challenge should be reduced rather than structured differently. Evidence that financing certainty can support delivery confidence does not automatically imply that a particular regulatory structure is required at Heathrow. HAL has identified real risks. What the evidence does not establish is that its preferred regulatory models are the necessary or appropriate response to those risks.

3.8 Separation of evidence, interpretation and recommendation

The Megaproject Outcome Drivers document is described as a fact base, but much of the material is interpretive. It goes beyond observation. It infers relationships between delivery structures, governance arrangements, incentives and outcomes, and links those

inferences directly to regulatory model design. That interpretive work is legitimate, but it should be visible. The CAA should be able to distinguish observed evidence from HAL's interpretation of causality and from HAL's regulatory preference. The Megaproject Outcome Drivers document does not consistently maintain that distinction. Statements about the benefits of single point accountability, streamlined governance or calibrated stakeholder engagement are often presented with the authority of generalised findings, even where the underlying evidence is mixed, context-dependent or open to alternative interpretation. The result is that the document reads as more conclusive than its reported method can support.

3.9 Summary: methodological quality review

The Megaproject Outcome Drivers document raises relevant issues for the CAA to consider, but it does not provide sufficient methodological transparency for the CAA to place independent decision-making weight on its conclusions. The problem is not simply that individual sources can be challenged. It is that the document does not provide the audit trail needed to understand how evidence was identified, selected, excluded, weighted or transferred to the Heathrow context.

This matters because the Megaproject Outcome Drivers document is used to support propositions that bear directly on regulatory model design. A decision-maker cannot readily determine whether the document reflects a balanced synthesis of the available megaproject evidence or a more selective narrative constructed around HAL's preferred delivery conditions. That uncertainty does not make the document irrelevant, but it limits the regulatory weight that can safely be placed on it.

The Megaproject Outcome Drivers document should therefore be treated as a stakeholder evidence submission that identifies issues for further examination, not as decision-grade evidence for narrowing, rejecting or materially constraining CAP3251 options. The next section tests HAL's most material conclusions on their own terms: what claim is being made, what evidence is relied upon, whether that evidence is fairly interpreted, and whether the conclusion drawn is proportionate to the evidence presented.

4 Headline conclusion testing

4.1 Basis for the headline conclusion testing

This section examines the main megaproject evidence base used in the Megaproject Outcome Drivers document. The focus is not the full regulatory model assessment, nor every argument HAL advances. It is the set of delivery claims that connect HAL's evidence on megaproject outcomes to its conclusions on structure, governance and incentives.

Five issues are tested: structure, airline decision rights, baseline maturity, regulatory acceleration and financeability. In each case, HAL identifies a real delivery issue. The test is how far those propositions take HAL's arguments. For example, evidence that supports integration does not necessarily settle ownership.

The purpose of this section is therefore to test the proportionality of the inference. Where the evidence supports HAL's conclusion, our analysis confirms that. Where it supports a narrower or different conclusion, the implication is that the model preference built on that evidence should be treated with corresponding caution.

4.2 STRUCTURE - Single-owner integrated delivery: integration authority does not require ownership concentration

The Megaproject Outcome Drivers document treats integration and ownership as closely linked. HAL brings together the 33 percent cost per owner estimate, the IPD benchmark, Heathrow T2, JFK and BER to support the proposition that a single integrated owner operator gives the strongest basis for reliable delivery.

The evidence does not need to be dismissed for this argument to be tested. Heathrow expansion clearly requires integration authority, accountable interfaces and common system standards. The question is whether the material HAL cites proves that these functions have to be delivered through ownership concentration.

This subsection tests that question in three steps. It starts with the 33 percent regression, because that is HAL's most direct quantitative support. It then turns to Integrated Project Delivery (IPD), where the performance logic is collaborative rather than corporate. It closes with JFK and BER, where the central problem is not third-party participation as such, but the absence of system planning, interface control or effective technical authority.

4.2.1 The 33% cost-per-owner claim: an exploratory three-project regression cannot support a general rule

Gil and Fu's governance study rests on three UK megaprojects: Heathrow T2, London 2012, and Crossrail. The authors describe it explicitly as exploratory. A regression built on three projects cannot establish a general causal rule and HAL's deployment of the 33 percent per-owner cost premium as though it were a settled empirical finding gives the underlying analysis more weight than it can bear.

Even on its own terms, the regression does not support the interpretation HAL places on it. Gil and Fu do not characterise megaproject cost evolution as a function of governance fragmentation causing execution waste. Their analysis traces cost changes to

governance adaptation over time to the renegotiation of value creation and distribution among actors whose cooperation was essential to pool resources, secure legitimacy, operational integration, and political alignment. Part of the cost escalation associated with additional governance actors therefore reflects the price of securing institutional cooperation that the project could not have proceeded without, representing, in all effects, a foundational characteristic of the polycentric nature of these projects. Treating the entire 33% premium as evidence of structural inefficiency caused by governance distribution conflates the cost of coordination with the cost of legitimacy and operational alignment. These are categorically different things, and the distinction matters directly for the regulatory question of whether to constrain stakeholder participation.

The specific case HAL invokes as a positive illustration, Heathrow T2, further complicates the interpretation. Gil and Fu describe T2 as a hybrid governance arrangement in which Star Alliance airlines were enfranchised through structured consultation rights, alliance-based coordination mechanisms, and de facto blueprint sign-off authority before construction could begin. This was not a purely unilateral owner-led structure. It incorporated governance characteristics specifically designed to align multiple long-term users' operational requirements with the asset being built. If T2 is one of the three projects in the regression, and T2's governance included systematic airline involvement, then the 33% premium cannot be attributed to multi-actor structures in the way HAL implies. Because T2 itself combined central owner control with structured airline involvement, the 33 percent figure does not isolate the effect of ownership fragmentation in the way HAL's interpretation requires.

The appropriate regulatory conclusion from this evidence is that clear governance arrangements and well-defined coordination rights matter for megaproject performance. The evidence does not support the more specific conclusion that all governance actors beyond a single owner-operator are net costs to delivery.

4.2.2 IPD evidence supports collaborative integration, not corporate concentration

Integrated Project Delivery frameworks have recorded a reduction in project cost overruns of approximately 15 percentage points compared with traditional Design-Bid-Build arrangements (Ibrahim et al., 2020; Lovallo et al., 2023). HAL uses this statistic to support the case for integrated delivery structures. The benchmark is relevant, but the institutional logic of IPD runs directly against the argument it is being used to support.

IPD is a collaborative delivery model structured around early involvement of key participants, shared commercial incentives, joint problem-solving, and risk-sharing across the owner, designers, contractors and major delivery participants (AIA, 2007; NASFA et al., 2010). Its performance benefits arise from aligning incentives across multiple parties, not from concentrating decision-making within a single executive hierarchy. An IPD framework is closer in institutional design to a structured multi-party arrangement with contractually embedded collaboration than to a fully centralised single-owner delivery model.

The IPD evidence is therefore better read as support for structured collaborative integration across delivery participants, rather than as evidence that exclusive development authority is required to secure the performance benefits associated with integration.

4.2.3 JFK and BER: system planning failures, not the cost of third-party participation

To reinforce its preference for an integrated delivery model, HAL draws on the delivery histories of John F. Kennedy International Airport (JFK) and Berlin Brandenburg Airport (BER) as cautionary examples of structural fragmentation. In HAL's framing, these cases illustrate how multi operator configurations and disaggregated delivery structures can contribute to lower asset utilisation, schedule delay and weaker service quality. These are legitimate risks to examine. However, the institutional and engineering record of both airports suggest that their outcomes cannot be explained by third party participation or multi contract delivery structures alone.

JFK's terminal estate developed through separate projects led by individual airlines or private terminal operators, with no overarching master plan. The Port Authority's *Vision Plan* links this history to a fragmented airport, inconsistent service levels, and weak inter-terminal connectivity. These are real deficits showing uncoordinated terminal development can damage whole-airport performance. But JFK's problem was the absence of airport-wide planning authority, not the presence of third-party participation. The Port Authority's response was not to reject private participation. It called for airport-wide master planning, common standards, better terminal interconnection, and continued redevelopment in partnership with private lessees. JFK's lesson is that system planning was absent, not that third-party participation was the problem. The risks the JFK history identifies can, in principle, be addressed through regulatory design, enforceable integration standards, common-use requirements and airport-wide planning obligations, rather than through consolidating all development authority in a single operator (Port Authority of New York and New Jersey Airport Advisory Panel, 2017).

BER is better read as a warning about weak integration, immature technical coordination and insufficient client-side challenge than as a general argument against distributed delivery. The Berlin inquiry records that after termination of the general planner, planning responsibilities were dispersed across a large number of individual offices, interface coordination was insufficient, and trades were not properly captured at their boundaries (Abgeordnetenhaus von Berlin, 2016; 2021). These were genuine integration failures. But the failure was not caused by the mere presence of multiple contractors. It was driven by client-side governance: technical challenge and cost discipline were weakened when they conflicted with the airport's preferred programme and budget. BER's difficulties arose from immature design being taken into construction, unresolved fire protection and smoke extraction problems, governance instability and weak technical oversight. The case therefore points to the need for sequencing discipline, empowered technical assurance and a client function that preserves independent challenge. It does not show that competitive or distributed delivery structures are structurally unworkable.

Both JFK and BER support the need for a clearly accountable integration authority and enforceable system-level standards. Neither establishes that development rights, financing exposure, construction control and operational authority must all be vested in the same corporate entity. The stronger lesson is one of governance design: parties carrying material financial, delivery or operational exposure need clearly defined roles, bounded rights and effective escalation routes. That is different from full institutional concentration under HAL.

4.3 GOVERNANCE - constrained airline decision rights: the evidence supports calibrated thresholds, not the marginalisation of airline input

The decision rights evidence raises a different issue. HAL is right to identify a delivery risk where approval rights are late, open ended or capable of reopening settled scope. That is not the same as showing that airline involvement is itself a problem.

This subsection tests that distinction against three strands of evidence: the general alignment literature, Gil and Fu's stakeholder cost metric, and the airport cases where airline involvement was either formalised within governance or left to operate through less clearly defined approval rights. The timing matters. Airline knowledge is most valuable before the asset is fixed in design and price; it becomes most disruptive when it re-enters the process after those decisions have closed.

4.3.1 Airline oversight as an integration asset

Large infrastructure programmes require clear decision rights, defined escalation routes and thresholds that prevent settled decisions from being re-litigated indefinitely. That principle is well supported by the megaproject literature, and BER and HS2 illustrate the consequences of weak technical authority, blurred accountability and ineffective escalation. The question is whether HAL's evidence supports a further inference: that airline approval mechanisms should be characterised primarily as sources of delay.

Airlines at Heathrow are not generic external stakeholders. They are the long-term operational users of the assets being planned, their charges fund the regulated revenue framework, and they hold practical knowledge about whether terminal layouts, baggage systems, stand configurations, passenger flows and connection processes will work in operation. They also have rich operational experience of how other airports work. Shenhar and Holzmann's total alignment framework supports clear roles, rules and communication protocols across project participants, but its argument is that alignment must be established early across the parties whose cooperation is required for delivery, not that stakeholder challenge should be minimised (Shenhar and Holzmann, 2017). Airline requirements excluded too early tend to reappear as redesign pressure, contested scope or compromised asset performance after opening.

The relevant regulatory question is therefore not whether airline involvement should be reduced, but how it should be structured, at what project stages, and against what decision criteria. HAL's evidence identifies a legitimate risk around poorly bounded approval rights. It offers much less support for treating structured airline input as a source of delivery inefficiency.

4.3.2 The 0.7-1.2% stakeholder cost metric: institutional cooperation conflated with inefficiency

The same Gil and Fu dataset that generates the 33 percent per-owner cost premium also produces the finding that each additional stakeholder admitted to megaproject decision rights is associated with a 0.7 to 1.2 percent cost increase. As discussed in Section 4.2.1, the dataset covers three projects and is described by the authors as exploratory. That limitation matters here as much as it does for the owner metric.

The more important issue is interpretation. Gil and Fu (2022) do not present stakeholder inclusion as a simple measure of administrative friction. Their explanation links cost escalation to governance adaptation, value negotiation and the need to secure cooperation from actors who control resources, approvals, operational knowledge or legitimacy. In that setting, cost forecast movement can reflect real scope negotiation and institutional alignment, not only inefficient process. This is partly a sequencing issue. User requirements can carry real cost because they define what the asset must do in operation. If those requirements are incorporated only after the budget has been fixed, they appear as cost growth. If they are tested earlier, they can be priced, challenged and traded off before the baseline is set.

For Heathrow, the practical issue is dependency. Airport infrastructure and airline needs have to be reconciled somewhere in the delivery process. If airline requirements are kept out too early, they do not disappear; they are more likely to return later as redesign pressure, contested scope or operational compromise. The Gil and Fu metric can therefore support caution over unbounded stakeholder rights. It cannot, on its own, show that airline participation is intrinsically inefficient. The stronger inference is that airline participation needs to be brought in early enough to shape a workable asset, while its approval rights need to be bounded tightly enough to avoid open-ended preference or repeated reopening of settled decisions.

4.3.3 Munich T2 and Heathrow T2: what structured airline involvement actually produced

The case evidence cited around decision rights is strongest when each case is kept to its proper failure mode. BER is relevant to technical authority. HS2 is relevant to escalation and governance architecture. Munich T2 and Heathrow T2 are more directly relevant to the structure of airline involvement. Read in this way, the evidence supports clearer decision rights and better bounded approval thresholds, but it does not support treating airline participation itself as the underlying delivery problem.

BER supports HAL's concern that technical decisions can be delayed or weakened when authority is not held by an empowered engineering and integration function. The Megaproject Outcome Drivers document uses the case to show technical decision-making being subordinated to political shareholder considerations, with delayed responses to safety and systems risks. That is a serious decision-rights failure. Its relevance to Heathrow, however, is specific. BER concerns unresolved safety and systems issues that were not escalated to, and resolved by, an effective technical authority (Abgeordnetenhaus von Berlin, 2016; Abgeordnetenhaus von Berlin, 2021). Airline approval rights at Heathrow concern a different question: how the operational users of the asset are involved in shaping terminal requirements before construction fixes them into place.

HS2 provides a related but distinct lesson. The Stewart Review found that HS2's governance was too complicated, blurred accountabilities and did not properly recognise the role of government as financial shareholder. It also called for clearer delegation, more efficient approvals, expert challenge closer to delivery detail and better forums for problem solving (Department for Transport, 2025). The Lovegrove Review reaches a consistent conclusion from the perspective of public-sector oversight: where

government carries the financial exposure, it needs effective information, challenge and governance rights, not a passive role at the edge of the delivery structure (Cabinet Office, 2026). This supports balanced decision rights for parties with real exposure to the project. It does not show that external stakeholder consultation was the primary source of cost escalation.

BER and HS2 help define the governance risk: authority must be clear, escalation must work, and parties with real exposure need a structured route into decisions. Munich T2 and Heathrow T2 bring the analysis closer to the specific Heathrow issue: how airline involvement is designed when the users of the asset have material operational exposure to the outcome. Munich T2 was developed through a formal airport-airline joint venture between Flughafen München GmbH and Lufthansa, with the airport operator holding 60 percent and Lufthansa 40 percent (Kerkloh, 2007; Kuchinke and Sickmann, 2007). This was not diffuse veto power. It was a structured governance arrangement in which the anchor airline held formal rights in the development of the terminal it would primarily occupy. The case shows that airline involvement can be embedded within integration rather than treated as external interference.

Heathrow T2 presents a more complex picture and requires careful handling. Gil and Fu describe a governance arrangement in which BAA was formally required to consult the Star Alliance airlines through constructive engagement, while the airlines held de facto blueprint sign-off authority because construction could not begin without their endorsement. They also record collective action problems, including disagreement over the scale and affordability of the campus. The T2 evidence therefore supports HAL's concern that airline involvement can create negotiation pressure when approval rights are not clearly bounded.

What it does not support is a conclusion in favour of curtailing structured airline input. T2 was the project in which operational requirements were systematically worked through before construction began, with the airlines whose operations would depend on the result. The governance friction Gil and Fu record was the friction of defining a major hub terminal's requirement with the users who would live with the outcome. That process produced a better-specified asset. The relevant decision question is not whether T2's governance created cost pressure but whether building to a unilaterally determined specification without that process would have produced a better or worse result. Munich T2 and Heathrow T2 therefore do not support a general case against airline involvement. They support a more precise governance lesson: airline input needs to be formalised early, tied to operational requirements and prevented from becoming an open-ended approval dependency once key decisions have been settled.

Taken together, the evidence supports clear, staged decision rights with defined thresholds and calibrated approval authority, governance structures that bring knowledge into the process early and under defined conditions, rather than governance structures designed to foreclose that knowledge. The cases, furthermore, do not support treating airline approval mechanisms as primarily sources of delay.

4.4 GOVERNANCE - FEED-based baseline lock-in: an industrial FEED regression cannot govern when Heathrow's regulatory baseline is fixed

HAL uses design maturity as a control point for the expansion model. In the Megaproject Outcome Drivers document, Front-End Engineering Design (FEED) evidence is not used only to support better project preparation, it is also used to support when cost, scope, schedule and further input can be treated as sufficiently settled.

This subsection tests that use of the evidence. It first separates the logic of front-end discipline from the governance question of baseline control. It then examines the FEED study itself and the infrastructure precedents that show why regulatory baseline readiness cannot be reduced to engineering definition.

4.4.1 Two claims HAL runs together: front-end discipline and FEED-based baseline control

HAL's argument compresses two different questions. The first concerns project discipline: what needs to be developed before major commitments are made. The second concerns governance authority: who decides when the evidence is mature enough for the baseline to close and for later challenge to be constrained.

The project discipline point is well founded. Major projects should not commit to cost, schedule and scope before requirements and constraints are sufficiently understood. Flyvbjerg and Gardner's "think slow, act fast" framing captures the same logic: weak front-end definition tends to reappear later as redesign, re-approval, disruption and cost escalation (Flyvbjerg and Gardner, 2023).

The governance point is different. A stronger design package may reduce uncertainty, but it does not settle planning consent, environmental obligations, airline interfaces, regulatory assumptions, estimate confidence or residual risk. At Heathrow, those factors can materially affect the final cost and schedule position even after engineering definition has advanced. Treating design maturity as the main threshold for baseline lock-in would therefore narrow a broader readiness question into an engineering one.

4.4.2 Why the industrial FEED regression cannot carry this weight: sector, sample, and explanatory power

El Asmar and Gibson examine 33 completed large industrial projects (oil and gas facilities, refineries, chemical plants, pharmaceutical plants, power plants, paper mills, steel and aluminium mills and similar assets) with a combined total installed cost of US$8.83 billion and an average project cost of approximately US$276 million (El Asmar and Gibson, 2018). The original paper's title — *"The Cost Impact of Front-End Engineering Design (FEED) Accuracy for Large Industrial Projects"* — states this sectoral boundary explicitly. Heathrow expansion is not a large industrial project. Its procurement context, regulatory framework, planning complexity, operational integration requirements, environmental constraints and funding structure have no material analogue in the study's sample.

The statistical result carries an equally clear limit. The relationship between FEED accuracy and cost change is statistically significant, but the model's R^2 is 0.181. More than four-fifths of the variation in cost outcomes across the sample is unexplained by FEED accuracy. The authors themselves identify design, construction, scope and other

project conditions as additional cost drivers. A metric with 18% explanatory power in a sample of 33 industrial projects cannot serve as a prescriptive cost control rule for a nationally regulated airport megaproject. HAL's 0.5 percentage point figure is a directional observation within a specific and narrow context. It should not be treated as a governing criterion for baseline-setting at Heathrow.

Msulwa's analysis of regulated megaprojects establishes why the transferability problem is institutional as well as statistical. Regulated megaprojects are shaped by legal, administrative, regulatory and financing arrangements that interact with technical design maturity in ways the FEED regression does not capture (Msulwa, 2022). An engineering design may reach a high FEED accuracy score while planning obligations, environmental requirements, user interface specifications and regulatory determinations remain live and unsettled. At Heathrow, where the cost and schedule position can be materially changed by planning consent conditions, surface access requirements, noise obligations and the outcomes of regulatory determinations, engineering accuracy is a necessary but insufficient indicator of baseline readiness. If HAL's FEED argument were used to justify fixing baselines or curtailing structured airline input before planning, environmental, user and regulatory variables are settled, that would be premature.

4.4.3 What the infrastructure precedents support: baseline realism, not lock-in triggers

Madrid T4 illustrates the difference between baseline realism and design accuracy as a commitment trigger. The Tribunal de Cuentas identified weaknesses in the information systems used to track investments under AENA's 1999 Master Plan. Delays were linked to works execution, expropriation proceedings, the time required to secure the Environmental Impact Declaration, new requirements introduced after works had commenced, and programming deficiencies (Cortes Generales, 2007). The Boletín Oficial del Estado (BOE) resolution reinforced this diagnosis by calling for planning procedures capable of producing more realistic cost and execution time forecasts, with particular attention to expropriation timelines and environmental approval processes (Comisión Mixta para las Relaciones con el Tribunal de Cuentas, 2007). Cost commitments became unrealistic at Madrid T4 because they were fixed against an immature understanding of land, environmental and institutional constraints, not because engineering design accuracy was insufficient.

Network Rail's SPEED framework supports adaptive gating rather than design-accuracy lock-in. SPEED replaced the previous approach, characterised as inflexible, with a system that allows project managers to adapt their method while maintaining rigour, targeting a halving of delivery time (Network Rail, 2024). HAL endorses SPEED-type adaptive governance elsewhere in its submission. That position is difficult to reconcile with treating a linear FEED accuracy metric as the trigger for fixing the regulatory baseline. Maturity is not normally a single pass or fail threshold. It is assessed against defined scope information, remaining uncertainty, estimate range and confidence. The Megaproject Outcome Drivers document uses FEED accuracy as evidence for front-end discipline, but it does not appear to convert that evidence into a Heathrow-specific maturity threshold, estimate classification, cost range or confidence level.

BER supports the need for sequencing discipline before public commitments are fixed. The Berlin inquiry finding that the fire protection system was inadequately planned and

coordinated when the opening date was announced confirms that immature safety-critical design should not be locked against public milestones (Abgeordnetenhaus von Berlin, 2021). For this subsection, the BER lesson does not extend beyond that specific point.

For Heathrow, the baseline should not be treated as ready simply because engineering definition has improved. A regulated airport baseline also depends on planning consents, environmental requirements, user interface specifications, regulatory determinations and a clearly stated estimate range. An internal FEED score cannot substitute for that wider assessment. Nor can maturity language justify closing structured inputs while the institutional variables most relevant to outturn cost remain unsettled.

4.5 GOVERNANCE - Regulatory acceleration: delivery duration and pre-commitment scrutiny are different problems

HAL's acceleration case draws on evidence that is relevant but not interchangeable. A long implementation period, a procurement delay, a permitting issue, a delayed funding decision and an assurance requirement are different delivery problems. Treating them as one generic case for faster regulation risks losing the practical distinction that matters most.

The subsection separates those mechanisms before drawing the inference. Flyvbjerg, Holm and Buhl concern cost escalation after the decision to build. The World Bank evidence gives more weight to weak feasibility, procurement capacity and scope change than to permits alone. Ofgem's Accelerating Strategic Transmission Investment (ASTI) framework shows earlier regulatory funding and assessment sequencing while retaining planning interfaces, full project assessment and output obligations. Tideway shows assurance embedded within delivery rather than sitting outside it. The issue is therefore not speed versus scrutiny, but whether acceleration is being applied to the right decision point with the right safeguards still in place.

4.5.1 The 4.6% delay cost and the World Bank evidence: neither points at independent regulatory review

Flyvbjerg, Holm and Buhl's transport infrastructure study estimates that cost escalation increases by an average of 4.64 percentage points for each additional year between the decision to build and the completion of construction (Flyvbjerg, Holm and Buhl, 2004). The finding concerns implementation duration. It is strong evidence that long delivery periods are expensive once a project has moved into execution. It does not identify independent pre-commitment scrutiny as a material delay source, and it does not disaggregate causes of extended duration in a way that would support that inference.

Flyvbjerg's separate work on the survival of the unfittest identifies the upstream problem the 2004 paper does not address. Projects reach approval with systematically understated costs and overstated benefits precisely where business cases are not challenged with sufficient rigour (Flyvbjerg, 2009). The two papers together describe a sequential problem: weak pre-commitment scrutiny allows poorly founded projects into execution; long execution periods then compound the consequences of initial underestimates. Accelerating past the scrutiny stage addresses neither problem — it increases the probability that the project entering execution rests on an unrealistic

baseline. HAL deploys the 4.6% figure to argue for faster approval. The full body of work from which that figure is drawn argues for the opposite sequencing: rigorous pre-commitment scrutiny to establish a credible baseline, then disciplined execution to minimise implementation duration. HAL's use of the 4.6% figure captures only one side of that sequencing argument.

The World Bank's analysis of infrastructure procurement delays further complicates HAL's position. The 13% figure representing permit and clearance approvals as a share of delay reasons relates to a subset of 29 contracts above US$100 million within a dataset of 312 contracts (World Bank, 2024). In that same subset, low-quality feasibility reports account for 38% of identified delay reasons, weak procurement capacity for 17%, and scope modification for a further 13%. Across the World Bank's broader review, the leading delay drivers are weak project and contract management, design and scope change, and funding constraints. Independent external approvals are a material delay channel but not the dominant one. A source that primarily diagnoses weak feasibility and implementing-entity capability does not provide meaningful support for treating independent regulatory scrutiny as the principal bottleneck.

The asymmetry in HAL's evidence is material. The World Bank evidence identifies weak feasibility and procurement preparation as primary drivers of infrastructure delay. This suggests that Heathrow's delivery risk is better addressed by strengthening cost assurance and feasibility testing before commitment, rather than by compressing those checks in the name of acceleration.

4.5.2 ASTI, HS2, and Tideway: what acceleration actually required

Ofgem's ASTI framework reduced regulatory assessment stages and brought earlier funding access for strategic onshore electricity transmission projects. HAL cites ASTI as a model for regulatory acceleration. The precedent is relevant, but its scope is narrower than HAL's interpretation suggests. ASTI accelerated regulatory funding and assessment sequencing; it did not remove planning consent, replace statutory approval processes or dispense with cost scrutiny.

ASTI projects were identified through national energy system planning and linked to the Government's 2030 offshore wind and net zero ambitions. Ofgem accepted the needs case for projects included in the ASTI framework, while recognising that projects could still evolve through planning and detailed design. The framework allowed pre-construction and early construction funding before all consents were complete, but full project assessment remained. Ofgem expected full funding requests after the relevant planning application had been submitted and aimed to decide full allowances within six months of planning permission (Ofgem, 2022). ASTI therefore supports earlier regulatory sequencing around a system-planned programme. It does not show that faster regulatory sequencing removes the need for planning, consenting and cost assessment to be addressed on their own terms.

ASTI also retained Price Control Deliverables, Licence Obligations, reporting requirements, delivery incentives, Cost and Output Adjusting Event reopeners and statutory consultation requirements. Ofgem did not make transmission investment faster by removing oversight. It made it faster by sequencing funding decisions earlier, reducing duplication between assessment stages and accepting some early funding risk

where the strategic need had already been established. For Heathrow, the ASTI precedent is therefore best read as support for earlier sequencing within a controlled framework: faster funding and assessment routes, but with cost assessment, output obligations and planning and consenting risks still addressed explicitly.

The Stewart Review's findings on HS2 support the point that programmes of exceptional scale require decision routes and governance structures designed for that scale, not standard departmental models (Department for Transport, 2025). HS2's governance was not fit for purpose: it blurred accountabilities, failed to recognise the government's role as financial shareholder, and did not evolve with the programme. The Review's diagnosis centres on an underpowered sponsor function and political disruption to schedule commitments. Independent technical challenge is not identified as a material driver of cost escalation. HS2 argues for tailored coordination and escalation arrangements commensurate with programme scale. It does not argue for reduced independent review.

Tideway provides the most directly relevant precedent for how independent assurance can be embedded in delivery rather than positioned against it. The Independent Water Commission identified Tideway's Independent Technical Assessor as a successful approach to supervising infrastructure delivery, capable of reviewing and assessing ongoing project performance objectively in relation to cost and schedule (Independent Water Commission, 2025). The Independent Technical Assessor was appointed within the project's liaison and licence framework, with duties to both the regulator and the project parties. Its role was to provide objective, independent review of project performance, primarily in relation to cost and schedule, and its quarterly assessments supported Ofwat's ability to monitor and challenge project costs. When Covid-related delay arose, Ofwat required Tideway to distinguish that delay from other causes before any licence amendment was agreed. Ofwat accepted five months of delay on the evidence provided (Ofwat, 2022). Independent scrutiny was embedded within the delivery framework and exercised proportionately. Tideway shows that independent assurance can be built into the delivery framework rather than slow it from the outside. The relevant lesson for Heathrow is that acceleration and scrutiny are not alternatives: faster sequencing still needs testing of cost, schedule, readiness and delay causation independently from HAL.

Approval speed, funding speed and delivery speed are different problems. ASTI shows that regulatory funding decisions can be sequenced earlier, but only within a framework that preserves planning consent, full project assessment, consumer protection and output obligations. Tideway shows that independent assurance can be embedded within delivery without being treated as an external obstacle. HAL's evidence therefore supports better-specified challenge and funding processes, not substituting speed for scrutiny or treating regulatory acceleration as a way around unresolved consenting and readiness risk.

4.6 INCENTIVES - RAB-based financeability: the precedents point to project-specific structures, not absorption into HAL's corporate RAB

This subsection tests whether HAL's financing evidence points specifically to its existing corporate RAB, or whether it points to design choices that can be separated from that

asset boundary, including ring fencing, construction period risk sharing and transparent assurance over expansion costs.

4.6.1 Tideway: why low-cost finance depended on project-specific design

Thames Tideway Tunnel achieved a regulated return of 2.497 percent. The National Audit Office linked that outcome to the Government Support Package, which reduced investor exposure to low-probability, high-impact risks and enabled Bazalgette Tunnel Limited to finance the project below the return that would otherwise have been possible (National Audit Office, 2017). HAL cites Tideway as evidence for the financeability benefits of RAB-based remuneration. The precedent is relevant, but the financing result depended on more than regulated revenue alone. It came from a project-specific structure combining bespoke licence conditions, the Specified Infrastructure Project Regulations, competitive procurement, a separate licensed Infrastructure Provider, and statutory government support for defined tail risks.

Tideway was financed through a dedicated project structure, separate from Thames Water's existing regulated business. Thames Water carried out early development work, but delivery was transferred to Bazalgette Tunnel Limited as a licensed Infrastructure Provider under the Specified Infrastructure Project Regulations. The financing model combined a standalone project licence, competitive procurement, regulated revenue attached to the new asset and government support for defined tail risks. Msulwa explains why separation was not incidental: Thames Water's debt levels, combined with the scale and uncertainty of the tunnel, made absorbing the project into its existing regulated business potentially damaging to its credit rating, its cost of debt and equity, and its ordinary service provision (Msulwa, 2022). The institutional response was to specify the project separately and attach the RAB to the new asset rather than to the incumbent's existing portfolio.

Financial separation did not require a sacrifice of delivery integration. Budzier's analysis records that Tideway was ring-fenced from Thames Water's balance sheet and directly regulated by Ofwat, while its delivery model combined three geographic NEC3 Option C contracts, a system integrator and an overarching Alliance Agreement (Budzier, 2026). Financial separation and delivery integration were both achieved by design. HAL's implicit suggestion that separating financial risk from the existing corporate RAB must create integration risk is not supported by how Tideway was actually structured.

The household bill evidence reinforces the same point. The NAO linked lower-than-expected financing costs to a reduction in expected bill impacts from earlier estimates of around £70-80 per year. The Independent Water Commission records that, after competitive SIPR procurement, the expected bill impact fell to around £20-25 per year (National Audit Office, 2017; Independent Water Commission, 2025). Those savings came from the combination of competitive procurement, a standalone project structure, regulated revenue, and targeted government support, not from maintaining the incumbent operator's control of the asset within its existing corporate RAB.

The Tideway evidence is better read not as confirmation that RAB-based remuneration supports financeability in general, but as a more specific claim: low-cost finance for a very large greenfield asset required a separate licensed Infrastructure Provider, competitive procurement, standalone asset regulation, and targeted government

support for tail risks. Whether a similar structure could be applied to Heathrow expansion is a question the Tideway evidence legitimately opens. Whether the same financing benefits are available to an expansion programme absorbed into HAL's existing airport-wide RAB is a question the Tideway evidence does not answer in HAL's favour.

4.6.2 The 28 basis points availability spread cannot be directly transferred to HAL's price-control recovery model

HAL also cites Thierie and De Moor's analysis of global infrastructure project finance debt deals, reporting that availability-based payment structures are associated with loan spreads around 28 basis points (bps) lower than demand-based structures (Thierie and De Moor, 2019). The result is relevant because it shows that lenders place value on revenue certainty.

The distinction between the two payment mechanisms is important. In an availability-based structure, repayment is linked primarily to whether the asset is available and performing. In a demand-based structure, repayment depends more directly on how much the asset is used. The 28 basis point difference is therefore best understood as evidence that more predictable revenues can reduce the cost of debt.

That is a useful financing insight, but it does not transfer directly to HAL's proposal. Heathrow expansion would recover costs through the airport price-control framework, where charges, allowed expenditure and traffic forecasts would form part of the regulatory settlement. This is not the same as an availability-payment structure. The evidence therefore supports careful design of a predictable revenue framework for expansion. It does not show that HAL's proposed model would attract the same financing discount observed in availability-based project finance, nor that revenue certainty has to be achieved by placing expansion within HAL's existing corporate RAB.

4.6.3 Nuclear RAB shows how construction risk can be shared; it does not resolve the asset boundary question

The UK Government's nuclear RAB consultation response established that a RAB model can reduce the cost of capital relative to a Contract for Difference by sharing construction risk with consumers and allowing investors to receive revenue during construction (BEIS, 2020). The shift from Hinkley Point C's CfD structure to Sizewell C's RAB model reflects this logic: construction costs are spread across consumers, taxpayers and private investors, rather than being carried entirely by private equity until revenue begins. HAL's use of this precedent to support construction-period revenue and shared construction risk as financeability tools is legitimate as far as it goes (DESNZ, 2025).

Nuclear power stations are sovereign strategic assets where government made a deliberate policy choice to share construction risk as the price of securing an essential national energy investment that private capital would not finance on acceptable terms through conventional structures. The asset boundary of who owns the plant and in what corporate structure, was determined by that policy choice, not through a regulatory process contested between an incumbent and its users. Heathrow expansion involves a privately owned regulated monopoly where the asset boundary is itself the contested regulatory question. The nuclear precedent demonstrates that governments can make large infrastructure assets financeable by changing the timing of revenues and the

distribution of construction risk. It does not establish the institutional perimeter that should apply to an airport expansion, and it does not show that financing benefits require construction risk to sit within HAL's existing corporate RAB rather than in a project-specific structure.

Sizewell C has reached final investment decision. Construction has not been completed, and cost and schedule outturn are not yet observable. The precedent can support arguments about financing design. It cannot yet support arguments about RAB-model delivery performance.

The evidence in this subsection establishes an important financing proposition: structured cost recovery, construction-period revenue and explicit risk sharing can reduce financing pressure for large capital projects. The evidence is not decisive on the institutional step that matters for CAP3251. Tideway points to a separate licensed Infrastructure Provider and a standalone project RAB. Thierie and De Moor address payment certainty and debt pricing, not corporate ownership. Nuclear RAB supports shared construction risk for strategic energy assets in a different ownership and policy context. None of these sources demonstrates that the financing benefits of RAB-based remuneration require blending Heathrow expansion into HAL's existing airport-wide regulated asset base.

4.7 Summary: headline conclusion testing

Section 4 shows that HAL's Megaproject Outcome Drivers document identifies genuine delivery risks but often draws a wider regulatory conclusion than the evidence can sustain. The material reviewed supports integration authority, disciplined front-end development, bounded decision rights, efficient sequencing and financeable cost recovery. Those are important delivery principles. They do not, however, settle the institutional questions that matter for CAP3251.

Across the five areas tested, the same pattern recurs. Evidence on integration is used to support ownership concentration. Evidence on stakeholder complexity is used to cast airline input mainly as delay risk. Evidence on FEED maturity is used to imply baseline lock-in before wider regulatory and operational readiness is established. Evidence on delay is used to support acceleration, even where the cited sources point to weak feasibility, procurement capability and immature baselines. Evidence on financeability is used to favour HAL's existing RAB, even where the strongest precedents point to project-specific risk allocation and ring-fencing.

The stronger conclusion is that HAL's evidence should be treated as a set of delivery tests, rather than as decisive support for one institutional model. A credible delivery structure needs accountable integration, structured airline involvement, mature and challengeable baselines, proportionate assurance and a financeable allocation of construction risk. Section 4 shows that those tests are supported by the evidence. It does not show that they can only be satisfied through HAL's proposed structure.

5 Megaproject overrun risk and uncertainty

5.1 Basis for this section

This section assesses the cost and schedule risk implications of HAL's own megaproject framing. HAL argues that Heathrow expansion should not be treated as normal airport capital expenditure, but as a large, complex, multi-faceted megaproject that requires a particular regulatory and delivery lens. HAL then uses that framing to argue against regulatory models which could introduce delay, fragmentation, interface risk or financing uncertainty.

That framing creates a clear need to examine the empirical risk profile implied by treating Heathrow expansion as a megaproject. If the programme is to be assessed through a megaproject lens, the CAA should understand not only the risks of alternative regulatory models, but also the baseline cost and schedule risks that are inherent in the programme itself. The Megaproject Outcome Drivers document acknowledges that megaprojects are prone to cost and schedule overruns, but its treatment of that evidence is generic. It does not provide a sufficiently specific view of airport-expansion risk, nor does it adjust the evidence for the maturity of Heathrow's current proposals. Both gaps matter, because the realised distribution of overruns on airport projects is heavy-tailed and because overrun risk is strongly affected by the maturity of the baseline from which the estimate is measured.

This section therefore presents a proportionate reference-class assessment of cost and schedule overrun risk on airport capital projects, adjusted for Heathrow's current maturity stage. The purpose is not to produce an independent cost or schedule forecast for the programme. It is to show that HAL's current cost and schedule assumptions should be read inside a realistic distribution of uncertainty, rather than as fixed, reliable or unusually immune to ordinary megaproject risk.

Cost overrun is measured as realised project cost relative to the estimate available at the relevant decision stage. Schedule overrun is measured as realised project duration relative to the estimated duration at the same decision stage. Where maturity is considered, overruns are measured separately from Strategic Outline Case, Outline Business Case and Full Business Case equivalent baselines.

5.2 Reference class and sample overview

This subsection introduces the reference classes used throughout the rest of the section and presents the cost-overrun and schedule-overrun summary statistics.

5.2.1 Reference class definition

The reference class is constructed from completed airport capital projects in OGP's international project database. The sample includes terminal upgrades/expansion, new terminals, and new airports. This covers a spectrum of airport capital projects that combine substantial passenger-handling infrastructure with airfield, surface-access and operational interface works at scale.

5.2.2 Sample composition

The cost-overrun sample contains 39 airport development programmes. The schedule-overrun sample contains 40. The two samples partly overlap and are broadly similar in regional and project-type composition.

All cost figures are normalised to a common currency and price year before cost overruns are calculated. Local-currency estimates and outturns are converted into pounds sterling and inflated to a single base year using World Bank GDP deflators. This ensures that the overrun ratios reflect project cost performance rather than inflation, exchange-rate movements or differences in reporting years.

Figures**Error! Reference source not found.** and **Error! Reference source not found.** summarise the geographic and temporal coverage.

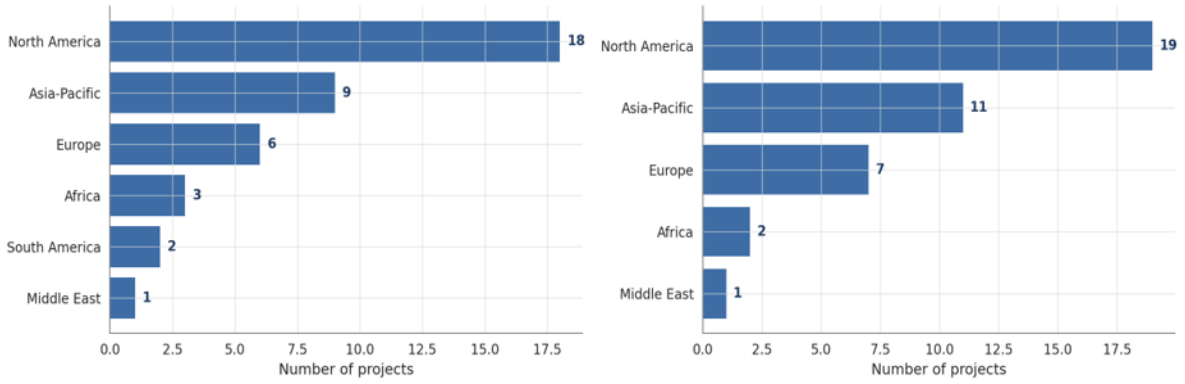

Figure 5-1 — Reference-class project counts by region (cost sample left, schedule sample right)

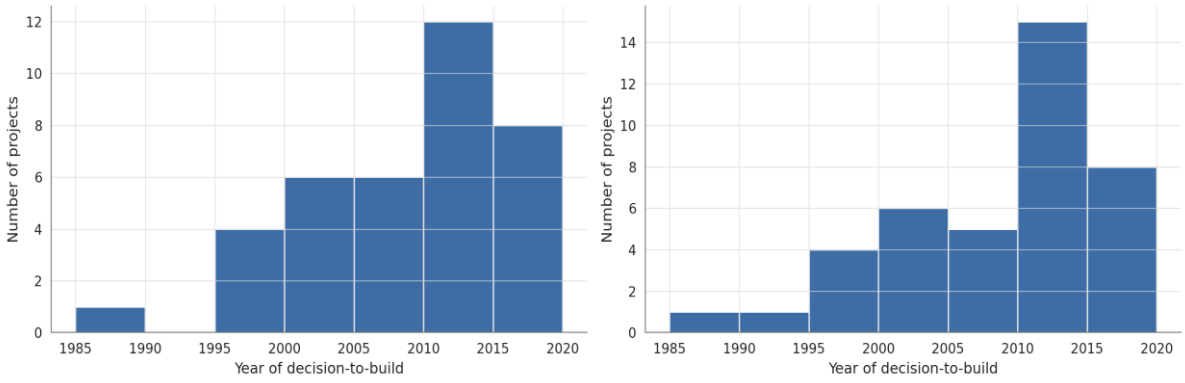

Figure 5-2 — Reference-class project counts by year of decision-to-build (cost sample left, schedule sample right)

5.2.3 Relevance to Heathrow expansion

The reference classes in Sections 5.2.4 and **Error! Reference source not found.** should be considered in relation to the scope of Heathrow's expansion. The disclosed scope of the third runway programme includes a new runway and major terminal extensions, but also substantial off-airport works that few reference-class projects share — relocation of a long stretch of the M25 motorway, displacement of the existing energy-from-waste plant, displacement of data centres, prisons, power and utility infrastructure, large-scale land assembly, surface-access reconfiguration and rail integration, all delivered alongside continuous live operations.

Statistical testing within the airport reference class pool (Annex 8.2) indicates that scope is the dominant correlate of cost overrun. Larger scope airport projects have statistically higher cost overrun than smaller scope airport projects. No other dimension tested produces a comparable signal. Heathrow's scope is materially larger and more interface-rich than the average project in the pool because its scope is greater than terminal expansions and new terminals, but below some new large airports. The pool average therefore likely understates the actual overrun risk a project of Heathrow's scope is likely to face. Therefore, the cost and schedule overrun statistics in Sections 5.2.4 and **Error! Reference source not found.** should be read as a lower bound for the programme's risk envelope, not as a central forecast — a point reinforced by the maturity-stage analysis in Section **Error! Reference source not found.**.

Consideration of scope is not necessary when interpreting schedule overrun because no factors tested statistically influenced schedule overrun, including scope.

5.2.4 Cost overrun in the reference class

The cost-overrun summary for the pooled reference class is shown in Table 5-1. The table reports the sample size, the median and mean cost overrun, and the share of projects exceeding an overrun threshold. These overrun figures are baselined at Full Business Case (FBC) maturity.

Table 5-1 - Cost overrun summary for airport expansion and new-build programmes at FBC

Reference class (at FBC-stage)	n	Median overrun	Mean overrun	> 0% overrun	> 25% overrun	> 50% overrun	> 100% overrun
Airport expansion and new-build programmes	**39**	**34%**	**49%**	**97%**	**62%**	**33%**	**15%**

The cost reference class shows that cost overruns are common in airport expansion and new-build programmes, even when measured from a relatively mature FBC-equivalent baseline.

Almost every project in the reference class exceeded its FBC-stage cost estimate. The median cost overrun is 34%, while the mean is 49%. The mean sits above the median because the distribution has a long upper tail. A majority of projects exceeded their baseline by more than 25%, one in three exceeded it by more than 50%, and around one in six more than doubled in cost.

The shape of the distribution is shown in Figure 5-3: a concentration of projects in the 0% to +50% range, a long upper tail running to +200%, and a mean displaced to the right of the median by that tail.

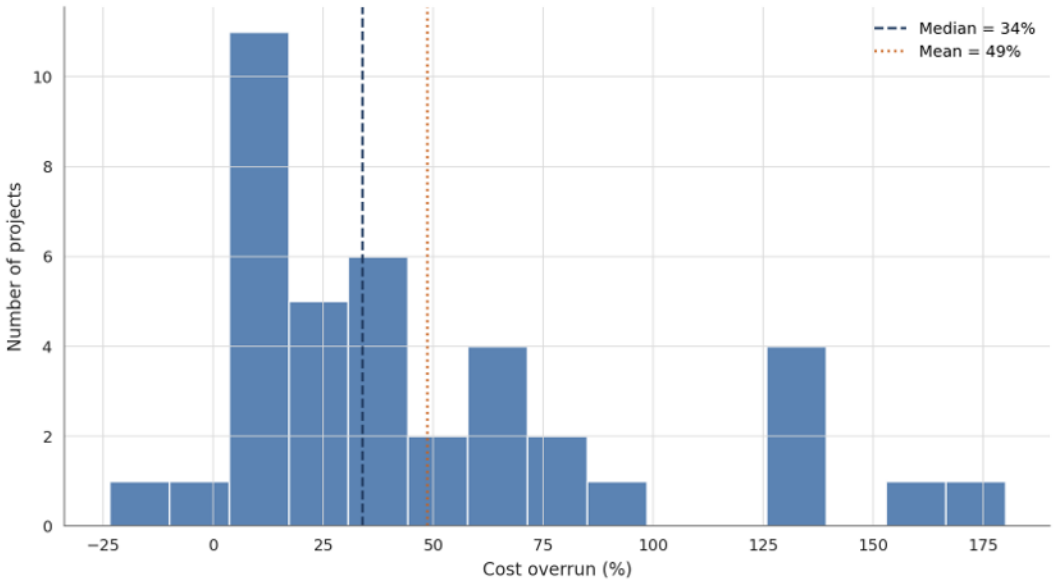

Figure 5-3 - Distribution of FBC-stage cost overruns, airport expansion and new-build programmes (n=39)

5.2.5 Schedule overrun in the reference class

Schedule overruns are also common, although less severe than cost overruns. As shown in Table 5-2, three-quarters of projects exceeded their FBC-stage schedule estimate. One third exceeded it by more than 25%, and 18% by more than 50%. The median overrun is 6%, while the mean is 20%, again reflecting an upper tail of delayed projects.

Table 5-2 – Schedule overrun summary for airport expansion and new-build programmes at FBC

Reference class (at FBC-stage)	n	Median overrun	Mean overrun	> 0% overrun	> 25% overrun	> 50% overrun	> 100% overrun
Airport expansion and new-build programmes	**40**	**6%**	**20%**	**75%**	**33%**	**18%**	**5%**

The distribution shape is shown in **Error! Reference source not found.**: the bulk of the sample sits close to plan, with a smaller upper tail extending past +100%.

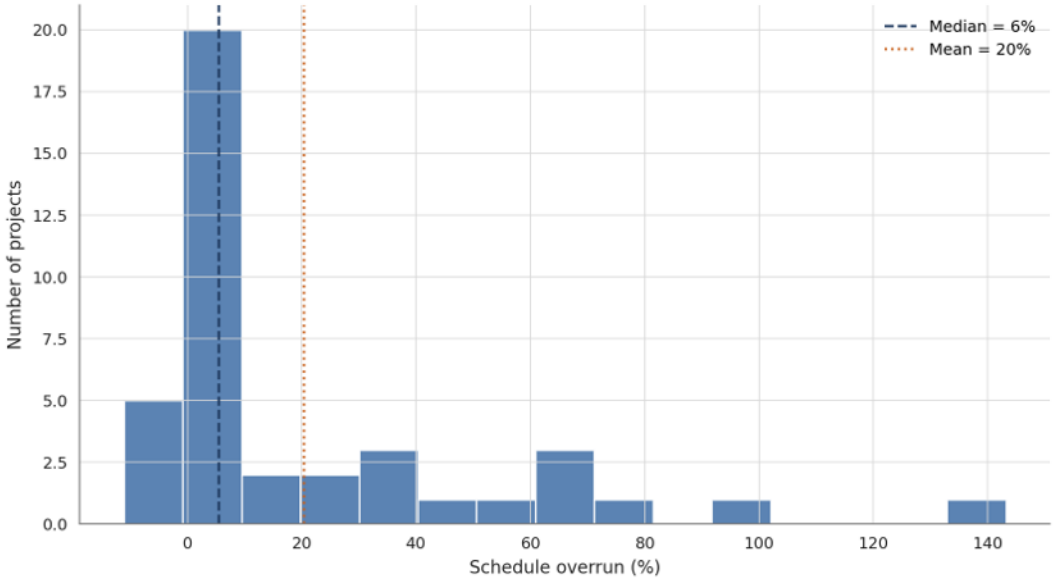

Figure 5-4 - Distribution of FBC-stage schedule overruns, airport expansion and new-build programmes (n=40)

5.3 Impact of maturity on overrun risk

HAL's use of megaproject overrun evidence is incomplete because much of the evidence it cites is measured from relatively mature decision points, such as final investment decision or equivalent. Heathrow expansion is not yet at that stage. Scope, planning consent, delivery strategy, procurement, construction interfaces and risk allocation remain under development.

This matters because estimates become more reliable as projects mature. An FBC-equivalent estimate is not directly comparable with an earlier-stage estimate. Applying mature-stage overrun evidence to an earlier-stage Heathrow estimate would understate uncertainty.

5.3.1 Estimate-maturity stages

The HM Treasury Five Case Model defines three principal business-case stages, each carrying a different level of underlying estimate maturity. The definitions used throughout this analysis are summarised in Table 5-3.

Table 5-3 – Maturity definitions (HM Treasury Green Book — Five Case Model)

Stage	Definition (HM Treasury Green Book — Five Case Model)
Strategic Outline Case (SOC) – Low Maturity	First formal business-case stage. Establishes the strategic context and the case for change; identifies high-level options; demonstrates strategic fit. Cost and schedule estimates are indicative only, derived from strategic-level information; scope, design and delivery arrangements are not yet developed. Used to secure approval to proceed with the development of an OBC.
Outline Business Case (OBC) – Medium Maturity	Revisits the strategic options and identifies the preferred option. Develops the strategic, economic, commercial, financial and management cases in outline. Cost

	and schedule estimates are based on outline design and an indicative procurement strategy. Used to secure approval to proceed with the development of an FBC.
Full Business Case (FBC) – High Maturity	Final business-case stage. Documents the agreed deal with the supplier and confirms the procurement decision. Cost and schedule estimates are based on detailed design, agreed scope and finalised procurement arrangements. Provides the baseline against which delivery is monitored.

This subsection presents the cost and schedule overrun distributions at each of these three estimate-maturity stages. The early- and medium-maturity views capture the additional uncertainty that earlier business-case stages carry, when scope, design, planning and consenting, and delivery arrangements are still being developed.

5.3.2 Maturity-adjusted cost-overrun distribution

The maturity-adjusted cost results show why HAL's current estimates should not be read as final delivery baselines. Even at FBC maturity, airport expansion and new-build projects show material cost overrun risk. At earlier maturity stages, the cost distribution widens. **Error! Reference source not found.** shows that maturity has limited effect around the middle of the distribution, where the three curves are close together. The material difference is in the tails, especially the upper tail. Earlier-stage estimates expose projects to a wider range of downside outcomes because scope, design, consents, procurement and risk allocation are less settled. Estimate maturity therefore matters most for downside risk, rather than for the median outcome.

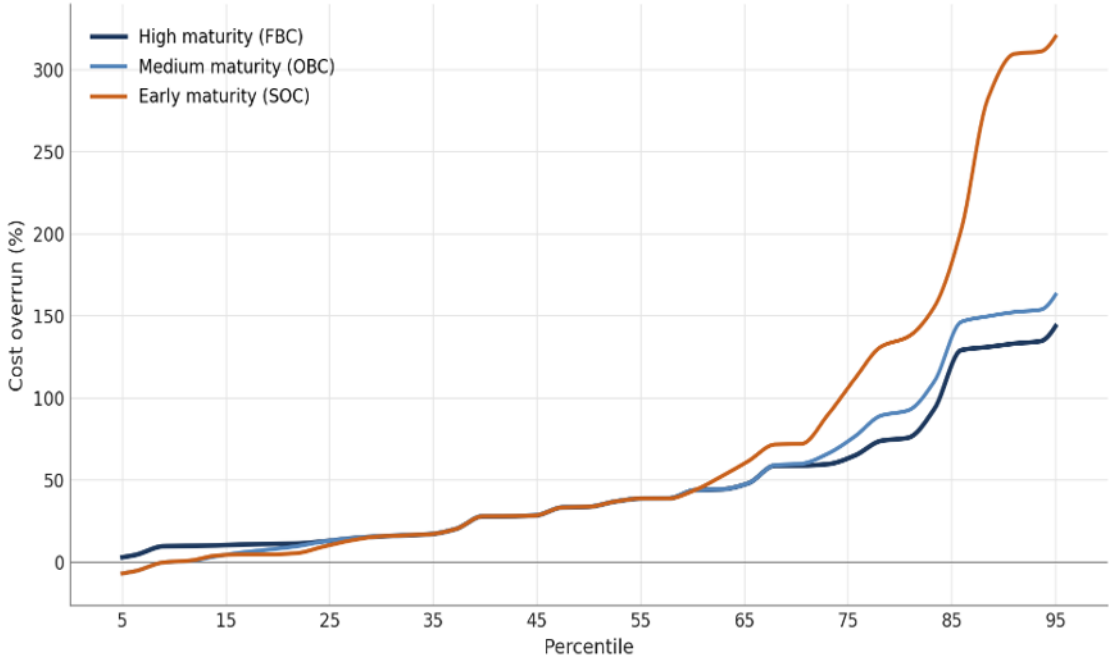

Figure 5-5 - Cost-overrun distribution at three estimate-maturity stages (airport expansion and new-build programmes, n=39)

The summary statistics at each maturity stage for the pooled cost overrun risk reference class are shown in Table 5-4. The mean cost overrun rises from approximately 49% at the

high-maturity FBC view to 52% at OBC and 78% at SOC; the share of projects exceeding +100% rises from 15% at FBC to 26% at SOC.

Table 5-4 – Cost overrun summary for airport expansion and new-build programmes at different maturity stages

Maturity stage	n	Median overrun	Mean overrun	> 0% overrun	> 25% overrun	> 50% overrun	> 100% overrun
High maturity (FBC)	39	34%	49%	97%	62%	33%	15%
Medium maturity (OBC)	39	34%	52%	92%	62%	33%	18%
Early maturity (SOC)	39	34%	78%	92%	62%	39%	26%

5.3.3 Maturity-adjusted schedule-overrun distribution

Error! Reference source not found. reports the equivalent maturity-adjusted curves for schedule overrun. The pattern is similar to cost, with the consequences of estimate maturity being concentrated in the extremes. However, between P70 and P95 the SOC overrun is slightly lower than OBC. This is a feature of the observed data, not an error.

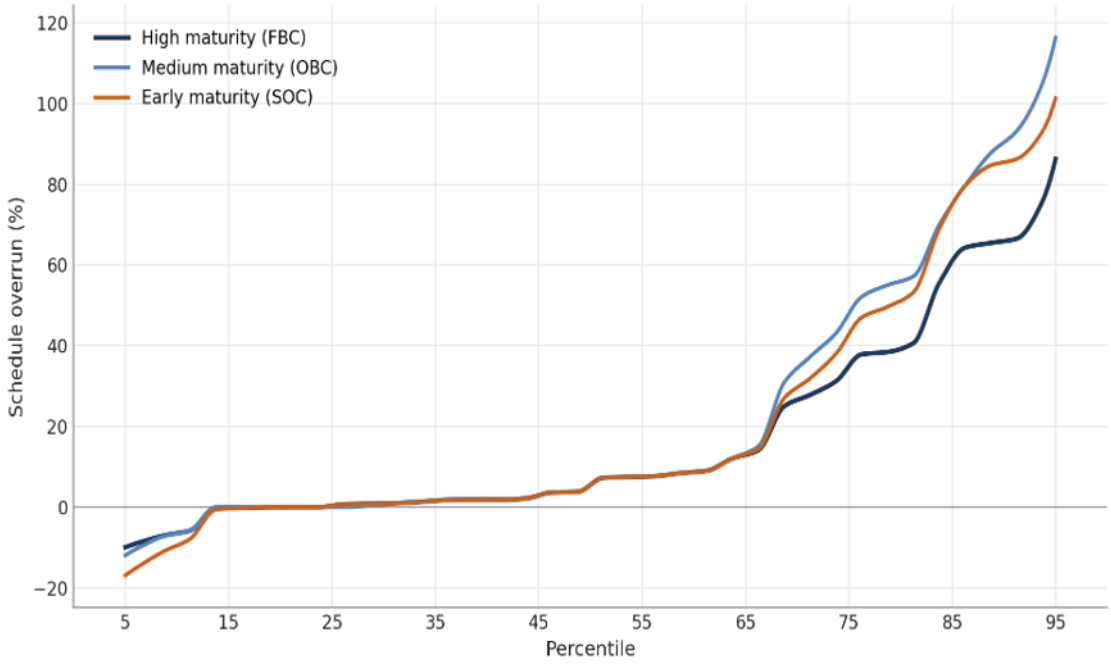

Figure 5-6 - Schedule-overrun distribution at three estimate-maturity stages (airport expansion and new-build programmes, n=40)

As shown in Table 5-5, the schedule maturity effect is less pronounced than cost. The mean schedule overrun rises from 20% at FBC to 26% at OBC and 23% at SOC. The exact ordering is less important than the regulatory point: schedule estimates made before

scope, consents, procurement and interfaces are settled remain more exposed to material uncertainty.

Table 5-5 – Schedule overrun summary for airport expansion and new-build programmes at different maturity stages

Maturity stage	n	Median overrun (%)	Mean overrun	> 0% overrun	> 25% overrun	> 50% overrun	> 100% overrun
High maturity (FBC)	40	6%	20%	75%	33%	18%	5%
Medium maturity (OBC)	40	6%	26%	73%	33%	25%	8%
Early maturity (SOC)	40	6%	23%	75%	33%	20%	5%

5.3.4 What maturity means for Heathrow expansion

Heathrow expansion is not yet at an FBC-equivalent baseline. It is closer to an SOC/OBC position, with scope, planning consent, delivery strategy, procurement, interfaces and risk allocation still under development. The FBC reference class should therefore be treated as a lower-bound risk view, not as the appropriate central benchmark for HAL's current estimates.

Prior option development and pre-development work may be relevant to an assessment of maturity, but it is not the same as FBC-equivalent maturity. The issue is not simply how much work has been undertaken, or how long the programme has been under consideration. The relevant question is which uncertainties have actually been closed. Option development can help identify, compare and refine potential solutions, but it does not by itself establish a fixed scope, a completed design, a procured delivery strategy, a consented planning position or settled risk allocation.

The same logic applies to any material fixed expenditure commitment made before an FBC-equivalent baseline is reached. Such commitments should be assessed against the uncertainty remaining at the point of commitment, not against the lower risk profile associated with a fully mature baseline. Where consent, scope, design, procurement, interfaces or risk allocation remain unresolved, fixed commitments remain exposed to the wider cost and schedule uncertainty characteristic of pre-FBC stages.

Read against the maturity-adjusted evidence, HAL's current estimates should therefore be viewed as one point within a broad range of plausible outcomes, rather than as a mature central expectation for the programme. The width of that range depends on estimate maturity, and at Heathrow's current stage it remains materially wider than under a fully developed FBC scenario. The FBC-adjusted view therefore represents the lower end of the risk envelope rather than its central case.

5.4 Inherent uncertainty in HAL's estimates

5.4.1 Current Heathrow project estimate evolution

Two ongoing Heathrow projects illustrate how estimates can evolve materially as projects move through maturity stages. These examples are not delivered outturns. They show estimate movement before completion and are intended to show that HAL is not immune to maturity-based cost escalation.

As shown in Table 5-6, the Terminal 2 Baggage System Replacement Programme estimate has risen from £xxx million at initial planning in 2021 to a current forecast of £xxxmillion in 2025. That is an increase of 56% before delivery. The project therefore illustrates the same maturity dynamic observed in the wider reference class: early estimates can move substantially as scope, design and delivery assumptions develop.

Table 5-6 - Cost evolution of the T2 Baggage System Replacement Programme

Estimate stage	Year	T2 Baggage System estimate (£m, money of the day)	T2 Baggage Programme estimate (£m, money of the day)
H7 Revised Business Plan	2021		£xxxm
P1 Initial Planning	2022	£xxxm	£xxxm
P2 Developed Estimate	2023	£xxxm	£xxx m
G3 Delivery Approval	2025	£xxxm	£xxx m
Current Forecast	**2025**	**£xxxm**	**£xxx m**

Multi-Storey Car Park 4 shows the same pattern on a smaller project (Table 5-7). The estimate has risen from a £5 million baseline in 2014 to a current forecast of £xx million in 2025. Again, the point is not that these projects forecast the third runway. The point is that Heathrow projects are not immune from estimate growth as maturity increases.

Table 5-7 - Cost evolution of Multi-Storey Car Park 4

Estimate stage	Year	Estimate (£m, money of the day)
Q6 Baseline	**2014**	**£xm**
G1 Concept Approval	2021	£xxm
G2 Developed Design	2022	£xxm
G2 Update	2023	£xxm

Current Forecast	**2025**	**£xxm**

Heathrow cites Terminal 5 and Terminal 2 as evidence that it can effectively deliver large terminal projects. However, these precedents do not provide a sufficient basis for assuming that the third-runway programme is insulated from airport megaproject risk.

The first limitation is maturity. The favourable performance of delivered Heathrow terminal projects is measured from relatively mature baselines, after substantial planning, design and delivery definition had already occurred. Heathrow expansion is not yet at an equivalent stage. A mature-baseline delivery record cannot be transferred directly to current expansion estimates without adjusting for the earlier maturity of the present proposals.

The second limitation is scope and complexity. The expansion programme is materially broader than a terminal delivery programme. The airport reference-class analysis indicates that wider scope is associated with greater cost-overrun risk. That does not mean Heathrow expansion will necessarily perform poorly, but it does mean that selected terminal precedents cannot be treated as a sufficient basis for assuming that Heathrow expansion will outperform the wider airport megaproject distribution.

The Heathrow-specific evidence therefore supports a cautious conclusion. Heathrow has delivered some major capital projects effectively, but it has also experienced material estimate movement on other projects as they have matured. Further, one or two precedents do not establish a general rule of outperformance. The relevant benchmark remains the performance of comparable airport programmes at similar levels of maturity, adjusted for the scale, scope and interface complexity of the third-runway proposals.

5.4.2 HAL's approach to front-end loading

HAL correctly identifies that front-end loading, design maturity and disciplined delivery reduce cost and schedule risk. That proposition is not in dispute. The issue is whether HAL's proposed regulatory approach gives those disciplines enough force. Front-end loading is not simply a label for early design work. It is a governance discipline: gates must operate as real hold points, estimates must not be fixed before scope and design are mature, and independent scrutiny must have the authority to require further work where readiness has not been demonstrated.

The empirical evidence supports this distinction. The Infrastructure and Projects Authority (IPA) has stated that projects with strong front-end loading are typically lower cost and faster than average comparators, and materially better than poorly defined projects (Smallwood, 2020). Construction Industry Institute (CII) research similarly finds better cost, schedule and change-control outcomes where owners achieve high front-end planning scores (CII, 2014). Those benefits arise because uncertainty is confronted before major commitments are made. They do not arise where policy or commercial pressure turns gate passage into a timetable milestone rather than a maturity test (Smallwood, 2020).

That distinction matters for Heathrow expansion. HAL's emphasis on speed, baseline certainty and streamlined challenge could support good delivery if tied to rigorous maturity gates and independent assurance. It could create additional risk if used to fix cost and schedule commitments before planning, scope, design, procurement, interfaces and risk allocation are sufficiently settled.

5.5 Summary: megaproject overrun risk and uncertainty

The reference-class evidence shows that cost and schedule escalation are normal features of airport megaproject delivery, not exceptional failures. Even from relatively mature FBC-stage baselines, airport expansion and new-build programmes frequently exceed their cost and schedule estimates, with cost outcomes showing a particularly wide upper tail.

Scope is the dominant correlate of cost overrun — no other factor tested produces a comparable signal — and Heathrow's programme extends well beyond the runway and terminal works into substantial off-airport infrastructure that few reference-class projects share. A programme of that scope is likely to sit towards the higher-risk end of the distribution, not around its centre.

Where estimates are made earlier, before scope, design, consents, procurement and risk allocation are settled, the range of plausible outcomes widens further. Heathrow expansion is not yet at an FBC-equivalent stage. HAL's current estimates should therefore be treated as early-stage reference estimates, not as fixed delivery baselines or central forecasts of final outturn.

The FBC reference-class results are thus best read as a lower-bound risk view for Heathrow, for two compounding reasons: they reflect a maturity level the expansion programme has not yet reached, and they are drawn from a pool of projects with lower average scope than Heathrow's own programme.

The Heathrow-specific evidence does not justify a more optimistic treatment. Recent in-flight Heathrow schemes show material estimate movement as projects progress through maturity stages, and HAL's delivered track record does not establish a consistent ability to sit outside the airport megaproject distribution. The point is not that Heathrow expansion will necessarily experience a particular level of overrun. It is that HAL has not demonstrated why the third runway programme should be treated as unusually accurate, unusually mature or unusually insulated from normal airport megaproject risk.

The regulatory implication follows directly. If Heathrow expansion is to be assessed through a megaproject lens, HAL's current estimates should not be treated as a mature regulatory baseline without transparent maturity evidence, estimate ranges and residual-risk disclosure.

6 Conclusions

The Megaproject Outcome Drivers document raises issues that are relevant to the CAA's consideration of Heathrow expansion, but it does not provide a sufficient evidential basis for placing decisive regulatory weight on HAL's preferred conclusions. The report identifies real delivery risks: integration failures, unclear decision rights, immature baselines, slow approvals and financing uncertainty are all capable of affecting a programme of this scale. The question is not whether those risks exist. It is whether HAL has shown that they require the particular institutional and regulatory responses it advances.

The methodological review indicates that the Megaproject Outcome Drivers document should be treated with caution. It is not a transparent, systematic or fully weighted synthesis of the relevant evidence base. Its framing gives prominence to HAL's delivery criteria, while the selection, exclusion, appraisal and weighting of evidence are not sufficiently visible. That does not make the document inadmissible as evidence. It does mean that it should be treated as a stakeholder submission that identifies matters for scrutiny, not as decision-grade evidence for narrowing or rejecting regulatory options.

The headline conclusion testing shows a recurring pattern. HAL's evidence often supports sensible delivery principles, but not the stronger institutional conclusions attached to them. The evidence supports clear integration authority, disciplined decision rights, mature baselines, better sequencing and credible financeability arrangements. It does not, without further evidence and independent testing, establish that those objectives require ownership concentration in HAL, reduced airline scrutiny, FEED-only baseline lock-in, weaker regulatory assurance or automatic absorption of expansion into HAL's existing corporate RAB.

The megaproject risk analysis shows the need for caution before relying on HAL's current estimates. If Heathrow expansion is to be assessed as a megaproject, the CAA should assess it against the empirical record of airport megaproject cost and schedule performance. That evidence indicates material overrun risk even from mature baselines, and Heathrow's current proposals remain at an earlier stage of maturity. The relevant benchmark is therefore not a fixed-point estimate, or the favourable outcome of selected Heathrow precedents, but a maturity-adjusted reference-class distribution that reflects the programme's scale, scope and interface complexity.

The practical implication is therefore one of evidential discipline. The Megaproject Outcome Drivers document should not be treated as establishing that particular regulatory models should be discounted, constrained or preferred. At this stage, the document is better read as identifying delivery risks that require further examination than as proving the institutional solution to those risks. The CAA should therefore place limited weight on HAL's stronger regulatory inferences unless they are supported by further, independently tested evidence.

7 References by section

7.1 Section 2 references

Civil Aviation Authority. (2025). CAP3195: Working paper on regulatory models. UK Civil Aviation Authority. https://www.caa.co.uk/CAP3195

Civil Aviation Authority. (2026). CAP3251: Heathrow capacity expansion – consultation on a shortlist of regulatory models. UK Civil Aviation Authority. https://www.caa.co.uk/CAP3251

Heathrow Airport Limited. (2026). CAP3195 – CAA Working Paper on Regulatory Models. Response by Heathrow Airport Ltd.

Heathrow Airport Limited. (2026). Megaproject outcome drivers: Benchmarking fact base. Supporting Document 1 to Heathrow Airport Limited's response to CAP3195.

7.2 Section 3 references

Civil Aviation Act 2012. (c.19). London: The National Archives. Available at: https://www.legislation.gov.uk/ukpga/2012/19

Civil Aviation Authority. (2025). CAP3195: Working paper on regulatory models. UK Civil Aviation Authority. https://www.caa.co.uk/CAP3195

Civil Aviation Authority. (2026). CAP3251: Heathrow capacity expansion – consultation on a shortlist of regulatory models. UK Civil Aviation Authority. https://www.caa.co.uk/CAP3251

Civil Aviation Authority. (2023). CAP2593: CAA Consumer Strategy. London: Civil Aviation Authority. Available at: https://www.caa.co.uk/our-work/publications/documents/content/cap2593/

Department for International Development. (2014). Assessing the strength of evidence: How to note. Department for International Development. https://www.gov.uk/government/publications/how-to-note-assessing-the-strength-of-evidence

Heathrow Airport Limited. (2026). CAP3195 – CAA Working Paper on Regulatory Models. Response by Heathrow Airport Ltd. 20 January 2026.

Heathrow Airport Limited. (2026). Megaproject outcome drivers: Benchmarking fact base. Supporting Document 1 to Heathrow Airport Limited's response to CAP3195. 20 January 2026.

HM Treasury. (2026). The Magenta Book: Central Government guidance on evaluation. HM Treasury. https://www.gov.uk/government/publications/the-magenta-book

Lewin, S., Booth, A., Glenton, C., Munthe-Kaas, H., Rashidian, A., Wainwright, M., Bohren, M. A., Tunçalp, Ö., Colvin, C. J., Garside, R., Carlsen, B., Langlois, E. V., & Noyes, J. (2018). Applying GRADE-CERQual to qualitative evidence synthesis findings: Introduction to the series. Implementation Science, 13(Suppl 1), Article 2. https://doi.org/10.1186/s13012-017-0688-3

Nutley, S., Powell, A., & Davies, H. (2013). What counts as good evidence? Alliance for Useful Evidence. Nesta. https://www.nesta.org.uk/report/what-counts-as-good-evidence/

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Royal Society & Academy of Medical Sciences. (2018). Evidence synthesis for policy: A statement of principles. The Royal Society and the Academy of Medical Sciences. https://royalsociety.org/-/media/policy/projects/evidence-synthesis/evidence-synthesis-statement-principles.pdf

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E., & Henry, D. A. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ, 358, j4008. https://doi.org/10.1136/bmj.j4008

Spencer, L., Ritchie, J., Lewis, J., & Dillon, L. (2003). Quality in qualitative evaluation: A framework for assessing research evidence. Cabinet Office. https://assets.publishing.service.gov.uk/media/5a8179c1ed915d74e33fe69e/Quality-in-qualitative-evaulation_tcm6-38739.pdf

Tong, A., Flemming, K., McInnes, E., Oliver, S., & Craig, J. (2012). Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ. BMC Medical Research Methodology, 12, Article 181. https://doi.org/10.1186/1471-2288-12-181

Whiting, P., Savović, J., Higgins, J. P. T., Caldwell, D. M., Reeves, B. C., Shea, B., Davies, P., Kleijnen, J., Churchill, R., & ROBIS Group. (2016). ROBIS: A new tool to assess risk of bias in systematic reviews was developed. Journal of Clinical Epidemiology, 69, 225–234. https://doi.org/10.1016/j.jclinepi.2015.06.005

Wong, G., Greenhalgh, T., Westhorp, G., Buckingham, J., & Pawson, R. (2013). RAMESES publication standards: Realist syntheses. BMC Medicine, 11, Article 21. https://doi.org/10.1186/1741-7015-11-21

7.3 Section 4 references

The references below use standardised publication details for the sources relied upon in this analysis. Where HAL's Megaproject Outcome Drivers document or CAP3251 materials use shortened titles, variant years, secondary descriptions or incomplete bibliographic details, the references have been checked against the identifiable published source and cited in that form. This keeps the evidential link to HAL's material while avoiding the repetition of minor inconsistencies in source descriptions.

7.3.1 Section 4.2 references

Abgeordnetenhaus von Berlin. (2016). Bericht des 1. Untersuchungsausschusses des Abgeordnetenhauses von Berlin, 17. Wahlperiode, zur Aufklärung der Ursachen, Konsequenzen und Verantwortung für die Kosten und Terminüberschreitungen des im

Bau befindlichen Flughafens Berlin Brandenburg Willy Brandt (BER). Drucksache 17/3000. Berlin: Abgeordnetenhaus von Berlin.

Abgeordnetenhaus von Berlin. (2021). Bericht des 2. Untersuchungsausschusses des Abgeordnetenhauses von Berlin, 18. Wahlperiode, zur Aufklärung der Ursachen, Konsequenzen und Verantwortung für die Kosten und Terminüberschreitungen beim Flughafen Berlin Brandenburg "Willy Brandt" (BER), Untersuchung II. Drucksache 18/4010. Berlin: Abgeordnetenhaus von Berlin.

Flyvbjerg, B. (2022). Heuristics for masterbuilders: Fast and frugal ways to become a better project leader. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4159984

Gil, N. and Fu, Y. (2022). Megaproject performance, value creation, and value distribution: An organizational governance perspective. Academy of Management Discoveries, 8(2). doi: 10.5465/amd.2020.0029.

A.I.A. (2007). Integrated project delivery: A guide. American Institute of Architects, California.

Ibrahim, M.W., Hanna, A.S. and Kievet, D. (2020). 'Quantitative comparison of project performance between project delivery systems', Journal of Management in Engineering, 36(6), 04020082. doi: 10.1061/(ASCE)ME.1943-5479.0000837.

Kerkloh, M. (2007). Munich Airport's Terminal 2: A successful airport airline collaboration. Journal of Airport Management, 1(4), pp. 330 to 337.

Kuchinke, B.A. and Sickmann, J. (2007). The joint venture Terminal 2 at Munich Airport and its consequences: an analysis of competition economics. in Fichert, F., Haucap, J. and Rommel, K. (eds) Proceedings of the 4th Conference on Applied Infrastructure Research. Berlin: INFER Research Perspectives, pp. 107 to 133.

Lovallo, D., Cristofaro, M. and Flyvbjerg, B. (2023). Governing large projects: A three-stage process to get it right. Academy of Management Perspectives, 37(2), pp. 138–156. doi: 10.5465/amp.2021.0129.

NASFA, COAA, APPA, AGC and AIA. (2010). Integrated Project Delivery for Public and Private Owners. Washington, DC: Associated General Contractors of America.

Port Authority of New York and New Jersey Airport Advisory Panel. (2017). A Vision Plan for John F. Kennedy International Airport: Recommendations for a 21st Century Airport for the State of New York. New York: Port Authority of New York and New Jersey.

7.3.2 Section 4.3 references

Abgeordnetenhaus von Berlin. (2016). Bericht des 1. Untersuchungsausschusses des Abgeordnetenhauses von Berlin, 17. Wahlperiode, zur Aufklärung der Ursachen, Konsequenzen und Verantwortung für die Kosten und Terminüberschreitungen des im Bau befindlichen Flughafens Berlin Brandenburg Willy Brandt (BER). Drucksache 17/3000. Berlin: Abgeordnetenhaus von Berlin.

Abgeordnetenhaus von Berlin. (2021). Bericht des *2.* Untersuchungsausschusses des Abgeordnetenhauses von Berlin, 18. Wahlperiode, zur Aufklärung der Ursachen, Konsequenzen und Verantwortung für die Kosten und Terminüberschreitungen beim

Flughafen Berlin Brandenburg “Willy Brandt” (BER), Untersuchung II. Drucksache 18/4010. Berlin: Abgeordnetenhaus von Berlin.

Department for Transport. (2025). Stewart Review: Major Transport Projects Governance and Assurance Review: The HS2 Experience. London: Department for Transport.

Gil, N. and Fu, Y. (2022). Megaproject performance, value creation, and value distribution: An organizational governance perspective. Academy of Management Discoveries, 8(2), pp. 165 to 189. doi: 10.5465/amd.2020.0029.

Kerkloh, M. (2007). Munich Airport’s Terminal 2: A successful airport airline collaboration. Journal of Airport Management, 1(4), pp. 330 to 337.

Kuchinke, B.A. and Sickmann, J. (2007). The joint venture Terminal 2 at Munich Airport and its consequences: an analysis of competition economics. in Fichert, F., Haucap, J. and Rommel, K. (eds) Proceedings of the 4th Conference on Applied Infrastructure Research. Berlin: INFER Research Perspectives, pp. 107 to 133.

Lovegrove, S. (2026). Review of implications for the Civil Service and wider public sector of findings of the James Stewart Review. London: Cabinet Office. Published 19 May 2026. Available at: https://www.gov.uk/government/publications/lovegrove-review-the-implications-for-the-civil-service-and-wider-public-sector-of-findings-of-the-james-stewart-review

Shenhar, A. and Holzmann, V. (2017). The three secrets of megaproject success: Clear strategic vision, total alignment, and adapting to complexity. Project Management Journal, 48(6), pp. 29 to 46.

7.3.3 Section 4.4 references

Abgeordnetenhaus von Berlin. (2021). Bericht des 2. Untersuchungsausschusses des Abgeordnetenhauses von Berlin, 18. Wahlperiode, zur Aufklärung der Ursachen, Konsequenzen und Verantwortung für die Kosten und Terminüberschreitungen beim Flughafen Berlin Brandenburg “Willy Brandt” (BER), Untersuchung II. Drucksache 18/4010. Berlin: Abgeordnetenhaus von Berlin.

Comisión Mixta para las Relaciones con el Tribunal de Cuentas. (2007). Resolución de 20 de marzo de 2007, en relación al Informe de fiscalización de las principales actuaciones realizadas por AENA en desarrollo del Plan Director del Aeropuerto de Madrid-Barajas, aprobado el 19 de noviembre de 1999 y de su situación a 31 de diciembre de 2004. BOE núm. 113, 11 May.

Cortes Generales. (2007). Diario de Sesiones de las Cortes Generales, Comisiones Mixtas, Comisión Mixta para las Relaciones con el Tribunal de Cuentas. VIII Legislatura, Núm. 85, sesión núm. 24, 20 February 2007.

El Asmar, M. and Gibson, G.E. Jr. (2018). The cost impact of front end engineering design (FEED) accuracy for large industrial projects. in Abdul-Malak, M., Khoury, H., Singh, A. and Yazdani, S. (eds) Responsible Design and Delivery of the Constructed Project. Fargo, ND: ISEC Press, CPM-06, pp. 1 to 6.

Flyvbjerg, B. and Gardner, D. (2023). How Big Things Get Done: The Surprising Factors that Determine the Fate of Every Project, from Home Renovations to Space Exploration and Everything in Between. London: Macmillan.

Msulwa, R. (2022). How do megaprojects influence institutional change?. Oxford Review of Economic Policy, 38(2), pp. 302 to 321. doi: 10.1093/oxrep/grac005.

Network Rail. (2024). Delivering an Efficient Railway: Our National Efficiency Plans 2024 to 2029. London: Network Rail.

7.3.4 Section 4.5 references

Department for Transport. (2025). Stewart Review: Major Transport Projects Governance and Assurance Review: The HS2 Experience. London: Department for Transport.

Flyvbjerg, B., Holm, M.K.S. and Buhl, S.L. (2004). What causes cost overrun in transport infrastructure projects?. Transport Reviews, 24(1), pp. 3 to 18.

Flyvbjerg, B. (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), pp. 344 to 367.

Independent Water Commission. (2025). Independent Water Commission: Final Report. London: Independent Water Commission.

Ofgem. (2022). Decision on Accelerating Onshore Electricity Transmission Investment. London: Ofgem.

Ofwat. (2022). Reasons for Amending Tideway's Project Licence. Birmingham: Ofwat.

World Bank. (2024). Drivers of Delays in Procurement of Infrastructure Projects. Washington, DC: World Bank.

7.3.5 Section 4.6 references

Budzier, A. (2026). Read Two Books: Organisational Design and the Integration Failure at the Heart of Megaproject Underperformance. Working paper. Oxford: Saïd Business School, University of Oxford.

Department for Business, Energy and Industrial Strategy. (2020). RAB Model for Nuclear: Government Response to the Consultation on a RAB Model for New Nuclear Projects. London: BEIS.

Department for Energy Security and Net Zero. (2025). Sizewell C Gets Green Light with Final Investment Decision. London: DESNZ.

Independent Water Commission. (2025). Independent Water Commission: Final Report. London: Independent Water Commission.

Msulwa, R. (2022). How do megaprojects influence institutional change?. Oxford Review of Economic Policy, 38(2), pp. 302 to 321. doi: 10.1093/oxrep/grac005.

National Audit Office. (2017). Review of the Thames Tideway Tunnel. London: National Audit Office.

Thierie, W. and De Moor, L. (2019). Determinants of bank loan spread in project finance. International Journal of Managing Projects in Business, 12(1), pp. 161 to 186.

7.4 Section 5 references

Construction Industry Institute. (2014). Front End Planning: Your Key to a Successful Project. Austin, TX. Available at: https://www.construction-institute.org/front-end-planning-your-key-to-a-successful-project/

HM Treasury. (2022). The Green Book: Central Government Guidance on Appraisal and Evaluation. London: HM Treasury. Available at: https://www.gov.uk/government/publications/the-green-book-appraisal-and-evaluation-in-central-government

Budzier, A. Franzen, S. Leed, A. Poker, D. (2025). Oxford Global Projects: Heathrow Benchmark Report. Civil Aviation Authority. Available at: https://www.caa.co.uk/commercial-industry/economic-regulation-and-competition-policy/heathrow-airport/capacity-expansion

Smallwood, N. (2020). Setting up for success: The Importance of Front-End Loading. Infrastructure and Projects Authority blog, 9 September. Available at: https://ipa.blog.gov.uk/2020/09/09/setting-up-for-success-the-importance-of-front-end-loading/

8 Annexes

8.1 Section 3 annexes

8.1.1 Source-to-framework table for the methodological quality review

The sources below comprise a mixture of reporting guidelines, methodological appraisal tools and policy evaluation frameworks. They are not applied as a checklist and do not impose academic review standards on HAL's document. Their purpose is to identify the practical safeguards that allow a reader to assess whether an evidence synthesis is transparent, balanced and proportionate to the conclusions drawn.

Source	Relevant principle	Risk mitigated	Relevance to HAL's Megaproject Outcome Drivers document	Resulting assessment theme
HM Treasury, The Magenta Book: Central Government guidance on evaluation, 2026	*Policy evaluation framework*. Government evaluation should be proportionate, transparent and useful for accountability, learning and decision-making. The Magenta Book also emphasises openness, preservation of evaluation materials, and clear reporting of findings.	Black-box analysis, lack of reproducibility, decision-maker overconfidence.	Although HAL's document is not a government evaluation, the Magenta Book provides a relevant UK public-sector reference point for proportionate transparency where evidence informs decisions affecting public and consumer interests.	Analytical transparency and proportionality.
DFID, Assessing the Strength of Evidence: How to Note, 2014	*Policy evidence guidance*. The strength of a body of evidence depends on quality, size, context and consistency. The guidance also recognises that	Selective presentation of evidence (cherry-picking), over-reliance on isolated examples, confirmation	HAL relies on a combination of academic studies, practitioner evidence and case studies. The CAA needs to know whether the overall body of evidence is sufficiently	Body-of-evidence strength and consistency.

	different research designs are suitable for different questions.	bias, weak body-of-evidence assessment.	large, consistent, objectively selected, high-quality and contextually relevant.	
PRISMA 2020 Statement	*Reporting guideline*. PRISMA 2020 supports transparent reporting of why a review was done, what methods were used, how studies were identified, selected, appraised and synthesised, and what was found. PRISMA is a reporting guideline, not a tool for assessing methodological quality.	Incomplete reporting, inability to understand search and selection, lack of reproducibility.	PRISMA should not be applied as a formal compliance checklist to HAL's report. It is still relevant because it identifies the basic information a reader needs to assess how an evidence base was assembled.	Reporting transparency.
ENTREQ, Tong et al., Enhancing transparency in reporting the synthesis of qualitative research, 2012	*Reporting framework for qualitative evidence synthesis*. ENTREQ identifies the reporting stages commonly associated with qualitative evidence synthesis, including searching and selecting qualitative research, quality appraisal and methods for synthesising qualitative findings.	Opaque qualitative synthesis, selective case-study interpretation, unclear appraisal of qualitative evidence.	The Megaproject Outcome Drivers document relies heavily on interpretive synthesis of literature, policy precedent and case studies. ENTREQ is relevant because those elements require transparent selection, appraisal and synthesis.	Qualitative and case-study synthesis transparency.
Spencer et al., Quality in Qualitative Evaluation, 2003	*Qualitative evaluation framework*. Qualitative evaluation should be contributory, defensible in design, rigorous in conduct and credible in claim. The framework assesses the credibility, rigour and relevance of qualitative evidence.	Weak-evidence risk, overclaiming, poor distinction between evidence and interpretation.	Several of HAL's conclusions depend on case-study interpretation and qualitative judgements about governance, stakeholder alignment and accountability. These judgements need to be supported by a defensible analytical route.	Qualitative rigour and credibility of claims.

GRADE-CERQual, Lewin et al., Applying GRADE-CERQual to qualitative evidence synthesis findings, 2018	*Confidence assessment framework for qualitative synthesis.* Confidence in qualitative evidence synthesis findings is assessed through methodological limitations, coherence, adequacy of data and relevance.	Transferability risk, false analogy, overconfidence in interpretive findings.	HAL draws lessons from other sectors and airport cases. CERQual provides a useful lens for asking whether those lessons are coherent, adequately supported and relevant to Heathrow.	Relevance, transferability and confidence.
ROBIS, Whiting et al., 2016	*Methodological appraisal tool (designed for systematic reviews).* ROBIS assesses risk of bias in systematic reviews, including study eligibility criteria, study identification and selection, data collection and appraisal, synthesis and findings, and interpretation.	Selection bias, eligibility-framing risk and interpretive bias.	HAL's report does not set out a full review protocol or eligibility framework. ROBIS supports asking whether bias could have entered through the review process itself, particularly through source selection, synthesis and interpretation.	Process integrity and selection bias.
AMSTAR 2, Shea et al., 2017	*Methodological appraisal tool (designed for systematic reviews).* AMSTAR 2 appraises systematic reviews that include randomised and non-randomised studies. It is not designed to generate an overall score; critical weaknesses may be hidden by aggregate ratings.	Superficial scoring, failure to identify critical weaknesses, inadequate search or risk-of-bias assessment.	AMSTAR 2 should not be applied mechanically. It is useful because it shows why evidence reviews should not be judged by appearance or volume of citations alone, but by whether critical methodological safeguards are present.	Critical-domain appraisal.
Royal Society and Academy of Medical Sciences, Evidence synthesis for policy:	*Policy evidence synthesis principles.* Good evidence synthesis for policy should be inclusive, rigorous, accessible and transparent. It should use the most comprehensive feasible body of	Framing bias, selection bias, publication bias, black-box evidence, inadequate uncertainty	This is directly relevant because the Megaproject Outcome Drivers document is a policy-facing evidence synthesis prepared for a contested regulatory process. It supports the proposition that	Policy-grade evidence synthesis; bias and uncertainty management.

a statement of principles, 2018	evidence, recognise and minimise bias, disclose methods, limitations and quality assurance, and communicate uncertainty and areas of contention.	disclosure and decision-maker overconfidence.	the CAA should not treat a synthesis as neutral simply because it cites real sources; it must be able to see how bias, limitations and uncertainty have been managed.	
RAMESES, Wong et al., realist synthesis standards, 2013	*Supporting standards*. Realist synthesis reports and tests explanations of how context, mechanisms and outcomes interact in complex interventions.	False analogy, context-blind policy transfer, unsupported causal claims.	RAMESES is supporting rather than core. It is useful where HAL transfers lessons from one sector or project context to Heathrow without clearly explaining the causal mechanism that would make the precedent applicable.	Causal mechanism and transferability.
Nutley, Powell and Davies, What Counts as Good Evidence? 2013	*Policy evaluation framework*. What counts as good evidence depends on the question being asked, the purpose of the decision and the context in which evidence is used.	Rigid or inappropriate evidence hierarchies, decision-maker overconfidence, poor fit between evidence type and decision question.	This is useful to keep the critique proportionate. It supports assessing whether HAL's evidence is fit for the regulatory question, rather than judging it by a rigid academic hierarchy.	Fitness for decision purpose.

8.1.2 Application of the methodological framework to the Megaproject Outcome Drivers document

Annex B draws on the principles established in Annex A to construct a structured assessment framework. Each row corresponds to one of the seven themes identified in Annex A — clarity of framing, transparency of selection, evidence quality, balance, transferability, proportionality and separation of evidence from interpretation — together with an overall reliability assessment. For each theme, the table sets out what a reliable evidence base would normally demonstrate, how the Megaproject Outcome Drivers document compares against that standard, the bias or reliability risk that follows, and the practical implication for the CAA.

Assessment theme	What a reliable evidence base would normally show	Initial application to the Megaproject Outcome Drivers document	Bias or reliability risk	Implication for the CAA
Clarity and neutrality of framing	A clear review question, an explanation of why the analytical framework was selected, and recognition of alternative ways of framing the issue.	HAL frames megaproject success around financeability, timeliness and quality end-result, overlaid with structure, governance and incentives. Those criteria are relevant, but the document does not explain why this framework is complete or neutral for the CAA's wider regulatory assessment.	Framing bias. The review may start from criteria that naturally support HAL's preferred emphasis on integrated control, delivery speed and financeability.	The CAA should not assume that HAL's framework is the appropriate regulatory framework. It should test the conclusions against its own statutory duties and evaluation criteria, including cost, efficiency, competition, affordability and user outcomes.
Transparency of evidence identification and selection	A reproducible account of how literature, reports and case studies were identified, selected and excluded.	The Megaproject Outcome Drivers document draws on megaproject literature, airport cases and UK infrastructure cases, and says examples shown are illustrative and not exhaustive. It does not provide search terms, databases, inclusion	Selection bias and selective evidence risk (cherry-picking). The reader cannot tell whether the evidence base is representative or curated.	The CAA should treat the evidence base as illustrative rather than representative unless HAL provides a fuller selection methodology.

		criteria, exclusion criteria or a log of excluded evidence.		
Appraisal of evidence quality and limitations	An assessment of the quality, limitations and evidential weight of each type of source.	The report relies on different evidence types, including academic studies, practitioner analysis, small-sample quantitative findings, case studies and policy precedents. It does not provide a systematic appraisal of source quality or limitations.	Weak-evidence risk. Sources of materially different reliability may be treated as if they provide equivalent support.	The CAA should distinguish evidence that identifies a plausible issue from evidence that is strong enough to support regulatory model conclusions.
Balance and treatment of mixed or contrary evidence	Evidence that supports, qualifies and contradicts each major finding should be reported and interpreted.	The document presents a set of key findings on structure, governance and incentives, but does not consistently show whether the underlying evidence is supportive, mixed, neutral or contrary.	Confirmation bias. Valid evidence may be presented one-sidedly, giving the impression of stronger support than exists.	HAL's conclusions should not be treated as established without independent testing of the case-study evidence on which they depend.
Relevance and transferability	The report should explain why evidence from one sector, country, ownership structure or project type transfers to Heathrow.	HAL uses lessons from airports and wider UK infrastructure precedents, including different regulated sectors and delivery models. The applicability of these examples to Heathrow's privately financed airport expansion is not always made explicit.	False analogy and transferability risk. A precedent may be real but only partly applicable.	The CAA should require a clear link between the precedent, the causal mechanism and the Heathrow context before treating the case as evidence for or against a regulatory model.

Proportionality of conclusions to evidence strength	Conclusions should be calibrated to the quality, consistency, adequacy and relevance of the evidence.	In some areas the report moves from evidence that an issue can create risk to stronger implications about which governance or regulatory models are preferable.	Overclaiming. Partial or context-specific evidence may be used to support stronger regulatory conclusions than it can bear.	The CAA should distinguish between evidence that identifies a risk to be managed and evidence that justifies rejecting, downgrading or materially constraining a CAP3251 option.
Separation of evidence, interpretation and recommendation	The report should distinguish observed facts, interpretation of causality and regulatory recommendation.	The Megaproject Outcome Drivers document is described as a fact base, but many propositions are interpretive judgements about how ownership, governance, incentives and financing affect outcomes.	Interpretation-as-fact risk. Judgement may be presented with the authority of factual reporting.	The CAA should treat HAL's conclusions as stakeholder interpretation unless the underlying evidence and causal reasoning are independently demonstrated.
Overall reliability for decision-making	The report should provide enough transparency for the regulator to place weight on its conclusions and to distinguish evidence, interpretation and uncertainty.	The document raises relevant issues but does not provide enough methodological transparency to establish that its conclusions are neutral, representative or sufficiently tested.	Decision-maker overconfidence. The report may look more objective and comprehensive than the reporting allows the reader to verify.	The CAA should treat the Megaproject Outcome Drivers document as a relevant stakeholder evidence submission, but not as decision-grade evidence for narrowing, rejecting or materially constraining regulatory model options unless the material conclusions are independently and more comprehensively tested.

8.2 Section 5 annexes

8.2.1 Multivariate regression

This annex reports the statistical checks used to interpret the pooled reference class. The tests are not used mechanically to decide whether pooling is permitted. They identify which dimensions materially affect overrun outcomes and therefore need to be considered when interpreting the reference class.

An OLS regression of overrun is estimated against four dimensions: project sub-category, with terminal upgrade or expansion as the baseline; region, with Europe as the baseline; project scale, measured as the log of added maximum annual passenger capacity; and decision year.

8.2.1.1 Cost-overrun pooling check

Sample restricted to projects with non-missing scale and year. n = 29, R^2 = 55%.

Cost — pooling regression coefficients	Coef. (pp)	p-value	95% CI
Sub-category: New terminal	51.9**	0.050	[0.0, 103.8]
Sub-category: New airport	118.2**	0.013	[27.5, 209.0]
Region: North America	-11.1	0.650	[-61.7, 39.4]
Region: Asia-Pacific	-55.4	0.134	[-129.4, 18.7]
Region: Middle East	24.6	0.809	[-185.8, 235.0]
Region: Africa	-50.9	0.131	[-118.4, 16.6]
Region: South America	-19.9	0.858	[-249.4, 209.5]
Max passenger capacity (log scale)	-13.4	0.169	[-32.9, 6.2]
Decision year (centred)	2.0	0.343	[-2.3, 6.3]

Reading: scope is the only statistically significant finding. New-airport projects sit about +118 percentage points above the terminal-upgrade baseline (p = 0.013) and new-terminal projects about +52 pp above it (p = 0.050).

8.2.1.2 Schedule-overrun pooling check

Sample restricted to projects with non-missing scale and year. n = 30, R^2 = 59%.

Schedule — pooling regression coefficients	Coef. (pp)	p-value	95% CI
Sub-category: New terminal	16.9	0.260	[-13.4, 47.2]
Sub-category: New airport	42.8	0.210	[-26.0, 111.6]
Region: North America	1.0	0.958	[-37.9, 39.9]
Region: Asia-Pacific	-13.9	0.583	[-65.5, 37.7]
Region: Middle East	58.3	0.211	[-14.9, 101.8]
Region: Africa	72.8	0.161	[-31.5, 177.0]
Max passenger capacity (log scale)	-2.7	0.795	[-24.0, 18.6]
Decision year (centred)	0.8	0.348	[-0.9, 2.6]

Reading: no tested factor reaches conventional significance. The direction of the sub-category coefficients is similar to cost, but the differences are not statistically robust in this sample.