Is Complexity Theory Right that Bigger Is Riskier? Evidence from Information Technology[1] [2]

Fioralba Ajazi, Daniel Nickelsen, Jens Schmidt, Bent Flyvbjerg, Maria Christodoulou

Preprint

Abstract

A key thesis in complexity theory is that larger ventures entail higher risk. We test the thesis for information technology (IT) cost and schedule risk. We find that, contrary to common belief, project size is not a reliable predictor of risk. The common assumption that smaller projects have lower risk should be rejected based on our findings. In fact, the smallest IT projects have the largest risk. This finding is significant, because it challenges a central tenet of complexity theory, and it shows project management practitioners that small is not necessarily safe.

Keywords: Complexity Theory; IT Projects; Project Management; Project Performance; Project Risk; Project Size; Cost Risk; Schedule Risk; IT Risk Management; Fat Tails.

[1] Fioralba Ajazi did the literature review, first hypotheses, analysis, and the bulk of writing. Jens Schmidt contributed hypotheses and rewrites. Bent Flyvbjerg contributed the idea for the paper, the data, supervision, and rewrites. Daniel Nickelsen and Maria Christodoulou did statistical analysis and rewrites.

[2] The authors wish to thank the Villum Foundation (Villum grant: 39735) and the IT University of Copenhagen for financial support.

1. Introduction

It is a key thesis in complexity theory that larger ventures entail higher risk. The Oxford English Dictionary defines complexity as "consisting of many and connected parts" (Oxford University Press, n.d.), with more parts and more connections making for more complexity, other things being equal. Parts that may drive complexity include people, technology, organizations, markets, politics, and output. Non-linear amplification between parts drives risk, according to the theory (West, 2017).

Large IT projects typically have more parts than smaller ones. They therefore carry more risk according to academics and practitioners (Zmud, 1980; McFarlan, 1981; Martin et al., 2005; Sauer et al., 2007; Peters and Verhoef, 2008; Iriarte and Bayona, 2020). Larger projects supposedly face higher uncertainty in delivering to expected cost, schedule, and benefits (McFarlan, 1981). Moreover, larger projects introduce greater complexity, higher task interdependencies, and increased volatility that negatively impact performance (Yetton et al., 2000; McFarlan, 1981; Martin et al., 2005; Jones, 2007; Peters and Verhoef, 2008; Gemino et al., 2007) according to common thinking.

In this paper, we test the common assumption of positive correlation between project size and risk. Specifically, we analyse the relationship between IT project size and IT project risk. We measure project size as estimated project cost (McFarlan, 1981, Hill et al., 2000; Shenhar, 2001; Yang et al., 2006; Morcov et al., 2021; Berg and Ritschel, 2023) at the time of the final investment decision (FID), and we measure risk as (a) cost risk (actual divided by estimated cost) and (b) schedule risk (actual divided by estimated duration), respectively. Specifically, we test the following two hypotheses:

H1: Increasing project size is associated with increasing cost risk.

H2: Increasing project size is associated with increasing schedule risk.

To test the two hypotheses, we compare the probability distributions of cost and schedule overrun, respectively, across different project sizes. We compare mean, median, variance, and tail parameters of the distributions to test whether cost and schedule risk increase with project size. We follow a conventional two-step approach: First, we evaluate resemblance across size groups. Second, we assess the direction (positive or negative) of the relationship between size and overrun.

We test H1 based on a sample of 5,094 IT projects and H2 based on a sample of 831 projects, using all available data in both cases. The data cover several industries and geographies, and multiple IT project types, including software development, BI implementation, ERP, and cybersecurity. The data are presented in detail in Section 3.

2. Background

Here we review the literature on the relationship between IT project size and performance. Appendix 1 provides our literature search criteria and Appendix 2 includes summaries of selected papers (Table A2.1 and Table A2.2 in Appendix 2). Most of the literature relies on survey-based data, self-reported after the fact by project managers, entailing well-known problems of validity in terms of subjective memory, availability bias, and other cognitive biases. Budget and schedule overrun are often measured as ordinal variables (i.e., prefixed intervals), which limits the level of detail for performance evaluation.

The literature varies in how project size is measured and in the methodology used to study the relationship between size and performance. Appendix 2 accounts for specific differences between the reviewed studies (Tables 5 and 6 in Appendix 2). The method for measuring

project size may influence the observed relationship between size and performance (Jørgensen et al., 2012). In the literature, project size is measured by (a) budget, (b) duration, (c) team size, (d) function points (size of software components), and/or (e) project complexity (number of interconnected systems, number of contractors involved, type of contracts, requirements creep, and time compression). We note that some of these size measures are interdependent. For example, team size, function points, and requirements creep affect cost.

McFarlan (1981, p. 143) finds that larger projects (measured by project cost, staffing levels, elapsed time, and number of organizational units affected) have higher risk (measured as cost overrun, schedule overrun, benefit shortfall, and shortfall in technical performance) than smaller projects. Project size and complexity are often considered together (McFarlan, 1981; Hill et al., 2000; Shenhar, 2001; Yang et al., 2006; Morcov et al., 2021; Berg and Ritschel, 2023). According to Morcov et al. (2021), for example, "In practice, size cannot be separated from complexity; it is strongly related to uncertainty (Zmud, 1980) and risk exposure (Wallace et al., 2004; Chiu and Wang, 2008)." Martin et al. (2005), however, treat size and complexity as separate variables. They define size by project cost, number of vendors, number of connected systems, team size, and project duration, while they define complexity by new technology adoption and expert availability. They show that as size increases, the likelihood of cost overrun increases. However, they did not find the relationship between size and schedule overrun to be statistically significant. Neither was the relationship between complexity and project performance (measured as system quality, information quality, system use, user satisfaction, individual impact and organizational impact). Jones (2007) measures software project size by function points, a metric for the amount of functionality in a software system. Peters and Verhoef (2008, pp. 21, 55) associate project size with the risk of cost and schedule overrun influenced by factors such as time compression and requirements creep. Jones (1995, p. 86) associates

size in function point with schedule overrun. Jørgensen (2019) finds a positive correlation between size and cost and schedule overrun in his study of agile methods in software development. Berg and Ritschel (2023) present similar conclusions about agile software development. These studies conclude that there is a negative relationship between size and performance. Dongus et al. (2015) found that project size (measured by team size) is only weakly correlated with project outcome (measured as "success", "performance", and "satisfaction"). Sauer et al. (2007) found a significant negative relationship between size (measured as team size, duration, and effort) and project performance. But Sauer et al. (2007) also found that "one-quarter of the projects underperformed, however small their size" and concluded that "there is a significant level of risk regardless of size."

In sum, the reviewed studies generally assume that large IT projects are more complex and more risky than smaller ones, but also that the relationship between size and performance is not straightforward. In this paper, we present a study based on observed (as opposed to self-reported) data documenting how project size (measured as estimated cost) relates to project performance (measured as cost and schedule overrun).

3. Data and methods

To study the relationship between size and overrun, we collected data on estimated and actual project cost. *Estimated cost* is measured at the time of the final investment decision (FID) and serves as a proxy for project size. *Actual cost* is measured at the date of going live. For each project, we converted both estimated and actual cost to USD and adjusted for inflation. *Cost overrun* is defined as actual cost divided by estimated cost. *Schedule overrun* is defined as actual duration divided by estimated duration.

We extracted the data from a database created by Flyvbjerg and colleagues, the largest of its kind and used in numerous studies over the past 20 years (Flyvbjerg et al., 2025). These data have not been used before for studying the relationship between size and performance, which is the original contribution of the present paper. Specifically, the dataset we used to analyse project size and cost overrun includes 5,094 IT projects, ranging in size from a few hundred to 8 billion USD in 2025 prices. The projects come from the public and private sectors in 64 countries on six continents and include a wide variety of IT project types as presented in Flyvbjerg et al. (2022). Information on project duration and schedule overrun was available for 831 IT projects, with duration ranging from a few weeks to 10 years. The projects studied were conducted from 1988 to 2017.

To simplify interpretation and make comparisons across project sizes more robust, we rank projects by size and divide them into five equally sized groups (quintiles). For readability, these size groups are labelled as very small (0-20%), small (20-40%), medium (40-60%), large (60-80%) and very large (80-100%). As robustness checks, we repeated the analyses for other numbers of size groups, including tertiles, quartiles, sextiles and septiles (Appendices 3 and 4).

In addition to project size, our data include contextual variables that may also influence overrun, namely *country*, *year of FID*, *project subtype*, and *region*. Because projects sharing such characteristics may be more similar to each other than to other projects, we use a multi-level (also known as hierarchical or mixed) statistical framework that accounts for these contextual influences when estimating the relationship between project size and overruns. In this way, the context is captured by the models, and we can separate more clearly the relationship between overruns and project size groups, thereby reducing the risk of confounding by underlying variables. For an introduction to multilevel modelling in the context of project management, see Ansar et al. (2014).

Specifically, to estimate how project size relates to project performance, we use regression models with project size groups as predictors and cost or schedule overrun as outcome. Using size groups as categorical predictors rather than size as a continuous variable allows for modelling possible non-linear relationships between size and overrun. This approach also makes the analysis robust with respect to extreme values.

Standard linear regression assumes normally distributed residuals with constant variance, assumptions that are not met by our data. We therefore use generalized linear mixed models (GLM models), which extend linear regression in three relevant ways that fit our data: (a) they allow variance to change with the mean, (b) they accommodate non-normal outcome distributions such as the positively skewed and fat-tailed overrun data observed here, and (c) they allow the contextual variables to be incorporated directly into the model.

A GLM model can be configured according to (a) which contextual variables are included, (b) whether their influence is represented as an overall pattern or as variation across contexts, (c) whether overrun variability is allowed to differ across size groups, and (d) which statistical distribution is used to model overrun. These possible configurations form a set of candidate models. We score each candidate model by a goodness-of-fit measure (Bayesian Information Criterion, BIC), which balances model fit against model complexity, favouring models that explain the data well without unnecessary parameters. This data-driven model selection approach helps to minimize residual confounding in order to better separate the effect of project size from contextual influences, while reducing the risk of overfitting, which strengthens our hypothesis testing. Full details of the GLM models and the model selection procedure are provided in Appendix 3.

To specifically examine the risk of large overruns, we fit tail distributions to the positive overrun tail (above a cut-off, $x_{\min} > 1$) in each size group. Because the observed upper tails

contain extreme overruns and visually resemble power-law behaviour (see the figures in Section 4), and because prior evidence shows that IT cost overruns can follow such behaviour (Flyvbjerg et al., 2022), we consider the Pareto Type I distribution (see Equation (A3.2) in Appendix 3) as our primary candidate model. We therefore test whether the Pareto Type I distribution is plausible for each size group and compare tail risk categories across project sizes using the power-law exponent α (see, for example, Table 3). For the Pareto I fits, we follow the widely used procedure of Clauset et al. (2009) in five steps: (1) estimating the scaling parameter α, (2) estimating the lower bound of power-law behaviour $x_{\min}$, (3) examining the uncertainty in the estimates of α and $x_{\min}$ by non-parametric bootstrapping (repeatedly refitting the model to resampled data), (4) testing the power-law hypothesis by semi-parametric bootstrapping, and (5) comparing the power-law model to alternative tail models (Flyvbjerg et al., 2022), including the exponential, Weibull, log-normal, and log-Weibull distribution. Again, we use the BIC to compare goodness of fit across all fitted distributions. This approach provides strong and converging evidence indicating the presence of a power-law distribution in the data (Clauset et al., 2009, Flyvbjerg et al., 2022).

Among the tail models in step (5), the log-Weibull distribution goes beyond the usual alternatives and provides additional insight: (a) the log-Weibull distribution includes Pareto I as a special case, and (b) it can distinguish between lower risk and even higher risk than a power-law.

We note that when relating cost overrun to project size, we relate the *ratio of actual to estimated cost* to *estimated cost*, which means that we, to some degree, relate estimated cost to itself. This can create mathematical coupling, meaning that the choice of size measure -- estimated or actual cost -- may itself induce an apparent negative or positive relationship between cost overrun and size, as noted by Jørgensen et al. (2012). As a robustness check, we therefore

carry out an additional analysis that removes this effect (see Appendix 3) and find that mathematical coupling does not affect our findings and conclusions.

More details of our methodology and results are provided in Appendices 3 and 4.

To ensure high data quality, we excluded self-reported survey data to avoid recall inaccuracies, response bias, and other forms of bias related to self-reporting. Instead, we relied on high-quality, written records of completed IT projects to maintain consistency in the dataset (Du et al., 2019). This may introduce selection bias, since organizations able to provide detailed, written data tend to be more mature with better performance, other things being equal.

4. Results

In this section, we present our findings on the relationship between project size (estimated cost, budget) and performance. In Section 4.1, we test hypothesis H1 for the relationship between size and cost overrun. In Section 4.2, we similarly test hypothesis H2 for the relationship between size and schedule overrun.

4.1. Relationship between size and cost overrun

Figure 1 shows a log-log scatterplot of project size and cost overrun. Projects are grouped into the five size groups (quintiles) mentioned above, separated by the vertical dashed lines. Eyeballing the plot, we see no clear trend in the relationship between project size and cost overrun.

Figure 1. Scatter plot of project size and cost overrun (log-log). The vertical dashed lines separate the five size groups (quintiles), each containing 20% of projects: very small (0-20%), small (20-40%), medium (40-60%), large (60-80%), and very large (80-100%).

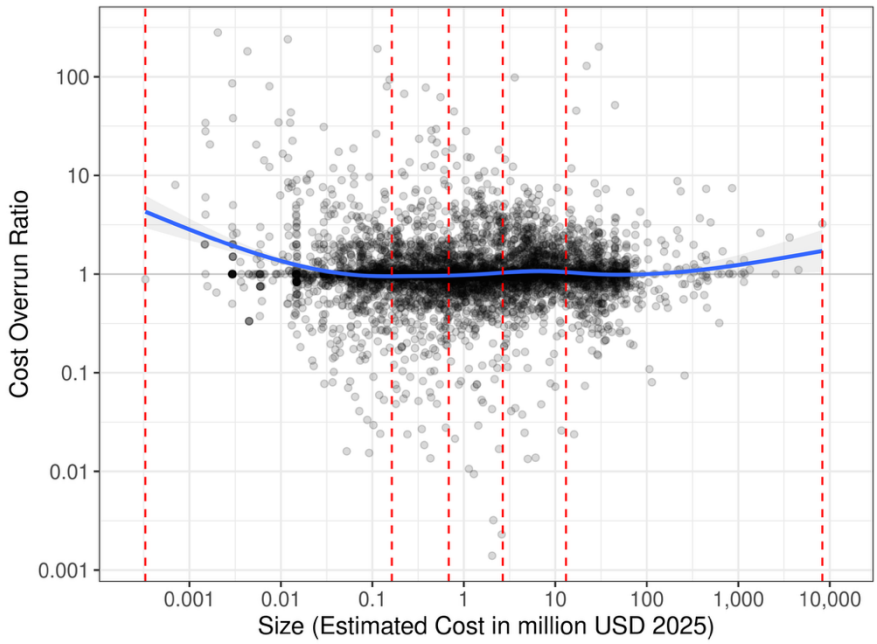


Table 1 presents the descriptive statistics of cost overrun by size groups. The mean and median show that cost overrun in very small projects is strongly right-tailed, which indicates high cost risk. This observation contradicts the common assumption that small projects are less risky than larger ones.

Table 1. Descriptive statistics of cost overrun by project size groups. Overrun is measured as actual divided by estimated cost.

Project Size	***N***	**Mean overrun**	**Median overrun**	**Share of projects with overrun > 1.5**
Very Small (0 – 20%)	1,019	2.92	1.00	21.6%
Small (20 – 40%)	1,019	1.53	0.98	16.4%
Medium (40 – 60%)	1,018	1.42	0.97	17.9%
Large (60% – 80%)	1,019	1.42	1.00	23.7%
Very Large (80% – 100%)	1,019	1.57	1.00	14.3%
Total IT	5,094	1.77	1.00	18.8%

Specifically, Table 1 shows that the largest mean cost overrun is found for very small projects, at 2.92 (192% overrun). The second-largest overrun is found for very large projects, at 1.57 (57% overrun). The median is close to 1 for all size groups, indicating that actual costs are centered near the original budget. The large differences between mean and median indicate that the mean is pulled up by a relatively small number of very large overruns in the right tail of the distribution. In the rest of this section, we test these descriptive patterns statistically.

First, we use non-parametric tests suited to skewed and fat-tailed data in order to examine the subset of projects that finished over budget. Specifically, we test whether cost overrun varies systematically with project size and, if so, whether the trend is positive or negative. The Kruskal-Wallis test indicates with strong statistical significance that differences in cost overrun between size groups are present ($p < 0.001$). The Jonckheere-Terpstra test further provides strong evidence of an overall negative trend ($p = 0.002$), that is, decreasing cost overruns with increasing project size. Dunn's pairwise tests between size groups confirm this pattern at similar significance levels. This finding constitutes the first statistical evidence against hypothesis H1 that increasing project size is associated with increasing cost risk, challenging the common assumption that larger projects are riskier than smaller ones.

Second, we use generalized linear mixed models (GLM models, see Section 3) to test whether the same negative trend holds for typical cost overrun when all projects are considered, and possible confounders (*Country* and *Year of FID*) are taken into account. When we control for these two potential confounders, we find no support for the common belief that very large projects have higher cost overrun than other size groups. In fact, cost performance for very small projects is worse than for any other size group ($p < 0.001$).

Table 2 compares project size groups. We see that the typical cost overruns of small, medium, large, and very large projects are all smaller than the typical overrun of very small projects. This result is highly statistically significant ($p < 0.001$). We further note a consistently decreasing trend with increasing project size, partly substantiated by non-overlapping confidence intervals. This finding further strengthens the evidence against hypothesis H1, even when accounting for contextual factors. In Appendix 3, we provide the full results and more detail of using GLM models to examine the expected cost overruns for projects of the different size groups, including a comprehensive robustness analysis, with additional support for our findings.

Table 2. Comparison of typical cost overrun for different project sizes adjusted for contextual factors (country and year of final investment decision). The second column shows how the size groups differ from very small projects with respect to typical overrun. The p-values show that differences in cost overrun between very small projects and other project sizes are highly statistically significant.

Comparison of cost overrun between very small project and other size groups			
Project size groups	**Typical overrun compared to very small projects**	**Confidence Interval** (at 99% level)	**p-value**
Very Small	× 1.00	0.87 – 1.15	–
Small	× 0.80	0.71 – 0.90	**< 0.001**
Medium	× 0.78	0.69 – 0.88	**< 0.001**
Large	× 0.72	0.64 – 0.80	**< 0.001**
Very Large	× 0.61	0.55 – 0.68	**< 0.001**

Third, we analyze the upper tail of the cost overrun distribution to examine cost risk for each size group. We test several tail models for plausibility for each size group using the procedure by Clauset et al. (2009) described in Section 3. We find that the Pareto I distribution is the best fit for all project size groups, i.e., it is the best suited distribution for our analysis (see Section 3 and Appendix 3).

Table 3 shows the estimated tail parameter, α, of the Pareto I for each size group. We see that very small projects have the lowest α-value, at 0.874, i.e., they are the riskiest in terms of tail risk. Distributions with $\alpha \leq 1$ theoretically have infinite mean and variance, which is the most extreme form of risk that exists for the Pareto I, called *wild risk* by Mandelbrot (1997), who pioneered the study of fat-tailed risk. Infinite mean and variance translates into infinite and thus unpredictable risk. The empirical and modeled tail risks are depicted in Figure 2.

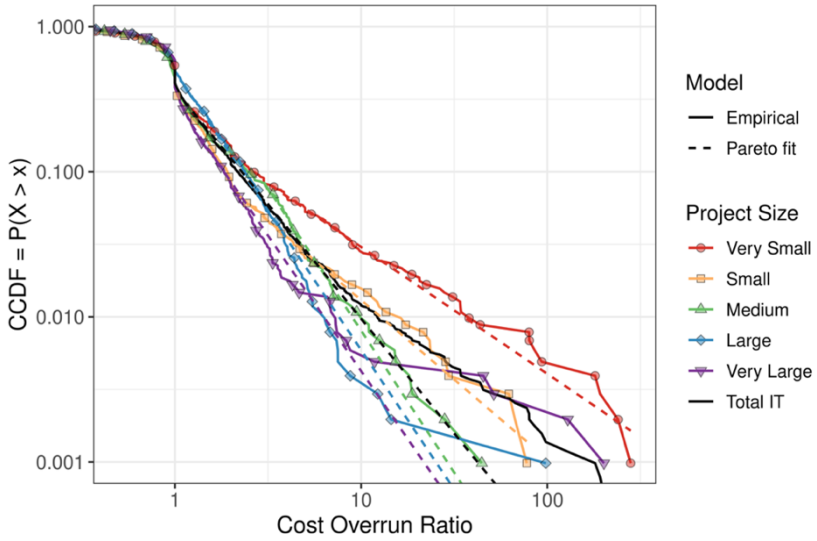
All other size groups have α values between 1 and 2, which implies finite mean but infinite variance, still a very high level of risk, although less extreme than for very small projects. For $1 < \alpha < 2$, the Law of Large Numbers states that the sample mean will converge to the population mean over time (or with increasing sample size), but convergence will likely be too slow (or require too large sample sizes) to be practical in real-world situations. Therefore, also in this category, risk remains unpredictable for practical purposes.

Across all size groups, the data are consistent with power-law tails (or, according to the log-Weibull tail classification, even higher-risk tails), which provide the best fit to the observed upper tail. Details and robustness checks are reported in Appendix 3.

Table 1. Tail risk of cost overrun by project size group. Very small projects have the lowest α, i.e., the highest risk of extreme overrun. Confidence intervals (CI) for α are based on 5,000 bootstrap samples.

Project Size Group	$x_{\min}$	α	99% CI for α
Very Small	3.273	0.874	0.693 – 1.319
Small	2.224	1.095	0.860 – 1.956
Medium	3.025	1.956	1.296 – 2.875
Large	1.105	1.922	1.710 – 2.888
Very Large	1.229	1.866	1.499 – 2.397
Total IT	1.113	1.579	0.959 – 1.690

Figure 2. Empirical complementary cumulative distribution functions (CCDFs) of cost overrun by project size group and for all IT projects combined, shown as a log-log plot so that Pareto tails appear as straight lines. Dashed lines show the fitted Pareto tails from Table 3. At each value on the horizontal axis, the CCDF gives the probability of observing cost overruns larger than x P(X > x).



In summary, we do not find evidence for hypothesis (H1) that increasing project size is associated with increasing cost risk. In contrast to conventional wisdom, we find that smaller projects have significantly higher cost risk than larger projects. Based on the tail analyses, we conclude that all project sizes are very high risk in terms of cost overrun. Moreover, very small projects, often assumed to be low risk, are in fact riskier than all other project sizes.

4.2. Relationship between size and schedule overrun

In this section, we study the relationship between project size and schedule performance.

Figure 3 shows a log-log scatterplot of schedule overrun and size. The vertical dashed lines separate the five size groups of very small, small, medium, large and very large projects. The plot shows no clear trend in the relationship between project size and schedule overrun.

Figure 3. Scatter plot of project size and schedule overrun (log-log). The vertical dashed lines separate the five size groups (quintiles), each containing 20% of projects: very small (0-20%), small (20-40%), medium (40-60%), large (60-80%), and very large (80-100%).

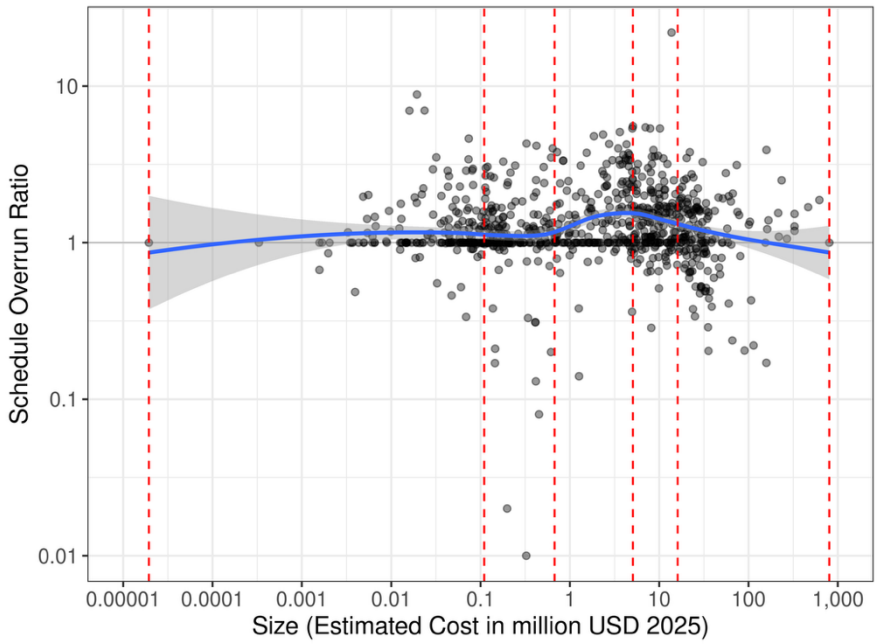


Table 4 presents the descriptive statistics of schedule overrun by project size group. The descriptive statistics show generally high schedule overruns across all size groups, but no clear pattern of variation with size. This does not support hypothesis H2 that larger projects are associated with higher schedule overrun than smaller projects.

Specifically, medium and large projects have the highest mean schedule overruns, at 75% and 67%, respectively, together with high rates of projects finishing with a delay of more than 50%. By contrast, very large projects have the lowest mean schedule overrun, at 26%.

For delayed projects we tested, non-parametrically, for increasing or decreasing trends in schedule overrun with project size. The Jonckheere-Terpstra test does not indicate a clear overall trend. Pairwise Dunn tests yield two statistically significant differences: medium projects have higher schedule overrun than (1) small projects ($p = 0.002$) and (2) very large projects ($p = 0.01$). These mixed results provide further evidence against hypothesis H2.

Table 4. Descriptive statistics for schedule overrun by project size groups. Overrun is measured as actual divided by estimated project duration.

Project Size	*N*	Mean overrun	Median overrun	Share of projects with overrun > 1.5
Very Small (0 – 20%)	167	1.37	1.00	20.4%
Small (20 – 40%)	166	1.30	1.06	21.7%
Medium (40 – 60%)	166	1.75	1.31	43.4%
Large (60% – 80%)	166	1.67	1.35	44.0%
Very Large (80% – 100%)	166	1.26	1.09	33.1%
Total IT	831	1.47	1.12	32.5%

For a more rigorous analysis, we again turn to generalized linear mixed models (GLM models) to compare typical schedule overrun between project size groups, as explained in Section 3. We consider a large set of candidate GLM models and select the best-fitting model for our analysis. Again, we find no evidence that schedule overrun consistently increases with project size. The evidence rather points partly in the opposite direction of hypothesis H2.

This comparison between project size groups is shown in Table 5. Compared with very small projects, small and very large projects show modest (second column) but statistically significant smaller typical schedule overruns ($p < 0.001$). By contrast, the higher schedule overruns for medium and large projects suggested by the descriptive statistics receive at most limited support ($p \approx 0.05$). Details and robustness checks are reported in Appendix 4.

Table 5. Comparison of typical schedule overrun for different project sizes. The second column shows how the size groups differ from very small projects with respect to typical overrun. The p-values in bold indicate differences in schedule overrun that are statistically significant.

Comparison of schedule overrun between very small project and other size			
Project size groups	**Typical overrun compared to very small projects**	**Confidence Interval** (at 99% level)	**p-value**
Very Small	× 1.00	0.92 – 1.09	–
Small	× 0.83	0.74 – 0.93	**<0.001**
Medium	× 1.09	0.98 – 1.21	0.036
Large	× 1.08	0.97 – 1.20	0.066
Very Large	× 0.89	0.80 – 1.00	**0.007**

Next, we analyze schedule risk for each size group by studying the upper tail of the schedule overrun distribution (overrun ratio greater than 1). We test several tail models for plausibility for each size group using the procedure by Clauset et al. (2009) described in Section 3. We find that the Pareto I distribution is the most adequate tail model for all but medium-sized projects, and conclude that Pareto I is the overall best suited distribution for our analysis (see Section 3 and Appendix 4). Among the size groups for which a Pareto tail remains plausible, very small projects have the highest estimated risk of extreme schedule overrun.

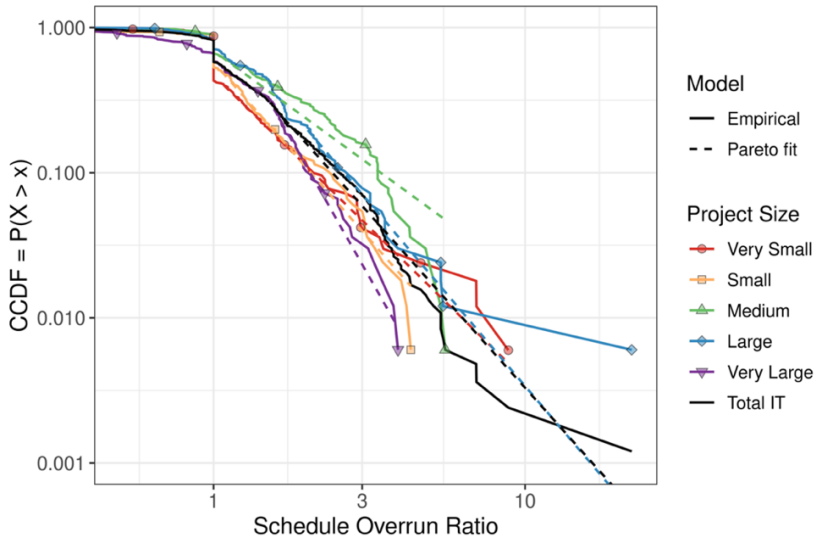
Table 6 and Figure 4 provide the details. Power-law tails are supported by the data for all size groups except medium projects ($p < 0.001$), for which the evidence instead points to a lighter, sub-exponential tail. For very large projects, the evidence for a Pareto tail is weaker, but it remains plausible. Among the remaining size groups, very small projects have the lowest estimated power-law exponent, with $\alpha = 2.097$, indicating the highest estimated tail risk of extreme schedule overrun. Figure 4 is consistent with this interpretation. Medium-sized projects initially show the fattest observed tail, but this pattern is not sustained in the far tail, which likely explains the low fitted alpha over a limited range of overruns rather than a stable power-law tail. Taken together, these results provide further evidence against hypothesis H2,

suggesting that extreme schedule overrun risk rather decreases than increases with project size. Details and robustness checks are reported in Appendix 4.

Table 6. Tail risk of schedule overrun by project size group. Among size groups with a plausible power-law tail, very small projects have the lowest α and therefore the highest estimated risk of extreme schedule overrun. Smaller values of α indicate fatter tails and higher tail risk. Size groups where a Pareto I distribution is not supported by the data ($p < 0.001$) are indicated by an asterisk (*). For each size group, α is estimated from a Pareto I distribution fitted to overruns above the data-driven lower cut-off x_{min}. Confidence intervals (CI) for α are based on 5,000 bootstrap samples.

Project Size Group	x_{min}	α	99% CI for α
Very Small	1.079	2.097	1.550 – 3.265
Small	1.061	2.436	1.970 – 4.828
Medium*	1.057	1.562	1.402 – 7.732
Large	1.909	2.527	1.573 – 4.464
Very Large	1.480	3.780	2.315 – 8.333

Figure 4. Empirical complementary cumulative distribution functions (CCDFs) of schedule overrun by project size group and for all IT projects combined, shown as a log-log plot so that Pareto tails appear as straight lines. Dashed lines show the fitted Pareto tails from Table 6. At each value on the horizontal axis, the CCDF gives the probability of observing schedule overruns larger than x , P(X > x).



Taken together, we find strong evidence against hypothesis H2. Moreover, some of the evidence even points in the opposite direction, suggesting that smaller projects are more prone

to extreme schedule overrun than larger projects. The smaller sample for schedule risk warrants caution in interpreting the finer differences between size groups. However, this does not change the main conclusion that schedule risk does not increase with project size.

5. Discussion

Large IT projects are generally considered more complex and therefore more risky than smaller projects. Our data do not support this view. We find that cost risk is high for all project sizes, and highest for the smallest projects, contrary to conventional wisdom. Schedule risk shows a similar pattern, with no trend of increasing risk with increasing size. If anything, smaller projects have higher schedule risk.

Previously, Flyvbjerg et al. (2022) documented that IT cost risk is fat-tailed, following a power-law distribution. Here we show, for the first time, that this applies to all project sizes, falsifying the common notion that smaller projects are less risky. Our finding is significant for both practice and research. It shows that practitioners are best advised to abandon the common planning assumption that smaller projects will be less risky. It further shows that scholars need to look beyond project complexity, as currently understood in the literature (see Section 2), to explain project risk.

Possibly, project complexity must be assessed relative to experience, or competence. A high risk project for an unexperienced organisation may be bread-and-butter routine (low risk) for an experienced one. In fact, selecting a team with proven successful experience in similar projects is a top heuristic in project management (Flyvbjerg and Gardner, 2023). This suggests that project complexity and size, and their implication for risk, cannot be assessed independently of team competence (another word for proven experience).

Two cognitive biases work against sound competence assessments of project organisations:

(1) The Dunning-Kruger effect makes overconfidence grow with a decreasing competence level (Dunning, 2011; Dunning and Kruger, 1999). In other words, the Dunning-Kruger effect makes self-assessment of competence unreliable. Dunning and Kruger even point to a double negative for project management: (a) a lack of experience makes a task more difficult, and (b) a lack of experience amplifies optimism bias by overestimating own competence. Dunning calls this "The double burden of incompetence" (2011, p. 260).

(2) Uniqueness bias, which is an over-propensity to consider a new project as unique rather than similar to past projects (Flyvbjerg, 2025). Uniqueness bias also causes problems for competence assessment: If new projects are considered unique by default, then it becomes difficult or impossible to determine what should count as relevant experience (competence).

Taking an outside view is a cure against behavioural biases in forecasting and planning (Kahneman and Tversky, 1977; Flyvbjerg, 2007). An outside view for identifying and assessing relevant organisational competences can be based on what we know about the causes of past project failure. The specific factors and causes behind project failure have been identified and reconfirmed by decades of research (Schmidt, 2023). Identifying and assessing general (i.e., non-unique) critical core competences for project management - the ones that make the difference between success and failure – is a future research potential.

It is a common project management practice to let the level of management attention vary with budget size: Bigger budgets, more scrutiny. For example, the Danish government - top ranked on the UN eGovernment Index (UN E-Government Survey, 2024) - exempt IT projects from external risk reviews if their budgets are below $7.5m (up from previously $1.5m). Our results suggest that - resource prioritisation aside - the general practice of exempting smaller

projects from risk assessment and other scrutiny is, in effect, turning a blind eye to significant risk. The doubtful value of budget size as a proxy for risk, as we have shown in this article, leaves a gap for future research. The identification of valid alternative indicators and proxies for risk is an opportunity for future research.

6. Conclusions

The present study finds clear evidence against the widely held assumption that smaller IT projects are less risky than larger ones. We draw four main conclusions:

First, all sizes of IT projects are highly risky in terms of cost and schedule overrun. Remarkably, however, smaller projects are riskier than larger projects.

Second, budget size alone is not a useful indicator of IT project risk. Budget thresholds are inadequate for deciding the appropriate level of management attention. This approach may lead to overlooking small but high-risk projects.

Third, research and practice must rethink IT risk management to effectively reflect the fat tails of cost and schedule risk. Methods that rely on a well-defined mean and variance, as is common, are inadequate for all project sizes. Our findings imply that extreme outcomes are not rare anomalies, they are inherent in the distribution of overruns. Conventional risk management practices that assume normally or near-normally distributed outcomes fail to account for, and underestimate, the likelihood of extremes. Governance frameworks, contingency planning, and escalation mechanisms that explicitly recognize the fat-tailed nature of IT project risk are needed.

Fourth, research and practice face the challenge of mitigating fat-tailed risk.

In this paper, we have analysed the relationship between IT project size and risk. In future research, we will study the relationship between project size and risk for non-IT projects.

REFERENCES:

Aaron, J. Shenhar, (2001). One Size Does Not Fit All Projects: Exploring Classical Contingency Domains. *Management Science* 47(3):394-414.

Ansar, A., Flyvbjerg, B., Budzier, A., & Lunn, D. (2014). Should we build more large dams? The actual costs of hydropower megaproject development. *Energy Policy, 69,* 43–56. https://doi.org/10.1016/j.enpol.2013.10.069

Berg, H. & Ritschel, J. D. (2023). The characteristics of successful military it projects: a cross-country empirical study. *International Journal of Information Systems and Project Management*, 11(2):25–44.

Chiu, C.-M. & Wang, E. T. (2008) Understanding web-based learning continuance intention: The role of subjective task value. *Information & management*, 45(3):194–201.

Claes, W. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE '14). *Association for Computing Machinery*, New York, NY, USA, Article 38, 1–10. https://doi.org/10.1145/2601248.2601268

Clauset, A., Shalizi, C. R., & Newman, M. E. (2009) Power-law distributions in empirical data. *SIAM Review*, 51, 4, 661-703.

Dongus, K., Ebert, S., Schermann, M., & Krcmar, H. (2015) What determines information systems project performance? a narrative review and meta-analysis. In 2015 48th Hawaii International Conference on System Sciences, pages 4483–4492. *IEEE*, 2015.

Du, J.; Wang, Q., & Shi, Q. (2019). Description–experience gap under imperfect information: information continuum and aggressive cost estimating in capital projects. *Engineering, Construction and Architectural Management,* 26(6), 1151-1170.

Dunning, D. (2011) The Dunning–Kruger effect: On being ignorant of one's own ignorance. In Advances in experimental social psychology, volume 44, pages 247–296. Elsevier.

Flyvbjerg, B. (2007) Curbing optimism bias and strategic misrepresentation in planning: Reference class forecasting in practice. *European Planning Studies*, 16(1):3–21.

Flyvbjerg, B., & 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. Penguin Random House, New York.

Flyvbjerg, B., Budzier, A., Aaen, J., Keil, M., & Zottoli, M. (2025). The Uniqueness of IT Cost Risk: A Cross-Group Comparison of 23 Project Types. *Available at SSRN 5247223*.

Flyvbjerg, B., Budzier, A., Lee, J. S., Keil, M., Lunn, D. & Bester, D. W., (2022). The empirical reality of IT project cost overruns: discovering a power-law distribution. *Journal of Management Information Systems*, 39(3), 607–639. https://doi.org/10.1080/07421222.2022.2096544, Available at SSRN: https://ssrn.com/abstract=4204819

Gemino, A., Reich, B., & Sauer, C. (2007) A Temporal Model of Information Technology Project Performance, *Journal of Management Information Systems*, *24*(3), 9-44. DOI: 10.2753/MIS0742-1222240301

Hill, J., Thomas, L. C., & Allen, D. E. (2000). Experts' estimates of task durations in software development projects. *International Journal of Project Management*, 18(1), 13–21. https://doi.org/10.1016/S0263-7863(98)00062-3

Iriarte, C., & Bayona, S. (2020) IT projects success factors: a literature review. *International Journal of Information Systems and Project Management*, 8(2), Article 4.

Jones, C. (1995) Patterns of large software systems: failure and success. IEEE Computer, 28(3):86–87, March.

Jones, T.C. (2007). Estimating software costs. 2nd Edition. New York: McGraw-Hill, Inc.

Jørgensen M., Halkjelsvik T., & Kitchenham B. (2012). How does project size affect cost estimation error? Statistical artifacts and methodological challenges. *International Journal of Project Management*, 30(7), 839–849.

Jørgensen, M. (2019). Relationships between project size, agile practices, and successful software development: results and analysis. *IEEE software*, 36(2), 39-43.

Kahneman, D. and Tversky, A. (1977) Intuitive prediction: Biases and corrective procedures. Technical report, Defence Advanced Research Projects Agency, December.

Kruger, J. and Dunning, D. (1999) Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121.

Mandelbrot, B. B. (1997). Fractals and Scaling in Finance: Discontinuity, Concentration, Risk. Springer.

Martin, N.L., Pearson, J.M., & Furumo, K.A. (2005). IS Project Management: Size, Complexity, Practices and the Project Management Office. Proceedings of the 38th Annual Hawaii International Conference on System Sciences, 234b-234b.

McFarlan, F. W. (1981). Portfolio Approach to Information Systems. *Harvard Business Review*,. 59(5), pp. 142-150.

Morcov, S., Pintelon, L., & Kusters, R. (2021). Definitions, characteristics and measures of IT project complexity - a systematic literature review. *International Journal of Information Systems and Project Management*, *8*(2), 5–21. https://doi.org/10.12821/ijispm080201

Oxford University Press. (2026, March). Complexity. In Oxford English dictionary. Retrieved March 19, 2026,

Peters, R.J., & Verhoef, C. (2008). Quantifying the yield of risk-bearing IT-portfolios. *Science of Computer Programming* 71, pp.17–56.

Sauer, C., Gemino, A.C., & Reich, B.H. (2007). The impact of size and volatility on IT project performance. Commun. *ACM*, 50, 79-84.

Schmidt, J. (2023), Mitigating risk of failure in information technology projects: Causes and mechanisms. *Project Leadership and Society*, 4:100097, December 2023.

Shenhar, A.J., (2001). One Size Does Not Fit All Projects: Exploring Classical Contingency Domains. *Management Science* 47(3):394-414.

Wallace, L., Keil, M., & Rai, A. (2004). Understanding software project risk: A cluster analysis. *Information & Management*, 42(1), 115–125.https://doi.org/10.1016/j.im.2003.12.007

West, G. Scale. (2017) *The Universal Laws of Growth, Innovation, Sustainability, and the pace of Life on Organisms, Cities, Economies, and Companies*. Penguin Press.

Yang, L.-R., O'Connor, J. T., & Wang, C.-C. (2006). Technology utilization on different sizes of projects and associated impacts on composite project success. *International Journal of Project Management, 24*(2), 96–102. https://doi.org/10.1016/j.ijproman.2005.07.002

Zmud, R.W. (1980). Management of large software development efforts. MIS Quarterly, 4, 2, 45–55

Yetton, P., Martin, A., Sharma, R. and Johnston, K. (2000) A model of information systems development project performance. Information Systems Journal, 10(4):263–289.

APPENDIX 1: LITERATURE REVIEW

To compile the initial list of studies, we used a combination of automated searches and snowballing techniques, as presented in Claes, W. (2014). We began by conducting a database search using primary keywords and commonly used terms. The selected keywords reflected project performance as the dependent variable and size as the explanatory variable.

The search was conducted in Scopus and Google Scholar, using terms related to project performance, such as "cost", “schedule”, "performance", and "overrun" alongside the independent variable keyword "project size". This initial search yielded a list of 438 studies spanning various sectors, including construction and transportation, that analyzed the relationship between project performance and project characteristics.

Next, we refined our search to focus specifically on studies involving projects within the IT/IS, software, or ICT domains. This narrowed the list to 221 studies. From this subset, we selected the most pertinent papers that, based on their abstracts, explicitly examined the relationship between project size and IT project performance (Table A1.1). Additionally, we reviewed and included seminal articles known to be relevant to IT project management, ensuring they met the criteria of evaluating project performance as a dependent variable of project size.

From this refined and consolidated list, we identified the most relevant studies and applied backward and forward snowballing techniques, which resulted in the addition of 29 more relevant papers (reported in the bibliography). These studies were incorporated into our background research to ensure a comprehensive review of the topic

Table A1.1. Refined list of articles identified through systematic search.

Title and author		Publication
1	The empirical reality of IT project cost overruns: discovering a power-law distribution. Flyvbjerg, B., Budzier, A., Lee, J. S., Keil, M., Lunn, D. & Bester, D. W., (2022).	Journal of Management Information Systems
2	The impact of size and volatility on IT project performance. Sauer, C., Gemino, A.C., & Reich, B.H. (2007).	Communications of the ACM
3	Is project management: Size, practices and the project management office. Martin, N.L., Pearson, J.M., & Furumo, K.A. (2005).	Journal of Computer Information Systems
4	How does project size affect cost estimation error? Statistical artifacts and methodological challenges. Jørgensen M., Halkjelsvik T.& Kitchenham B. (2012).	International Journal of Project Management
5	IS project management: Size, complexity, practices and the project management office. Martin, N.L., Pearson, J.M., & Furumo, K.A. (2005).	Proceedings of the Annual Hawaii International Conference on System Sciences
6	The characteristics of successful military IT projects: a cross-country empirical study. Berg, H.. & Ritschel, J. D. (2023).	International Journal of Information Systems and Project Management
7	Quantifying the yield of risk-bearing IT-portfolios. Peters, R.J. and Verhoef, C., (2008).	Science of Computer Programming
8	One Size Does Not Fit All Projects: Exploring Classical Contingency Domains. Aaron J. Shenhar, (2001).	Management Science
9	IT projects success factors: a literature review. Iriarte, C. & Bayona, S. (2020).	International Journal of Information Systems and Project Management

APPENDIX 2: SUMMARY OF BACKGROUND STUDIES

Table A2.1. Summary of findings (Part 1)

Paper / Author	Measure of Size	Conclusion	Sample Size	Data Source	Comments
The characteristics of successful military IT projects: a cross-country empirical study. Berg, H., & Ritschel, J. D. (2023)	Size as a cost: proxy of complexity. Greater code amount increases expected cost; project size was examined as an independent variable.	1. Size ≈ performance: larger projects show increased mean cost overruns % and more delays. 2. Project size and contract type affect IT project success. 3. Size (monetary investment) acts as a proxy for project complexity. 4. Skilled personnel and project priority differences influence success.	25 projects, 0.8-351 MUSD	Survey and document. (stated and revealed data)	Size ≈ cost and schedule studied only via mean cost overrun %. No regression.
IS Project Management: Size, Complexity, Practices and the Project Management Office. Martin, N.L., Pearson, J.M., & Furumo, K.A. (2005)	1. Size = cost, duration, system count, team size, vendor count. 2. Complexity = new tech, expertise availability. 3. PM practice. 4. PMO presence.	1. Size predicts budget attainment (larger projects less likely on budget). 2. Size not significant for schedule. 3. Larger size reduces quality. 4. Complexity not significant for quality. Conclusion: larger projects imply harder budget/quality control due to higher cost, team size, vendors, and duration.	129 IT projects (max cost 75M$, max duration 28 months, max systems 387, team 98, vendors 90)	Survey (budget, schedule, quality attainment). Stated data.	No explicit statement if budgets are estimated or actual.
Experts' estimates of task durations in software development projects. J. Hill, L.C. Thomas, D.E. Allen (2000)	Estimated time as cost determinant.	Experts overestimate small tasks and underestimate large ones. Smaller tasks imply better schedule adherence; larger imply worse. Schedule overrun ratio: big projects > 1 (underestimation); small < 1 (overestimation).	500 projects	Document, Database. Revealed data.	Uses estimated working hours; can be used to compute estimated cost.
Impact of size and volatility on IT project performance. Sauer, C., Gemino, A.C., & Reich, B.H. (2007)	1. Effort (person-months) 2. Duration 3. Team Size 4. Budget	Budget: smaller projects perform better; larger projects show worse performance. Effort: more effort implies higher risk of overrun. Duration: longer imply higher risk. Team Size: larger imply more cost/schedule overrun. Conclusion: significant risk exists regardless of size.	412 projects	Survey. Stated data	Risk increases with duration; smaller budgets perform best.

Table A2.2 Summary of findings (Part 2)

Paper / Author	Measure of Size	Conclusion	Sample Size	Data Source	Comments
Relations between Project Size, Agile Practices and Successful Software Development. Jørgensen, M. (2019).	Size as budget.	Increased project size decreases performance for both agile and non-agile, but agile performs better across all sizes.	196 IT projects	Survey Stated data	Not explicit if budget is estimated or actual.
Technology utilization on different sizes of projects and associated impacts on composite project success. Li-Ren Yang, James T. O'Connor, Chu-Ching Wang (2006)	Size as Total Installed Cost (TIC) = actual cost.	1. Tech usage has limited impact on cost/schedule success of large projects. 2. Medium-size projects benefit significantly from tech use. 3. Small projects show positive correlation between tech use and success.	207 IT projects	Survey. Stated data	They classify success as cost/schedule overrun <1.
A Temporal Model of Information Technology Project Performance. Gemino et al. (2008)	Size measured via duration, budget, effort, and relative size.	Project size and complexity directly increase volatility risk. Volatility risk inversely affects project performance. Larger projects imply more changes, lower project performance (budget/schedule).	194 projects, avg budget < $5M, 150 person-months, 15 months duration	Survey. Stated data	
What determines information systems project performance? A narrative review and meta-analysis. Dongus, K., Ebert, S., Schermann, & M., Krcmar, H. (2015)	Project size (as project size variable).	Project size and technological uncertainty have only marginal effects on IS project performance.	157 empirical studies	Meta data	Not explicit if size refers to budget, estimated, or actual.

APPENDIX 3: DETAILS AND ROBUSTNESS ANALYSIS FOR COST OVERRUN

The main text reports the main result: cost risk does not increase with project size, and the smallest projects have the highest cost risk. This appendix gives additional detail and reports our robustness checks for the cost overrun analysis in Section 4.1. The analogous schedule overrun details and robustness checks are reported in Appendix 4. An extended version with additional information about the analysis is available upon request.

The appendix has four purposes. First, we address a possible mathematical coupling problem in the usual cost overrun ratio, because estimated cost is both the denominator of cost overrun and the measure of project size. Second, we give more detail on the generalized linear mixed models (GLMMs) used to estimate typical cost overrun while accounting for contextual variables. Third, we explain the Pareto tail analysis in more detail, including how the lower cut-off $x_{\min}$ and the tail exponent α were estimated. Fourth, we summarize the additional robustness checks, including re-baselining, jittering rounded values, using different numbers of size groups, and removing extreme observations.

A3.1. Mathematical coupling and residual cost overrun

The cost overrun measure we use is

$$CO_i = A_i/E_i,$$

where A_i is actual cost and E_i is estimated cost for project i. Project size is also measured by E_i. This means that estimated cost appears on both sides of the relationship when we relate CO_i to project size. This can induce a mechanical relationship, even if there is no substantive size effect. This is the mathematical coupling issue discussed by Jørgensen et al. (2012).

On the log scale, cost overrun is

$$\log(CO_i) = \log(A_i) - \log(E_i),$$

so any misjudgment in the estimated cost E_i enters directly into the dependent variable CO_i. To check whether our results are driven by this construction, we complement our analysis with a residual cost overrun measure that removes this direct algebraic link.

The residual measure is based on the predictive question: given the estimated cost of a project, how well can we predict its actual cost? We first fit the log-log model

$$\log(A_i) = \beta_0 + \beta_1 \log(E_i) + \epsilon_i,$$

where β_1 is the elasticity of actual cost with respect to estimated cost. If $\beta_1 = 1$, actual cost scales proportionally with estimated cost. If $\beta_1 < 1$, actual cost grows less than proportionally with size. If $\beta_1 > 1$, actual cost grows more than proportionally with size.

The fitted predicted cost $\widehat{A_i}$ and the residual $\widehat{\epsilon_i}$ are

$$\widehat{A_i} = \exp\left(\widehat{\beta_0} + \widehat{\beta_1} \log(E_i)\right),$$

$$\widehat{\epsilon_i} = \log(A_i) - \log\left(\widehat{A_i}\right).$$

We then define residual cost overrun as

$$RCO_i = A_i / \widehat{A_i} = \exp(\widehat{\epsilon_i}).$$

This ratio compares actual cost with the cost predicted from project size. Thus, $RCO_i > 1$ means that the project cost turns out to be more than predicted from its estimated cost, and $RCO_i < 1$ means that the project cost is less than predicted from its estimated cost.

The residual $\widehat{\epsilon_i}$ is uncorrelated with $\log(E_i)$ by construction in the fitted predictive model. Therefore, RCO_i removes the direct mathematical coupling between size and the ordinary cost overrun ratio. The question then becomes whether the distribution of RCO_i, including its center, spread and upper tail, still differs across project sizes.

The usual cost overrun ratio and residual cost overrun ratio are closely related. Substituting the predictive model gives

$$\log(RCO_i) = \log(CO_i) - \widehat{\beta_0} + \left(1 - \widehat{\beta_1}\right)\log(E_i),$$

or, equivalently,

$$RCO_i = CO_i \exp\left(-\widehat{\beta_0}\right) E_i^{1-\widehat{\beta_1}},.$$

Thus, when $\widehat{\beta_1}$ is close to 1, residual cost overrun is close to the usual cost overrun ratio, except that it is corrected for the fitted scaling relation between estimated and actual cost.

In our data, actual cost scales almost proportionally with estimated cost, with $\widehat{\beta_1} \approx 0.99$ and $R^2 \approx 0.901$. This means that estimated cost is a strong predictor of actual cost in absolute dollar terms. Nevertheless, the remaining residual variation is large enough to matter for risk analysis. The residual cost overrun ratio is therefore useful as a robustness check: it studies cost performance after removing the direct dependence on estimated cost.

Figure A3.1. Log-log plot of estimated cost E_i and actual cost A_i. Each point is one IT project. The fitted line shows the expected actual cost conditional on estimated cost. Vertical dashed lines show the quintile cut-offs used to define the five project-size groups.

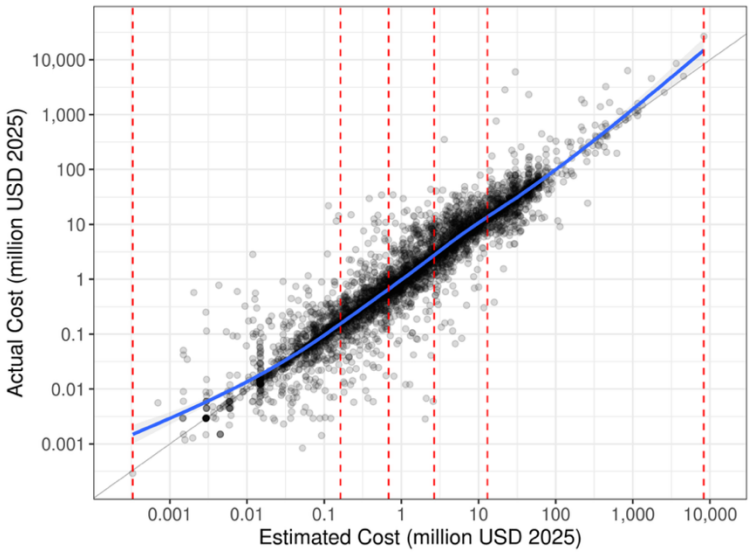


Figure A3.2. Residual cost overrun RCO_i by project size. Residual cost overrun is actual cost divided by the cost predicted from estimated cost. This removes the direct mathematical coupling between estimated cost and the usual cost overrun ratio. Vertical dashed lines show the quintile cut-offs used to define the five project-size groups.

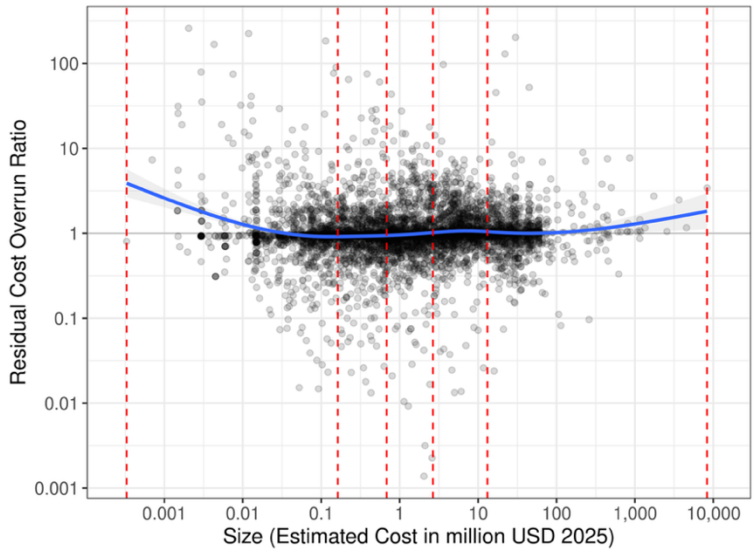


Table A3.1. Descriptive statistics of residual cost overrun (RCO) by project size group. Residual cost overrun is defined as actual cost divided by the cost predicted from estimated cost. Values above 1 indicate projects that cost more than expected after accounting for project size.

Project Size	N	Share of projects with overrun > 1.5	Mean overrun	Median overrun
Very Small (0-20%)	1,019	19.6%	2.75	0.94
Small (20-40%)	1,019	15.0%	1.48	0.95
Medium (40-60%)	1,018	17.1%	1.39	0.95
Large (60-80%)	1,019	23.2%	1.41	0.99
Very Large (80-100%)	1,019	14.4%	1.58	1.01
Total IT	5,094	17.9%	1.72	0.96

A3.2. Non-parametric checks

As an initial check, we applied non-parametric rank-based tests to the subset of projects with cost overrun greater than one. These tests are useful because cost overrun data are skewed and fat-tailed, and they do not require normal residuals. We used the Kruskal-Wallis test to test whether distributions differ across size groups, Dunn's tests for pairwise comparisons, and the Jonckheere-Terpstra test to test for an ordered trend across project sizes.

The Kruskal-Wallis test indicates clear differences between size groups ($\chi^2 = 37.7$, df $= 4$, $p < 0.001$). The Jonckheere-Terpstra test supports a negative ordered trend across project size groups ($p = 0.002$), while an increasing trend is not supported ($p = 1$). Dunn's pairwise tests are consistent with this interpretation: no higher cost overrun for larger projects is indicated where significant. Rather, at the 1% level, Very Small projects have higher cost overrun than Small and Very Large projects, and Very Large projects have lower cost overrun than Medium and Large projects. The comparisons Very Small versus Large, Small versus Medium, and Small versus Large are weaker, but also do not support an increasing size-risk relation. This provides first evidence against H1. However, rank tests are not our main analysis because they neither model contextual variables nor target the mean or the upper tail specifically.

The same pattern is found for residual cost overrun. The Kruskal-Wallis test indicates clear differences between size groups ($\chi^2 = 273.5$, df $= 4$, $p < 0.001$). Dunn's pairwise tests show that the significant differences consistently point in the same direction: larger size groups tend to have lower residual cost overrun. At the 1% level, this applies to Very Small versus Small, Large, and Very Large projects, and to all comparisons involving Very Large projects. The comparison between Very Small and Medium projects points in the same direction but is weaker ($p = 0.021$). The Jonckheere-Terpstra test further supports a negative ordered trend across project size groups ($p = 0.001$), while an increasing trend is not supported ($p = 1$).

Overall, the non-parametric tests do not support the hypothesis that cost risk increases with project size. If anything, the ordered evidence points in the opposite direction.

A3.3. GLMM model selection and interpretation

The descriptive means show large differences between size groups, but they do not control for contextual variables. Projects from the same country, time period, subtype, or region may be more similar to each other than to other projects. We therefore used GLMMs to estimate the

relationship between project size and cost overrun while accounting for such contextual clustering.

Since the outcome variable (cost overrun ratio) is positive and skewed, ordinary linear regression is not appropriate. We used log-normal GLMMs because they model positive outcomes, allow skewness, and provide interpretable multiplicative effects on the overrun ratio. The log-normal model is not intended to model the extreme upper tail perfectly. The upper tail is modelled separately in the Pareto analysis below. The GLMM is used for the bulk and the typical level of cost overrun while controlling for context.

To select this model for our cost overrun analysis, we compared a broad set of candidate models. Each model included project size group as the main predictor. We then varied whether *Country*, *Year of FID*, *Subtype*, and *Region* entered as fixed effects, random effects, or not at all. Because *Country* and *Region* are not independent contextual classifications, region-based models were considered separately from country-based models. For each mean structure, we fitted two dispersion structures: common dispersion across all size groups and size-group-specific dispersion. This produced 45 mean structures times 2 dispersion structures, or 90 candidate models for each outcome.

Models were compared by AIC (Akaike information criterion) and BIC (Bayesian information criterion or Schwarz criterion). We use BIC as the main criterion because it penalizes extra parameters more strongly than AIC. This is useful here as many contextual models are possible, and we want a model that improves fit without adding unnecessary complexity. AIC-selected models were used as an additional check and led to the same qualitative conclusion.

Candidate models were compared on a common complete-case subset for the contextual variables, so that BIC differences were not driven by changing sample composition. After

model selection, the selected model was fitted to all observations with data available for the variables retained in that model.

The top 10 candidate models are shown in Table A3.2. The best BIC model for cost overrun uses a log-normal distribution, random intercepts for Country and Year of FID, and size-group-specific dispersion, as shown in Table A3.3.

Table A3.2. GLMM model-selection results for cost overrun. Candidate models differ in which contextual variables are included, whether these variables enter as fixed or random effects, and whether dispersion is constant or varies by project-size group. Models are sorted by BIC. Lower BIC values indicate better support after penalizing model complexity. BIC and AIC are shown as difference to the best scoring models, respectively. Only the top 10 candidate models out of a total of 90 are shown.

Effects Structure	**AIC**	**BIC**	l	k	n	Δ**AIC**	Δ**BIC**	**Dispersion**
Quintiles fixed, Country + YearFID random	3831.9	3899.2	-1904.0	12	2012	25.7	**0.0**	~ quintiles
Quintiles fixed, Country + Subtype + YearFID random	3833.9	3906.8	-1904.0	13	2012	27.7	7.6	~ quintiles
Quintiles fixed, Country + Subtype random	3844.0	3911.3	-1910.0	12	2012	37.8	12.1	~ quintiles
Quintiles + Subtype fixed, Country random	3839.0	3945.5	-1900.5	19	2012	32.8	46.3	~ quintiles
Quintiles fixed, Country random	3886.6	3948.3	-1932.3	11	2012	80.5	49.1	~ quintiles
Quintiles + Subtype fixed, Country + YearFID random	3838.7	3950.8	-1899.3	20	2012	32.5	51.6	~ quintiles
Quintiles fixed, Country + YearFID random	3927.1	3972.0	-1955.6	8	2012	121.0	72.8	~ 1
Quintiles fixed, Country + Subtype + YearFID random	3929.0	3979.5	-1955.5	9	2012	122.8	80.3	~ 1
Quintiles + YearFID fixed, Country random	3806.2	3985.6	-1871.1	32	2012	**0.0**	86.4	~ quintiles
Quintiles fixed, Country + Subtype random	3942.3	3987.1	-1963.1	8	2012	136.1	87.9	~ 1

The detailed fit results in Table A3.3 complement the results shown and discussed in the main text: With very small projects as the baseline, the fitted cost-overrun ratio is 1.826. The

other size groups are estimated as multiplicative factors relative to this baseline. Thus, after controlling for country and year, the fitted cost overrun is highest for very small projects and decreases with project size. All size groups have lower fitted cost overrun than very small projects, with $p < 0.001$ in the baseline model.

Table A3.3. Full results of the best BIC GLMM for cost overrun. The model uses a log-normal distribution, project size group as fixed effect, random intercepts for Country and Year of FID, and size-group-specific dispersion. For this model specification, 4834 data points are available. Estimates for non-baseline size groups are multiplicative factors relative to very small projects.

Fit result for best BIC-GLMM (CostOverrun) *Log-normal: CostOverrun ~ quintiles + (1 | Country) + (1 | YearFID) -- dispersion: ~quintiles*					
Parameter (Project size group)	**Coefficient**	**SE**	**99% Confidence Interval**	***z***	***p***
Fixed effects:					
Very Small (intercept)	1.8261	0.0964	(1.5941, 2.0920)	11.4132	< .001
Small	0.7991	0.0362	(0.7111, 0.8981)	-4.9465	< .001
Medium	0.7774	0.0358	(0.6904, 0.8753)	-5.4667	< .001
Large	0.7150	0.0312	(0.6389, 0.8001)	-7.6825	< .001
Very Large	0.6115	0.0261	(0.5478, 0.6826)	-11.5229	< .001
Dispersion:					
Very Small (intercept)	0.7155	0.0610	(0.5585, 0.8725)	11.7367	< .001
Small	-0.4988	0.0786	(-0.7012, -0.2964)	-6.3487	< .001
Medium	-0.4710	0.0803	(-0.6777, -0.2643)	-5.8688	< .001
Large	-0.6950	0.0757	(-0.8899, -0.5000)	-9.1820	< .001
Very Large	-0.9518	0.0738	(-1.1418, -0.7618)	-12.9013	< .001
Random effects variances:					
SD τ_C (Intercept: Country)		0.1209			
SD τ_T (Intercept: YearFID)		0.0912			

A possible concern with these results is that it uses very small projects as the baseline. To check that the result is not an artefact of this choice, we refitted the same best-BIC specification with each size group used as the reference category.

This more comprehensive comparison between project size groups is included as Table A3.4, which shows the GLM model results as pairwise comparisons between project size

groups. Each cell gives the typical cost overrun of the column group relative to the row group, after adjusting for country and year of final investment decision. The pattern is clear: typical cost overrun generally decreases with project size with strong statistical evidence.

Table A3.4. Pairwise comparison of typical cost overrun between project size groups, estimated with the GLM model adjusting for country and year of final investment decision. Each cell shows the estimated typical cost overrun of the column group relative to the row group. Values below 1 indicate lower overrun than the row group (shaded in green); values above 1 indicate higher overrun (shaded in red). Bold values indicate high statistical evidence ($p < 0.001$).

	Very Small	Small	Medium	Large	Very Large
Very Small	1	**0.8**	**0.78**	**0.72**	**0.61**
Small	**1.25**	1	0.97	**0.89**	**0.77**
Medium	**1.29**	1.03	1	0.92	**0.79**
Large	**1.4**	**1.12**	1.09	1	**0.86**
Very Large	**1.64**	**1.31**	**1.27**	**1.17**	1

We repeated the same GLMM analysis for residual cost overrun RCO_i. This is the direct robustness check for mathematical coupling. The best BIC model has the same contextual structure as for CO_i. The fitted residual cost overrun level is again highest for very small projects, as evident from Table A3.5. This finding is retained also after changing the baseline.

Thus, the same qualitative result holds after removing the direct algebraic link between estimated cost and cost overrun. This is an important robustness result: the finding that very small projects have the highest cost risk is not just caused by estimated cost appearing in the denominator of the cost overrun ratio.

Table A3.5. Full results of the best BIC GLMM for residual cost overrun. The model uses a log-normal distribution, project size group as fixed effect, random intercepts for Country and Year of FID, and size-group-specific dispersion. Estimates for non-baseline size groups are multiplicative factors relative to very small projects.

Fit result for best BIC-GLMM (ResidualCostOverrun) *Log-normal: ResidualCostOverrun ~ quintiles + (1 | Country) + (1 | YearFID) -- dispersion: ~quintiles*					
Parameter (Project size group)	**Coefficient**	**SE**	**99% Confidence Interval**	***z***	***p***
Fixed effects:					
Very Small (intercept)	1.7331	0.0914	(1.5131, 1.9852)	10.4310	< .001
Small	0.8159	0.0369	(0.7261, 0.9168)	-4.4945	< .001
Medium	0.8032	0.0369	(0.7135, 0.9043)	-4.7633	< .001
Large	0.7484	0.0326	(0.6689, 0.8374)	-6.6472	< .001
Very Large	0.6506	0.0277	(0.5830, 0.7261)	-10.0861	< .001
Dispersion:					
Very Small (intercept)	0.6602	0.0608	(0.5035, 0.8169)	10.8516	< .001
Small	-0.4750	0.0785	(-0.6771, -0.2728)	-6.0524	< .001
Medium	-0.4354	0.0802	(-0.6419, -0.2290)	-5.4324	< .001
Large	-0.6466	0.0756	(-0.8413, -0.4519)	-8.5543	< .001
Very Large	-0.8872	0.0737	(-1.0769, -0.6974)	-12.0428	< .001
Random effects variances:					
SD τ_C (Intercept: Country)		0.1211			
SD τ_T (Intercept: YearFID)		0.0912			

Diagnostic plots showed no strong remaining trend in the residual-versus-fitted plots for the best BIC GLMMs. The Q-Q plots confirmed that the log-normal GLMMs do not fully capture the far upper tail, which is why tail risk is analysed separately with Pareto-type models.

A3.4. Pareto tail analysis

The GLMM analysis concerns typical cost overrun. Cost risk, however, is also strongly driven by rare but very large overruns. We therefore analyze the right tail separately.

For each project-size group, we fit a Pareto Type I model to the upper tail of the cost overrun distribution. We use the survival-function parameterization

$$\Pr(X > x \mid X \geq x_{\min}) = (x/x_{\min})^{-\alpha}, \quad x \geq x_{\min}\,, \tag{A3.1}$$

which implies the probability density function

$$f(x) = \frac{\alpha}{x_{\min}}\left(\frac{x}{x_{\min}}\right)^{-(\alpha+1)}, \quad x \geq x_{\min}\,. \tag{A3.2}$$

Here, $x_{\min}$ is the lower cut-off for the tail and α is the Pareto tail exponent. Smaller α means a fatter tail and higher risk of extreme overruns.

Under this parameterization:

- $\alpha \leq 1$ means infinite mean and infinite variance.
- $1 < \alpha \leq 2$ means finite mean but infinite variance.
- $\alpha > 2$ means finite mean and finite variance.

Thus, the value of α is directly interpretable as a risk category. Values below or near 1 indicate the most extreme form of cost risk. Values between 1 and 2 still indicate very high risk, because the variance is infinite and sample averages can be highly unstable in practice.

We follow the Clauset et al. (2009) procedure, implemented using the poweRlaw work-flow in R, with additional checks. The procedure has three main steps.

Step 1 estimates $x_{\min}$ and α. The lower cut-off $x_{\min}$ is selected by minimizing the Kolmo-gorov-Smirnov distance between the empirical tail and the fitted Pareto tail. We then use a selection-aware bootstrap with 5000 simulations to obtain 99% confidence intervals for both $x_{\min}$ and α. This is important because $x_{\min}$ is itself estimated from the data and is not known in advance.

Step 2 checks whether the Pareto tail is plausible. We use a semi-parametric bootstrap goodness-of-fit test based on the Kolmogorov-Smirnov distance. In this test, a high p-value does not prove that the Pareto model is true. It means that the observed distance between the data and the fitted Pareto tail is not unusually large compared with data generated from the

fitted model. In line with Clauset et al. (2009), we treat $p > 0.10$ as "Pareto plausible" and $p \leq 0.10$ as evidence against a Pareto tail.

Step 3 compares the Pareto model with alternative tail models. We compare against the log-normal, exponential, and Weibull distributions using likelihood-ratio tests at a common $x_{\min}$. We also compute ΔAIC and ΔBIC relative to the Pareto model. Negative values favor the alternative; positive values favor the Pareto model.

For the log-normal alternative, we additionally checked whether the fit was only achieved by pushing the parameters to extreme values. This is important because a log-normal distribution can resemble a power law over any finite range, especially in finite data. We therefore used generous parameter bounds in the log-normal fit. The log-scale standard deviation was allowed to range from 0.05 to 10, where the upper value is extremely large on the original ratio scale. The log-mean was allowed to range from the smallest to the largest observed log-value in the fitted tail, further extended by two empirical log-standard deviations on both sides. If the fitted log-normal reaches such a boundary, we treat this as a warning sign that the log-normal is not a stable or suitable tail model, even when the bootstrap goodness-of-fit p-value alone would classify it as plausible.

In addition to the Pareto distributions and the alternatives, we fit a log-Weibull tail model,

$$\Pr(X > x \mid X \geq x_{\min}) = \exp\left(-(\log(x/x_{\min})/\theta)^{\beta}\right)$$

The log-Weibull is useful because it includes the Pareto tail as a special case. When $\beta = 1$, it becomes a Pareto tail with $\alpha = 1/\theta$. When $\beta < 1$, the tail is even heavier than a power law ("super-heavy"). When $\beta > 1$, the tail is lighter than a power law, but still heavy compared with exponential-type tails. We therefore use it to check whether the data are consistent with "power-law or heavier" tails.

One technical point is important. The poweRlaw package internally uses the density-exponent convention. In the paper, we report the Pareto Type I survival exponent α, which is the exponent relevant for the mean and variance categories above. Therefore, the package output is transformed to the reported α.

Before fitting tails, we filter out values close to one. For cost overrun and residual cost overrun, the main analysis uses values above 1.1 for the tail search. This avoids letting the many observations at or near one, distort the tail fit and the KS statistic. For the likelihood-ratio comparisons, we use a fixed $x_{\min} = 2.0$ for cost and residual cost overrun as a bias-variance trade-off: it keeps enough tail observations for useful comparison while focusing on clearly substantial overruns. For the alternative goodness-of-fit checks, $x_{\min}$ is optimized separately by the KS criterion.

The main cost-overrun tail results are reported in the main text. They show that very small projects have the lowest α, with $\alpha = 0.874$. This implies infinite mean and variance under the fitted Pareto model. Other size groups have α values mostly between 1 and 2, implying finite mean but infinite variance. Thus, all size groups have very high cost risk, but very small projects have the most extreme estimated tail risk. This result is illustrated in Figure A3.3.

Figure A3.3. Pareto I fit result for cost overrun by project size group and for all IT projects combined. Markers show the estimated values of α and $x_{\min}$ reported in Table 3. Vertical lines indicate 99% confidence intervals for α, and horizontal lines represent uncertainty in $x_{\min}$.

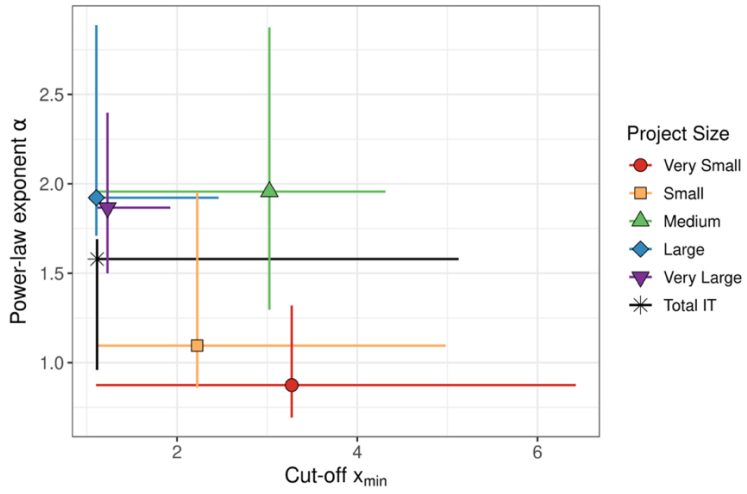


Table A3.6 adds the full goodness-of-fit information for the cost overrun tail fits. The Pareto Type I model is plausible for all five project-size groups. The pooled Total IT fit is not plausible, instead, a super-heavy tail law given by the log-Weibull distribution is plausible.

Table A3.6. Selection-aware Pareto Type I tail-fit summary for cost overrun. The table reports the estimated lower cut-off $x_{\min}$, Pareto exponent α, number of observations above the cut-off, and semi-parametric bootstrap goodness-of-fit p-value. The final column gives the log-Weibull tail classification, where $\beta \leq 1$ supports a power-law or heavier tail and $\beta > 1$ indicates a lighter, but still heavy-tailed, log-Weibull fit.

Project size group	N	$x_{\min}$	α	N_{tail}	GoF p	**Plausibility**	**Log-Weibull tail class**
Very Small	321	3.273	0.874	82	0.456	power-law plausible	power-law / super-heavy plausible
Small	291	2.224	1.095	69	0.530	power-law plausible	power-law / super-heavy plausible
Medium	314	3.025	1.956	85	0.364	power-law plausible	plausible, but $\beta > 1$
Large	418	1.105	1.922	415	0.120	power-law plausible	plausible, but $\beta > 1$
Very Large	284	1.229	1.866	216	0.221	power-law plausible	super-heavy plausible
Total IT	1628	1.113	1.579	1586	0.000	power-law not plausible	super-heavy plausible

Table A3.7 reports the corresponding selection-aware bootstrap confidence intervals for $x_{\min}$ and α. The uncertainty in $x_{\min}$ is large, which is expected in tail fitting, but the ranking of tail risk remains informative. Very small projects have the lowest estimated α, and hence the fattest estimated tail.

Table A3.8 compares the Pareto Type I model with alternative tail models at a common cut-off, $x_{\min} = 2.0$. This common cut-off is used only for model comparison, so that the likelihood-ratio tests compare the models on the same observations. The exponential model is clearly disfavored in most groups, which rules out a thin-tailed interpretation. Log-normal and log-Weibull alternatives can sometimes fit similarly well or better, but this does not weaken the risk conclusion: these alternatives are still heavy-tailed over the relevant range and do not support increasing cost risk with project size.

Table A3.7. Selection-aware bootstrap confidence intervals for Pareto Type I cost overrun tail fits, based on 5000 bootstrap samples accounting for uncertainty in both the lower cut-off x_{min} and the Pareto exponent α.

Project size group	$x_{\min}$ **lower**	$x_{\min}$	$x_{\min}$ **upper**	α **lower**	α	α **upper**
Very Small	1.102	3.273	6.425	0.693	0.874	1.319
Small	1.103	2.224	4.982	0.860	1.095	1.956
Medium	1.102	3.025	4.313	1.296	1.956	2.875
Large	1.101	1.105	2.460	1.710	1.922	2.888
Very Large	1.103	1.229	1.925	1.499	1.866	2.397
Total IT	1.101	1.113	5.123	0.959	1.579	1.690

Table A3.8. Comparison of Pareto Type I tail fits with alternative tail distributions for cost overrun at the common cut-off $x_{min} = 2.0$. Alternatives are log-normal, exponential, Weibull, and log-Weibull. Likelihood-ratio statistics and p-values compare each alternative with the Pareto model. ΔAIC and ΔBIC are relative to Pareto Type I; negative values favor the alternative, positive values favor Pareto Type I

Project size group	$x_{\min}$	N_{tail}	Alternative	LR statistics	p (1-sided)	p (2-sided)	ΔAIC	ΔBIC
Very Small	2	130	lognormal	1.058	0.145	0.290	1.955	4.823
Very Small	2	130	exponential	6.878	0.000	0.000	190.480	190.480
Very Small	2	130	weibull	2.751	0.003	0.006	6.471	9.338
Very Small	2	130	logweibull	-	-	-	1.822	4.689
Small	2	86	lognormal	0.250	0.401	0.803	2.003	4.457
Small	2	86	exponential	4.318	0.000	0.000	77.560	77.560
Small	2	86	weibull	1.933	0.027	0.053	6.393	8.848
Small	2	86	logweibull	-	-	-	-2.880	-0.426
Medium	2	138	lognormal	-0.670	0.748	0.503	-0.095	2.832
Medium	2	138	exponential	1.686	0.046	0.092	34.381	34.381
Medium	2	138	weibull	-0.413	0.660	0.680	0.250	3.178
Medium	2	138	logweibull	-	-	-	-0.414	2.513
Large	2	140	lognormal	0.071	0.472	0.943	2.001	4.943
Large	2	140	exponential	1.292	0.098	0.196	89.025	89.025
Large	2	140	weibull	0.635	0.263	0.525	6.604	9.546
Large	2	140	logweibull	-	-	-	0.381	3.322
Very Large	2	81	lognormal	0.087	0.465	0.930	2.004	4.398
Very Large	2	81	exponential	4.146	0.000	0.000	178.312	178.312
Very Large	2	81	weibull	3.445	0.000	0.001	18.818	21.212
Very Large	2	81	logweibull	-	-	-	-7.299	-4.904
Total IT	2	575	lognormal	0.586	0.279	0.558	2.023	6.378
Total IT	2	575	exponential	7.442	0.000	0.000	829.134	829.134
Total IT	2	575	weibull	5.978	0.000	0.000	54.513	58.867
Total IT	2	575	logweibull	-	-	-	-4.227	0.127

Tables A3.9 and A3.10 give additional checks where the Pareto Type I, log-normal, and log-Weibull models are each fitted at their own KS-optimized cut-off. The log-normal often remains plausible, but the boundary hit rates show that this plausibility can depend on extreme parameter values. The log-Weibull fits remain useful because they distinguish between power-law-like, heavier-than-power-law, and lighter-than-power-law tails. Overall, the

conclusion is unchanged: cost overrun is fat-tailed in all size groups, and very small projects have the most extreme estimated tail risk.

Table A3.9. Parameter estimates for Pareto Type I (Par-1), log-normal (LN), and log-Weibull (LW) tail fits for cost overrun, with each model fitted at its own KS-optimized lower cut-off $x_{\min}$. For Pareto Type I, the fit result is α. For the log-normal, the fit result gives the log-mean μ and log-scale standard deviation σ. For the log-Weibull, the fit result gives θ and β. Brackets give bootstrap confidence intervals.

Project size group	N	**Par-1** $x_{\min}$	**Par-1** N_{tail}	**Pareto-1 fit result**	**LN** $x_{\min}$	**LN** N_{tail}	**Log-normal fit result**	**LW** $x_{\min}$	**LW** N_{tail}	**Log-Weibull fit result**
Very Small	321	3.2727	81	$\alpha = 0.863$ [0.659, 1.170]	3.0000	85	μ = -1.103 [-1.526, 1.858]; $\sigma = 2.277$ [1.109, 2.721]	2.3684	116	$\theta = 1.02$ [0.766, 1.351]; $\beta = 0.887$ [0.754, 1.088]
Small	291	2.2237	68	$\alpha = 1.079$ [0.799, 1.507]	4.3478	31	$\mu = 0.378$ [-0.863, 2.515]; $\sigma = 1.64$ [0.598, 2.387]	2.1310	75	$\theta = 0.854$ [0.601, 1.167]; $\beta = 0.932$ [0.756, 1.239]
Medium	314	3.0250	84	$\alpha = 1.933$ [1.487, 2.636]	1.4932	184	μ = -0.633 [-1.023, 0.634]; $\sigma = 1.274$ [0.786, 1.508]	1.6000	162	$\theta = 0.807$ [0.681, 0.945]; $\beta = 1.311$ [1.129, 1.56]
Large	418	1.1053	414	$\alpha = 1.918$ [1.694, 2.182]	1.2199	335	μ = -0.778 [-0.854, 0.187]; $\sigma = 1.019$ [0.682, 1.118]	1.2199	335	$\theta = 0.558$ [0.487, 0.634]; $\beta = 1.121$ [1.01, 1.257]
Very Large	284	1.2286	215	$\alpha = 1.858$ [1.566, 2.222]	1.1786	236	μ = -1.168 [-1.013, 0.193]; $\sigma = 1.168$ [0.722, 1.231]	1.1026	283	$\theta = 0.471$ [0.39, 0.565]; $\beta = 0.88$ [0.782, 0.999]
Total IT	1628	1.1131	1585	$\alpha = 1.578$ [1.483, 1.686]	12.5304	47	$\mu = 2.497$ [0.332, 3.653]; $\sigma = 1.409$ [0.726, 2.34]	1.1027	1621	$\theta = 0.601$ [0.559, 0.645]; $\beta = 0.911$ [0.866, 0.957]

Table A3.10. Bootstrap goodness-of-fit comparison for Pareto Type I, log-normal, and log-Weibull tail fits for cost overrun, with each model fitted at its own KS-optimized lower cut-off x_{min}. The log-normal boundary hits column reports the share of bootstrap fits in which the log-normal fit reached one of the chosen parameter bounds. High values indicate unstable log-normal fits, even if the goodness-of-fit p-value is acceptable.

Project size group	***N***	**Par-1 GoF *p***	**Pareto-1 verdict**	**LN GoF *p***	**LN verdict**	**Log-Normal boundary hits**	**LW GoF *p***	**LW verdict**	**Log-Weibull tail class**
Very Small	321	0.988	power-law plausible	0.265	log-normal plausible	0.461	0.908	log-Weibull plausible	power-law / super-heavy plausible
Small	291	0.976	power-law plausible	0.728	log-normal plausible	0.343	0.994	log-Weibull plausible	power-law / super-heavy plausible
Medium	314	0.694	power-law plausible	0.372	log-normal plausible	0.384	0.550	log-Weibull plausible	plausible, but $\beta > 1$
Large	418	0.345	power-law plausible	0.383	log-normal plausible	0.503	0.871	log-Weibull plausible	plausible, but $\beta > 1$
Very Large	284	0.477	power-law plausible	0.001	log-normal not plausible	0.608	0.412	log-Weibull plausible	super-heavy plausible
Total IT	1628	0.002	power-law not plausible	0.892	log-normal plausible	0.125	0.186	log-Weibull plausible	super-heavy plausible

To check mathematical coupling at the tail level, we repeated the tail analysis for residual cost overrun. If the main tail result were only mechanical, this analysis should change it.

Figure A3.4. Pareto Type I tail fits for residual cost overrun by project-size group and for all IT projects combined. Markers show the estimated lower cut-off x_{min} and Pareto exponent α. Vertical lines show 99% bootstrap confidence intervals for α, and horizontal lines show uncertainty in x_{min}.

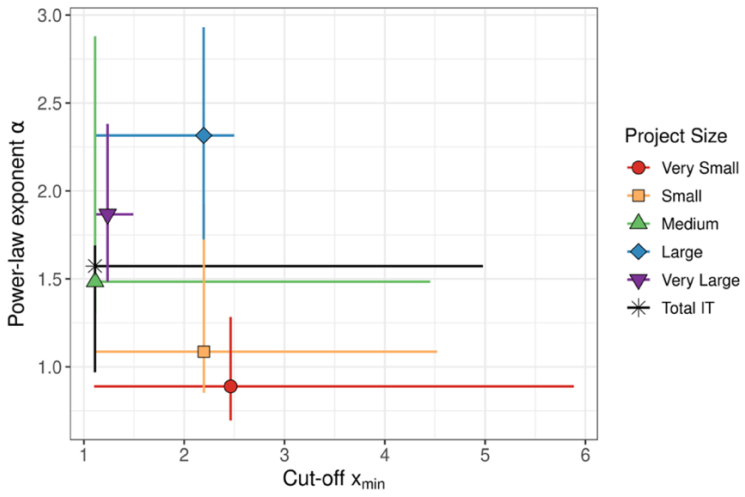
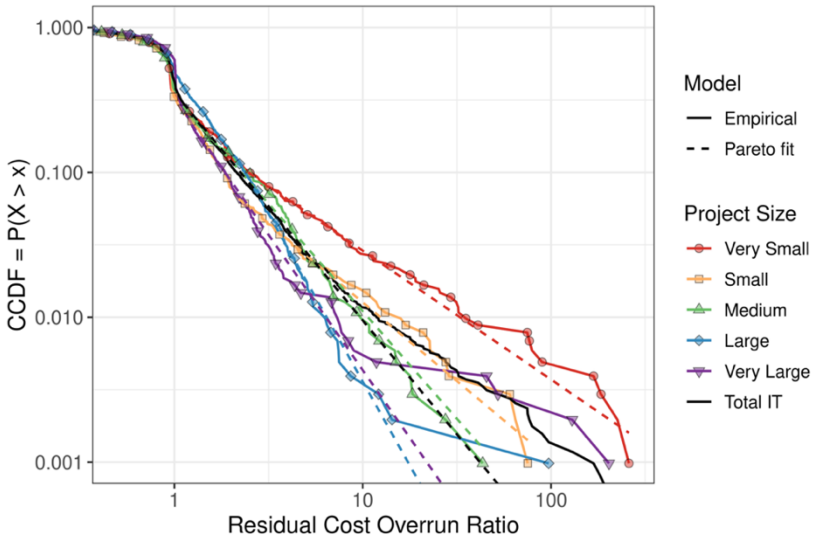


Figure A3.5. Empirical complementary cumulative distribution functions (CCDFs) and fitted Pareto Type I tails for residual cost overrun. The plots are shown on doubly logarithmic axes, where Pareto tails appear approximately as straight lines. Dashed lines show the fitted Pareto tails above the estimated x_{min}.



It does not. Figures A3.4 and A3.5 show that very small projects again have the lowest fitted α and remain visibly heavy-tailed after correcting for the fitted relation between estimated and actual cost. Table A3.11 gives details: Pareto I is plausible for all groups except medium, but this exception does not point toward increasing risk with size; the lowest α is again found for very small projects, followed by small projects.

Table A3.11. Selection-aware Pareto Type I tail-fit summary for residual cost overrun. The table reports the estimated lower cut-off $x_{\min}$, Pareto exponent α, number of observations above the cut-off, and semi-parametric bootstrap goodness-of-fit p-value. The final column gives the log-Weibull tail classification.

Project size group	***N***	$x_{\min}$	α	N_{tail}	**GoF** p	**Plausibility**	**Log-Weibull tail class**
Very Small	299	2.463	0.889	102	0.440	power-law plausible	power-law / super-heavy plausible
Small	271	2.196	1.086	67	0.510	power-law plausible	power-law / super-heavy plausible
Medium	304	1.112	1.484	292	0.002	power-law not plausible	plausible, but $\beta > 1$
Large	410	2.194	2.315	123	0.712	power-law plausible	plausible, but $\beta > 1$
Very Large	290	1.236	1.867	218	0.330	power-law plausible	power-law / super-heavy plausible
Total IT	1574	1.111	1.573	1538	0.000	power-law not plausible	super-heavy plausible

The bootstrap intervals in Table A3.12 show that uncertainty in $x_{\min}$ remains substantial, as expected in tail fitting. However, the broad pattern remains the same: residual cost overrun is fat-tailed, and the smallest projects remain among the riskiest groups.

Table A3.12. Selection-aware bootstrap confidence intervals for the Pareto Type I tail fits for residual cost overrun based on 5000 bootstrap samples accounting for uncertainty in both $x_{\min}$ and α.

Project size group	$x_{\min}$ **lower**	$x_{\min}$	$x_{\min}$ **upper**	α **lower**	α	α **upper**
Very Small	1.101	2.463	5.884	0.695	0.889	1.284
Small	1.102	2.196	4.522	0.852	1.086	1.928
Medium	1.101	1.112	4.452	1.300	1.484	2.880
Large	1.100	2.194	2.499	1.724	2.315	2.931
Very Large	1.102	1.236	1.492	1.484	1.867	2.381
Total IT	1.100	1.111	4.979	0.969	1.573	1.691

We conducted the same alternative-model checks as for ordinary cost overrun. These checks again do not support a thin-tailed interpretation. The exponential model is generally a poor description of the tail. The log-normal can be plausible in some cases, but often with unstable parameter behavior. The log-Weibull checks support the broader conclusion that the tail behavior is power-law-like or heavier for the smallest groups and remains far from normal or thin-tailed.

A3.5. Additional robustness checks for cost overrun

We also ran several additional robustness checks that are not shown in full, because they returned similar results to what is already shown, and therefore did not change the qualitative conclusion.

First, we changed the number of project size groups. The main analysis uses quintiles because five equally sized groups are easy to interpret and still give enough observations per group. We repeated the analysis using tertiles, quartiles, sextiles, and septiles. The conclusion did not depend on the exact cut-points used for grouping estimated cost.

Second, we jittered rounded overrun values. Many observations are reported with limited precision, and many values are close to one. This can create ties, which matter for rank tests and KS-based tail fitting. In the jittered analysis, small random noise was added only below the reported precision, using fixed seeds for reproducibility. The results remained qualitatively unchanged.

Third, we removed extreme observations from both ends of the distribution. We repeated the analysis after removing the lowest and highest 0.1% of observations, and again after removing the lowest and highest 0.5%. This is a conservative check, especially for the upper tail, because it removes observations that are most important for extreme-risk estimation. The main conclusion still holds.

Fourth, we explored alternative GLMM distributional families, including Gamma-type and Tweedie-type alternatives. These alternatives did not provide a better basis for the main analysis. The log-normal GLMM remained the most useful model for comparing typical overrun while controlling for context, and the extreme tail was then analyzed separately using Pareto and related tail models.

Taken together, the robustness checks show that the cost overrun result is not driven by one modelling choice. It is not driven by mathematical coupling, by the choice of reference group, by the exact number of size groups, by rounded values, by the most extreme observations, or by the selected GLMM family. The evidence consistently goes against H1: increasing project size is not associated with increasing cost risk.

APPENDIX 4: ROBUSTNESS ANALYSIS FOR SCHEDULE OVERRUN

This appendix details the schedule overrun analysis in Section 4.2. The methods are the same as in Appendix 3: non-parametric tests, GLMM model selection, model diagnostics, Pareto Type I tail fitting, and comparison with alternative tail distributions. We therefore do not repeat the full methodological explanation here.

Schedule overrun has no mathematical-coupling problem: it is actual duration divided by estimated duration, while project size is measured by estimated cost. Estimated cost is therefore not the denominator, so no residual correction is needed.

A4.1. GLMM model selection and interpretation

Table A4.1 shows the schedule-overrun GLMM model-selection results. The best BIC model is simpler than the cost-overrun model. It uses project-size group as fixed effect, common dispersion, and no retained contextual random effect. Adding contextual random effects or size-group-specific dispersion improves AIC in some cases, but not enough to overcome the stronger BIC penalty for extra parameters.

The full results of the best BIC model are shown in Table A4.2. The fitted schedule overrun is 1.50 for very small projects. Relative to this baseline, small and very large projects have lower fitted schedule overrun. Medium and large projects have higher fitted values, but with weaker support. The pattern is therefore mixed and does not support a monotonic increase in schedule overrun with project size.

We also considered the second-best BIC model, which allows dispersion to differ by project-size group. It has a lower AIC than the best BIC model, but a higher BIC. Here, some group differences appear in dispersion rather than only in the fitted mean. This does not change the substantive conclusion. It shows that schedule overrun differs across size groups in both level and spread, but not in a way that supports increasing schedule risk with project size.

Table A4.1. GLMM model-selection results for schedule overrun. Candidate models differ in contextual variables, random effects, and dispersion structure. Models are sorted by BIC. Lower BIC values indicate better support after penalizing model complexity. Intermediate models before the best AIC model are omitted.

Effects Structure	**AIC**	**BIC**	l	k	n	ΔAIC	ΔBIC	**Dispersion**
Quintiles fixed	1273.7	1300.2	-630.8	6	615	13.8	**0.0**	~ 1
Quintiles fixed	1261.5	1305.7	-620.7	10	615	1.6	5.5	~ quintiles
Quintiles fixed, Country random	1275.0	1306.0	-630.5	7	615	15.1	5.8	~ 1
Quintiles fixed, YearFID random	1275.3	1306.2	-630.6	7	615	15.4	6.0	~ 1
Quintiles fixed, Subtype random	1275.7	1306.6	-630.8	7	615	15.8	6.4	~ 1
Quintiles fixed, Region random	1276.1	1307.1	-631.1	7	615	16.2	6.9	~ 1
Quintiles fixed, Country + YearFID random	1275.5	1310.8	-629.7	8	615	15.6	10.7	~ 1
Quintiles fixed, Region random	1262.4	1311.0	-620.2	11	615	2.5	10.8	~ quintiles
Quintiles fixed, YearFID random	1263.2	1311.9	-620.6	11	615	3.3	11.7	~ quintiles
Quintiles fixed, Subtype random	1263.5	1312.1	-620.7	11	615	3.6	11.9	~ quintiles
(52 more candidate models)								
Quintiles + Country fixed	1259.9	1423.5	-592.9	37	615	0.0	123.3	~ quintiles

Table A4.2. Full results of the best BIC GLMM for schedule overrun. The model uses a log-normal distribution, project-size group as fixed effect, and common dispersion. All 831 data points can be used for fitting. Estimates for non-baseline size groups are multiplicative factors relative to very small projects.

Fit result for best BIC-GLMM (ScheduleOverrun) *Log-normal: ScheduleOverrun ~ quintiles -- dispersion: ~1*					
Parameter (Project size group)	**Coefficient**	**SE**	**99% Confidence Interval**	***z***	***p***
Fixed effects:					
Very Small (intercept)	1.5035	0.0489	(1.3828, 1.6348)	12.5503	< .001
Small	0.8254	0.0367	(0.7362, 0.9255)	-4.3201	< .001
Medium	1.0901	0.0447	(0.9808, 1.2116)	2.1023	0.036
Large	1.0777	0.0439	(0.9703, 1.1970)	1.8369	0.066
Very Large	0.8918	0.0381	(0.7990, 0.9955)	-2.6822	0.007
Dispersion:					
(Intercept)	0.8900		(0.8002, 0.9899)		

For consistency and to avoid hand-picking from the near-best models, we retain the best BIC model as the main GLMM result. It also keeps the division of our analysis clear: the GLMM compares typical schedule overrun, while the Pareto analysis studies the upper tail.

Re-baselining the best BIC schedule-overrun model in Table A4.3 confirms the mixed pattern: medium and large projects are similar to each other, small and very large projects are lower, and there is no monotonic increase from smaller to larger projects. Thus, the model does not support the hypothesis (H2) that schedule risk increases systematically with project size.

Table A4.3. Pairwise comparison of typical schedule overrun between project size groups, estimated with the GLM. Each cell shows the estimated typical schedule overrun of the column group relative to the row group. Values below 1 indicate lower overrun than the row group (shaded in green); values above 1 indicate higher overrun (shaded in red). Bold values indicate high statistical evidence ($p < 0.001$).

	Very Small	**Small**	**Medium**	**Large**	**Very Large**
Very Small	1	**0.83**	1.09	1.08	0.89
Small	**1.21**	1	**1.32**	**1.31**	1.08
Medium	0.92	**0.76**	1	0.99	**0.82**
Large	0.93	**0.77**	1.01	1	**0.83**
Very Large	1.12	0.93	**1.22**	**1.21**	1

Diagnostic plots showed no strong remaining trend in the residual-versus-fitted plots for the best BIC GLMMs. The Q-Q plots confirmed that the log-normal GLMMs do not fully capture the far upper tail, which is why tail risk is analysed separately with Pareto-type models.

A4.2. Pareto tail analysis for schedule overrun

The schedule tail analysis follows the same Clauset-type procedure as in Appendix 3. For the likelihood comparisons, we use a smaller cut-off $x_{\min} = 1.3$ due to a narrower range of schedule overrun ratios compared to cost overrun.

Figure A4.1 and Tables A4.4 and A4.5 present the selection-aware Pareto Type I tail-fit summary. Power-law tails are plausible for very small, small, large, and very large projects, but not for medium projects. Conservative estimation and considerably less data for schedule

overrun cause wider confidence intervals. Among the groups where a Pareto tail is plausible, very small projects have the lowest α and thus the fattest tail. This does not support H2. If anything, the tail evidence again points away from larger projects being systematically riskier.

Figure A4.1. Pareto I fit result for schedule overrun by project size group and for all IT projects combined. Markers show the estimated values of α and $x_{\min}$ reported in Table 6. Vertical lines indicate 99% confidence intervals for α, and horizontal lines represent uncertainty in $x_{\min}$.

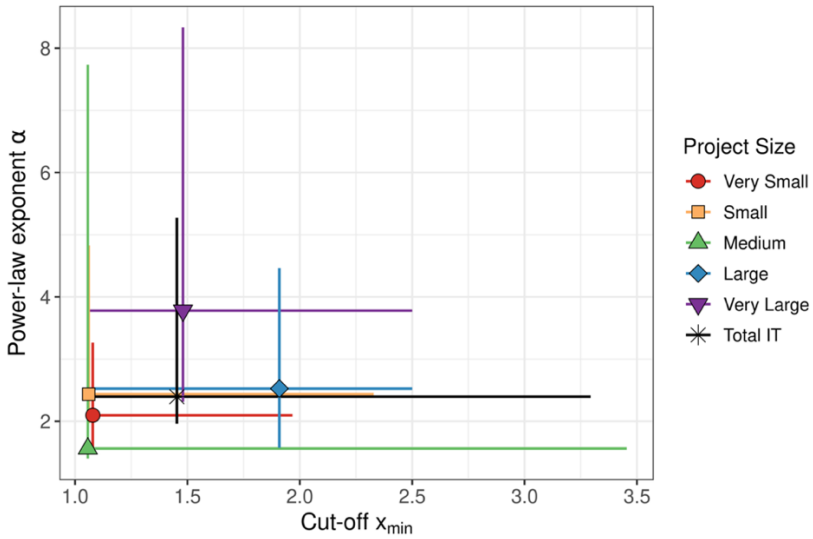


Table A4.4. Selection-aware Pareto Type I tail-fit summary for schedule overrun. The table reports the estimated lower cut-off x_{min}, Pareto exponent α, number of observations above the cut-off, and semi-parametric bootstrap goodness-of-fit p-value. The final column gives the log-Weibull tail classification.

Project size group	***N***	$X_{\min}$	α	N_{tail}	**GoF *p***	**Plausibility**	**Log-Weibull tail class**
Very Small	69	1.079	2.097	67	0.747	power-law plausible	power-law / super-heavy plausible
Small	84	1.061	2.436	83	0.114	power-law plausible	power-law / super-heavy plausible
Medium	105	1.057	1.562	105	0.000	power-law not plausible	not plausible
Large	114	1.909	2.527	37	0.838	power-law plausible	power-law / super-heavy plausible
Very Large	91	1.480	3.780	56	0.390	power-law plausible	plausible, but $\beta > 1$
Total IT	463	1.453	2.397	279	0.095	power-law not plausible	plausible, but $\beta > 1$

Table A4.5. Selection-aware bootstrap confidence intervals for the Pareto Type I tail fits for schedule overrun. Confidence intervals are based on 5000 bootstrap samples and account for uncertainty in both x_{min} and α.

Project size group	$x_{\min}$ lower	$x_{\min}$	$x_{\min}$ upper	α lower	α	α upper
Very Small	1.057	1.079	1.967	1.550	2.097	3.265
Small	1.053	1.061	2.329	1.970	2.436	4.828
Medium	1.057	1.057	3.455	1.402	1.562	7.732
Large	1.051	1.909	2.500	1.573	2.527	4.464
Very Large	1.057	1.480	2.500	2.315	3.780	8.333
Total IT	1.057	1.453	3.294	1.961	2.397	5.272

Table A4.6 compares Pareto I with alternative models. Lighter models fit better only for the medium group, consistent with the goodness-of-fit result that Pareto I is not plausible for medium projects. Other groups do not show a pattern of increasing tail risk with project size.

Table A4.6. Comparison of Pareto Type I tail fits with alternative tail distributions for schedule overrun at the common cut-off $x_{min} = 1.3$. Alternatives are log-normal, exponential, Weibull, and log-Weibull. ΔAIC and ΔBIC are relative to Pareto Type I; negative values favor the alternative, positive values favor Pareto Type I.

Project size group	$x_{\min}$	N_{tail}	Alternative	LR statistics	p (1-sided)	p (2-sided)	ΔAIC	ΔBIC
Very Small	1.3	45	lognormal	-0.232	0.592	0.816	1.891	3.697
Very Small	1.3	45	exponential	1.329	0.092	0.184	7.357	7.357
Very Small	1.3	45	weibull	0.160	0.437	0.873	2.228	4.035
Very Small	1.3	45	logweibull	-	-	-	1.984	3.790
Small	1.3	49	lognormal	-0.938	0.826	0.348	-0.171	1.721
Small	1.3	49	exponential	-0.767	0.778	0.443	-2.661	-2.661
Small	1.3	49	weibull	-0.981	0.837	0.327	-0.790	1.102
Small	1.3	49	logweibull	-	-	-	1.814	3.706
Medium	1.3	85	lognormal	-2.145	0.984	0.032	-11.110	-8.668
Medium	1.3	85	exponential	-2.803	0.997	0.005	-15.032	-15.032
Medium	1.3	85	weibull	-2.299	0.989	0.022	-13.727	-11.284
Medium	1.3	85	logweibull	-	-	-	-1.737	0.705
Large	1.3	86	lognormal	-0.107	0.542	0.915	1.927	4.381
Large	1.3	86	exponential	1.076	0.141	0.282	25.767	25.767
Large	1.3	86	weibull	0.372	0.355	0.710	3.489	5.943
Large	1.3	86	logweibull	-	-	-	-0.800	1.654
Very Large	1.3	65	lognormal	-1.852	0.968	0.064	-5.059	-2.885
Very Large	1.3	65	exponential	-2.425	0.992	0.015	-6.215	-6.215
Very Large	1.3	65	weibull	-2.002	0.977	0.045	-4.427	-2.252
Very Large	1.3	65	logweibull	-	-	-	-5.671	-3.496
Total IT	1.3	330	lognormal	-1.364	0.914	0.173	-6.898	-3.099
Total IT	1.3	330	exponential	0.629	0.265	0.529	17.976	17.976
Total IT	1.3	330	weibull	-1.242	0.893	0.214	-6.090	-2.291
Total IT	1.3	330	logweibull	-	-	-	-5.491	-1.692

Tables A4.7 and A4.8 repeat the alternative-model checks at KS-optimized cut-offs for each model. These checks support the same conclusion. The medium group remains the main non-Pareto case, while the other size groups are compatible with power-law or related heavy-tailed behavior. The schedule-overrun results are therefore less extreme and less clean than the cost-overrun results, but they do not support increasing risk with project size.

Table A4.7. Parameter estimates for Pareto Type I (Par-1), log-normal (LN), and log-Weibull (LW) tail fits for schedule overrun, with each model fitted at its own KS-optimized lower cut-off x_{min}. For Pareto Type I, the fit result is α. For the log-normal, the fit result gives the log-mean μ and log-scale standard deviation σ. For the log-Weibull, the fit result gives θ and β. Brackets give bootstrap confidence intervals.

Project size group	N	**Par-1** $x_{\min}$	**Par-1** N_{tail}	**Pareto-1 fit result**	**LN** $x_{\min}$	**LN** N_{tail}	**Log-normal fit result**	**LW** $x_{\min}$	**LW** N_{tail}	**Log-Weibull fit result**
Very Small	69	1.0787	66	$\alpha = 2.066$ [1.533, 2.925]	1.0569	68	μ = -0.859 [-1.066, 0.418]; $\sigma = 0.944$ [0.439, 1.15]	1.1100	64	$\theta = 0.464$ [0.32, 0.647]; $\beta = 0.97$ [0.768, 1.278]
Small	84	1.0607	82	$\alpha = 2.407$ [1.852, 3.268]	1.2850	49	μ = -0.032 [-0.659, 0.653]; $\sigma = 0.652$ [0.307, 0.952]	1.0607	82	$\theta = 0.43$ [0.317, 0.559]; $\beta = 1.094$ [0.902, 1.423]
Medium	105	1.0566	104	$\alpha = 1.547$ [1.219, 2.035]	1.0566	104	$\mu = 0.384$ [-0.498, 0.678]; $\sigma = 0.635$ [0.437, 0.967]	1.0566	104	$\theta = 0.7$ [0.568, 0.845]; $\beta = 1.342$ [1.12, 1.674]
Large	114	1.9091	35	$\alpha = 2.39$ [1.608, 3.909]	1.7097	42	μ = -0.37 [-0.556, 0.952]; $\sigma = 0.89$ [0.338, 1.111]	2.0435	31	$\theta = 0.38$ [0.211, 0.609]; $\beta = 0.912$ [0.672, 1.416]
Very Large	91	1.4800	55	$\alpha = 3.713$ [2.708, 5.412]	1.0851	84	$\mu = 0.302$ [-0.394, 0.51]; $\sigma = 0.419$ [0.285, 0.666]	1.1379	77	$\theta = 0.464$ [0.374, 0.56]; $\beta = 1.528$ [1.254, 1.965]
Total IT	463	1.4531	278	$\alpha = 2.388$ [2.052, 2.805]	1.6857	189	μ = -0.179 [-0.333, 0.655]; $\sigma = 0.786$ [0.48, 0.901]	1.0952	423	$\theta = 0.542$ [0.485, 0.601]; $\beta = 1.226$ [1.121, 1.356]

Table A4.8. Bootstrap goodness-of-fit comparison for Pareto Type I, log-normal, and log-Weibull tail fits for schedule overrun, with each model fitted at its own KS-optimized lower cut-off. The log-normal boundary hits column reports the share of bootstrap fits in which the log-normal fit reached one of the chosen parameter bounds. High values indicate unstable log-normal fits, even if the goodness-of-fit p-value is acceptable.

Project size group	***N***	**Par-1 GoF *p***	**Pareto-1 verdict**	**LN GoF *p***	**LN verdict**	**Log-Normal run-away rate**	**LW GoF *p***	**LW verdict**	**Log-Weibull tail class**
Very Small	69	0.997	power-law plausible	0.485	log-normal plausible	0.452	0.999	log-Weibull plausible	power-law / super-heavy plausible
Small	84	0.596	power-law plausible	0.570	log-normal plausible	0.254	0.519	log-Weibull plausible	power-law / super-heavy plausible
Medium	105	0.029	power-law not plausible	0.099	log-normal not plausible	0.002	0.091	log-Weibull not plausible	not plausible
Large	114	0.626	power-law plausible	0.602	log-normal plausible	0.447	0.914	log-Weibull plausible	power-law / super-heavy plausible
Very Large	91	0.721	power-law plausible	0.132	log-normal plausible	0.002	0.265	log-Weibull plausible	plausible, but $\beta > 1$
Total IT	463	0.383	power-law plausible	0.840	log-normal plausible	0.447	0.104	log-Weibull plausible	plausible, but $\beta > 1$

A4.3. Summary for schedule overrun

The schedule-overrun robustness checks support the same general conclusion as the main text. The evidence does not show that schedule risk increases with project size. The GLMM results are mixed rather than monotonic, and the tail results show that among the Pareto-plausible groups, very small projects have the fattest fitted tail. Schedule risk is less extreme than cost risk and the evidence is less clean, but it still goes against H2.