Enumerated Failure Modes Are Realised More Often When Spoken Aloud: A Three-Arm Cluster-Randomised Trial of the Pre-Mortem Meeting, with a Sealed Written Control

Ottoline Vasseur[1] Bertil Haugsrud[2]

[1]The Enumeration Registry, Dunkeld

[2]Department of Delivery Assurance, Kvitsøy Institute of Technology

o.vasseur@enumeration-registry.org

Received 22 May 2026; revised 4 July 2026; accepted 29 July 2026

Abstract

In every family somebody says: don't jinx it. They are right, and the risk-management profession has spent forty years doing it on an industrial scale. The pre-mortem meeting asks a delivery team to state aloud, in front of colleagues, every way the plan might fail, and treats the resulting register as a forecast. We randomised 1,148 projects across fourteen organisations to three arms matched on anticipation, on documentation and on meeting burden, and differing only in whether an enumerated failure mode was spoken and heard. Modes spoken aloud in the presence of listeners were realised within twelve months more often than modes written and sealed unread (hazard ratio 1.61, 95% CI 1.34 to 1.94). The hazard rose with the specificity of the wording, by 1.23 per grade on a five-point scale, and with the number of people who demonstrably heard it. Attendance, once audibility was conditioned upon, did nothing. Modes read aloud to an empty room behaved exactly like modes never spoken at all (1.02). Failures that no register had named were unaffected. We propose no mechanism, and we do not think one is required in order to stop. Enumerate in writing. Seal the register. Circulate it as a document and read it to nobody. A risk register is an order form.

Keywords: pre-mortem; prospective hindsight; risk registers; cluster-randomised trial; utterance.

1 Introduction

Here is the practice, as it stands. A delivery team is brought into a room and told that the plan has failed. Not that it might fail: that the year is over, the programme is dead, and the only question left is what killed it. Each person writes down the causes, and then — this is the step that matters, and it is the step every facilitation guide insists upon — each person reads their causes out to the room. The facilitator consolidates. A register is produced. The register is circulated, reviewed at gate meetings, and treated, reasonably enough, as a forecast.

The practice is thirty years old, taught in every programme-management curriculum we can find, and defended on grounds that are difficult to argue with. People generate more concrete failure explanations when the failure is stated as accomplished than when it is stated as possible [7, 4]. Saying the causes aloud surfaces the concerns that the quiet members of a team would otherwise carry home [5]. The register that results is longer, sharper and better covered than the one produced by any other elicitation method [2]. All of this is true, and none of it is in dispute here.

Here is the other thing, which is also true. In every household the reader has ever sat in, somebody has said *don't jinx it*. The score is not spoken before the whistle. The uneventful journey is not remarked upon until the car is parked. The offer is not described as accepted until it is signed, and the person who describes it that way is told to stop. The ethnographic record puts the injunction in forty-one societies and notes that what is proscribed is with remarkable consistency the *naming*, never the thinking [1]. It is filed under superstition and has stayed filed.

These two positions are in flat contradiction, and so far as we can establish nobody has remarked on it. One of them lives in a management textbook and the other lives at a bus stop.

The contradiction is decidable, and the observational literature already contains the answer with the wrong sign attached to it. Register richness is a durable positive predictor of subsequent programme failure. The projects with the longest and most diligently maintained risk registers fail more often, and the margin does not shrink under adjustment for size, sector, budget or duration [3, 6]. The literature treats this as reverse causation — troubled projects write more — and the reading is a sensible one. It has also never been tested. To test it you must randomise the meeting.

We randomised the meeting. What follows is a three-arm cluster-randomised trial in 1,148 delivery projects, in which every element of the pre-mortem was held constant across arms except one: whether a written failure mode was spoken, and whether anyone heard it.

2 Related work

Prospective hindsight. The generative effect is well established. Instructing a participant that an outcome has occurred, rather than that it may occur, increases both the number and the concreteness of the causes they produce [7], an effect that replicates across domains and survives incentive manipulation [4]. Note the dependent variable throughout: what the participant *says*. No study in this tradition follows the plan.

The pre-mortem as procedure. The meeting was formalised as a management instrument in the late 1990s [5] and given the facilitation protocol still in general use six years later [2]. Both papers, and the two dozen that follow them, measure the register: item count, specificity, coverage of a reference hazard taxonomy, and participant report of psychological safety. The measured object is the document. The project is not an outcome in this literature.

Register efficacy. Where projects have been followed, they have been followed observationally. The association is consistent and has the same sign everywhere it has been looked for: richer registers, worse outcomes [3]. The standard reading is reverse causation, argued most carefully by Nyström [6], who also observes that the association is stronger for specific items than for vague ones and treats that gradient as further evidence of the same confounding — a team in trouble writes in detail. The gradient is real. We recover it below under randomisation, where reverse causation cannot produce it.

The injunction. Aldreth's comparative survey [1] remains the standard reference: apotropaic speech taboos in forty-one societies, with the structural convergence noted above. It is an anthropology paper. It has been cited fourteen times, never outside anthropology, and never as a hypothesis.

3 The sealed-enumeration trial

3.1 Design

A multicentre, three-arm, cluster-randomised, parallel-group, actively controlled, outcome-assessor-blinded pragmatic trial. The unit of randomisation is the project, because the exposure is a meeting and a meeting cannot be allocated to some of the people in the room [9]. The trial is reported to the cluster-randomised extension of the standard reporting statement [10].

Fourteen organisations across four jurisdictions and six sectors — civil infrastructure, public-sector information technology, pharmaceutical scale-up, broadcast and live events, municipal transport, offshore energy — contributed projects with budgets from €0.4M to €710M. Enrolment ran from January 2023 to March 2025, with follow-up to March 2026. Randomisation was central, computer-generated, stratified by sector and budget band, and concealed until baseline data were locked. The protocol and the analysis plan were lodged with the registry on 14 November 2023 [12], during enrolment and before any outcome had been adjudicated.

3.2 Arms

Projects were allocated 2:1:2.

Arm S (spoken). The standard pre-mortem, run to the published protocol [2]. Individual written enumeration, then each participant reads their modes aloud in turn, then consolidation.

Arm V (voiced, solitary). Individual written enumeration, then each participant reads their modes aloud — alone, in an acoustically isolated booth, into a recorder. The sentences are produced. No person hears them.

Arm E (sealed). Individual written enumeration. Sheets are collected face down, sealed, and couriered unread to an offsite registry. No failure mode is uttered in the room, and none is uttered by the facilitator.

3.3 What is held constant

The design's load is carried by what the arms share, so the list is worth stating. All three arms convene the same people for the same scheduled duration, under the same trained facilitator, working from an identical script up to the enumeration step and an identical script after it. All three produce individually written failure modes. All three receive a consolidated register nine working days later. That register is typeset by the offsite registry, identical across arms in format, length distribution and delivery channel, and circulated to the project team and its steering board as a document.

Teams in arms V and E were instructed, in writing, not to read the register aloud. Compliance was audited (§8).

The arms are therefore matched on anticipation: everyone enumerates. They are matched on documentation: everyone ends up holding the same list. They are matched on meeting burden: everyone loses the same afternoon. Where they differ is that in arm S a human being heard the sentence, in arm V a sentence was spoken to nobody, and in arm E no sentence was spoken.

3.4 Facilitators

Facilitators cannot be blinded to allocation, since they run the meeting. Thirty-one facilitators were used, each rotated across all three arms, each blinded to outcome, and each without contact with the project after the index meeting. Facilitator enters the models below as a shared frailty; the variance component is small (ICC 0.011) and did not materially alter any estimate.

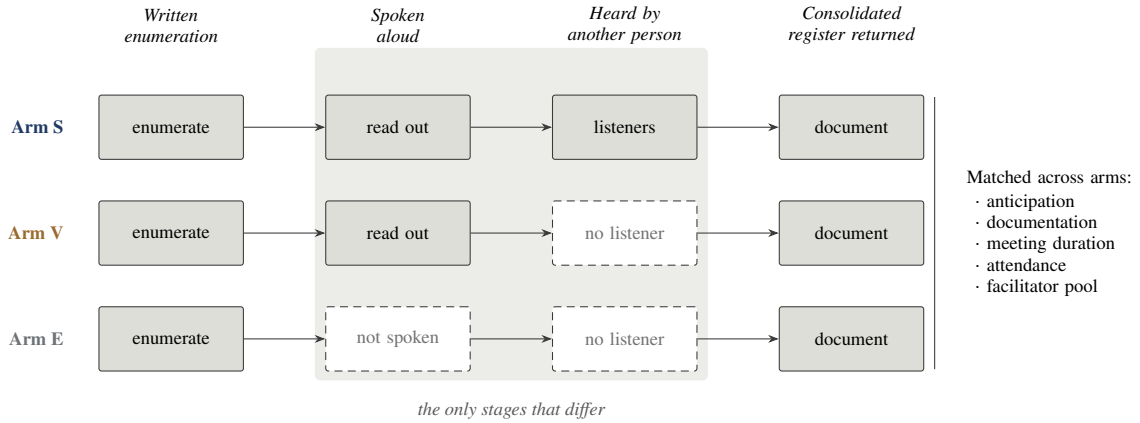

Figure 1: Structure of the three arms. Every stage but two is common to all arms: each arm enumerates in writing, and each receives the same consolidated register as a document nine working days later. Arm V produces the sentence and supplies no listener. Dashed boxes mark a stage that does not occur.

4 Outcomes and adjudication

4.1 Primary outcome

Time from the index meeting to the first adjudicated *realisation* of an enumerated failure mode, within twelve months.

A mode counts as realised when a documented event satisfies the actor, the mechanism and the artefact stated in the enumeration, within a tolerance of ±25% on any quantity and ±6 weeks on any date. Where a match is ambiguous, the strictest available reading is scored, which is to say the reading least favourable to realisation; this follows established practice for adjudicating composite outcomes whose components are only partly specified in advance [8]. Criteria were fixed in the registered protocol and not amended.

Adjudication was by two independent assessors working from the register and the project's incident record with all allocation-identifying material redacted, including meeting minutes and attendance sheets. This redaction is possible only because every arm produces the same document type — another reason for returning a register to arm E. A third assessor resolved disagreements. Agreement was substantial ($\kappa = 0.79$).

4.2 Secondary and falsification outcomes

Cumulative incidence of any enumerated mode at twelve months; count of modes realised per project; project cancellation or descoping, treated as a competing event.

The falsification outcome is the one we would ask of anybody reporting a result like this. Every incident in every project's record was coded, offsite and after database lock, against the union of all registers in the trial. Incidents matching no enumeration in any arm form the *unnamed* failure set. If the exposure is really that arm S contains worse projects, unnamed failures must move with arm. If the exposure is utterance, unnamed failures cannot move at all.

Table 1: Baseline cluster characteristics.

	Arm S $n = 459$	Arm V $n = 230$	Arm E $n = 459$
Budget, median (€M)	8.4	8.1	8.6
Planned duration, median (mo)	19	18	19
Team size, median	24	23	24
Attendance at index meeting	11.2	10.9	11.3
Modes enumerated, mean	8.9	8.6	8.7
Prior pre-mortem experience	61.4%	60.9%	62.1%
Civil infrastructure	22.4%	22.2%	22.7%
Public-sector IT	27.7%	27.8%	27.5%
Pharmaceutical scale-up	12.2%	12.6%	12.0%
Broadcast and live events	14.6%	14.3%	14.4%
Municipal transport	13.1%	13.0%	13.3%
Offshore energy	10.0%	10.0%	10.2%

5 Statistical methods

The primary analysis is a Cox proportional-hazards model for time to first realisation, stratified by sector, with cluster-robust variance at the organisation and a shared frailty on facilitator. For project i in organisation j under facilitator k,

$$\lambda_{ijk}(t) = \lambda_{0c(i)}(t)\, u_k \exp\!\left(\beta_S S_i + \beta_V V_i + \gamma^{\top} z_i\right), \qquad (1)$$

with $c(i)$ the sector stratum of project i, $u_k \sim \Gamma(\theta^{-1}, \theta)$, z_i the pre-specified baseline covariates of Table 1, and arm E the reference. Proportional hazards were assessed on scaled Schoenfeld residuals; the global test was not significant ($p = 0.34$).

Cancellation or descoping before any realisation is a competing event and was treated as such rather than as censoring. Subdistribution hazards [11] are the pre-specified sensitivity; cause-specific hazards are reported alongside.

The dose–response analyses operate at the level of the individual utterance, with a project-level frailty v_i. For mode

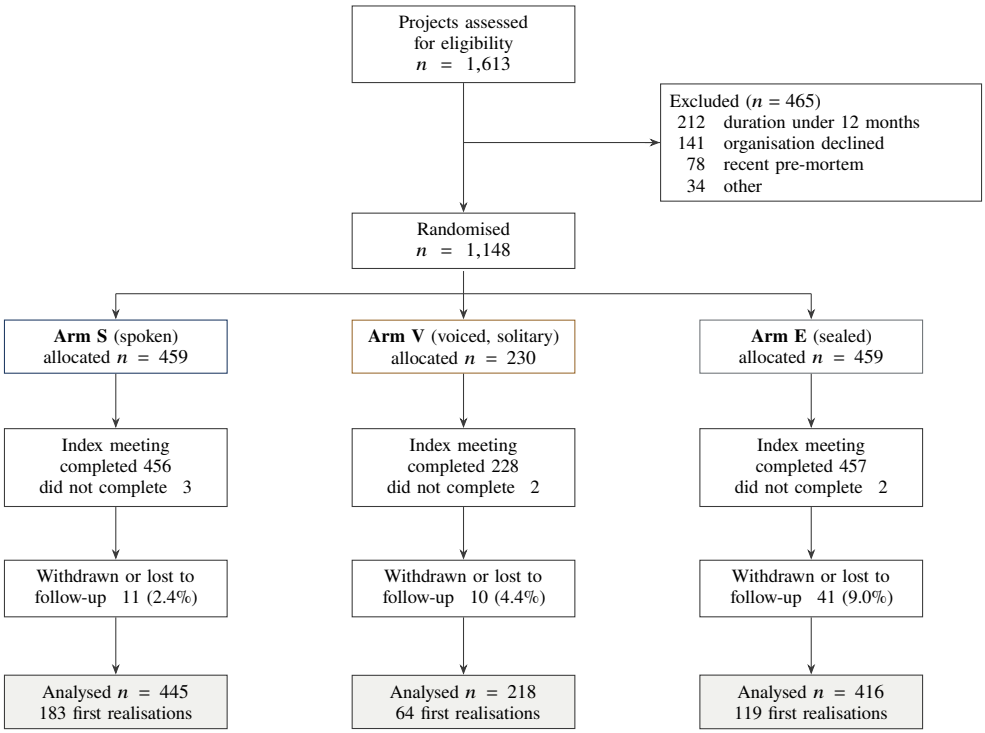

Figure 2: Cluster flow. Attrition after randomisation was differential (9.4% of arm-E clusters against 3.1% of arm-S clusters, counting failures to complete the index meeting); it is addressed in §8.

m in project i,

$$\lambda_{im}(t) = \lambda_0(t)\, v_i \exp\!\big(\beta_g\, g_{im} + \phi \log_2 a_{im} + \psi n_i + \gamma^{\top} z_i\big), \quad (2)$$

where $g_{im} \in \{1, \dots, 5\}$ is the specificity grade, a_{im} the received audience, and n_i the meeting attendance. Entering $\log_2 a$ and n simultaneously is the point of the specification: n is what a manager can observe, and a is what varies.

Specificity. Two coders, blinded to arm and to outcome, graded every enumerated mode on a five-point scale running from a bare hazard class (grade 1; *something with the vendor*) to a named actor, mechanism, artefact, date and quantity (grade 5; *the vendor's Tallinn office ships the billing adapter three weeks after the March gate, costing two sprints*). Agreement was $\kappa = 0.84$. Grade enters both as an ordinal trend and as five indicators, so that departure from linearity is testable.

Received audience. The trial ran through the hybrid-meeting period, which supplies variation in who actually heard a given sentence. For each utterance in arm S we recovered, from conference-platform telemetry and per-participant inbound audio logs, the number of persons whose client received and rendered the audio stream over that interval. Reception failures — muted inbound, dropped connection, late join, an utterance falling inside a participant's own microphone-check window — are set by network and client conditions and not by what was about to be said. We checked this rather than asserting it: specificity grade is balanced across audibility strata ($p = 0.71$), as are speaker seniority, agenda position and utterance duration. Received audience therefore varies within a single meeting, across sentences from a single register, and the same room supplies both the exposed and the unexposed.

The within-project contrast. Facilitators work down the enumeration list in a centrally randomised order and meetings are time-boxed, so in arm S some modes are reached before time expires and some are not. Within a project, *reached* is therefore assigned at random, holding constant the team, the plan, the sector, the budget, the facilitator and every unmeasured project-level characteristic whatever. This is the analysis that project selection cannot reach.

Multiplicity and power. Hierarchical gatekeeping: primary, then the two dose–response tests, then subgroups, which are reported as interaction tests and described as exploratory. A target of 1,400 clusters gave 0.90 power against HR 1.45 at $\alpha = 0.05$, assuming twelve-month control incidence 0.29 and ICC 0.03. Enrolment closed at 1,148, giving 0.81.

6 Results

6.1 Primary

Of 1,613 projects assessed, 1,148 were randomised and 1,079 analysed (Figure 2). Baseline characteristics were

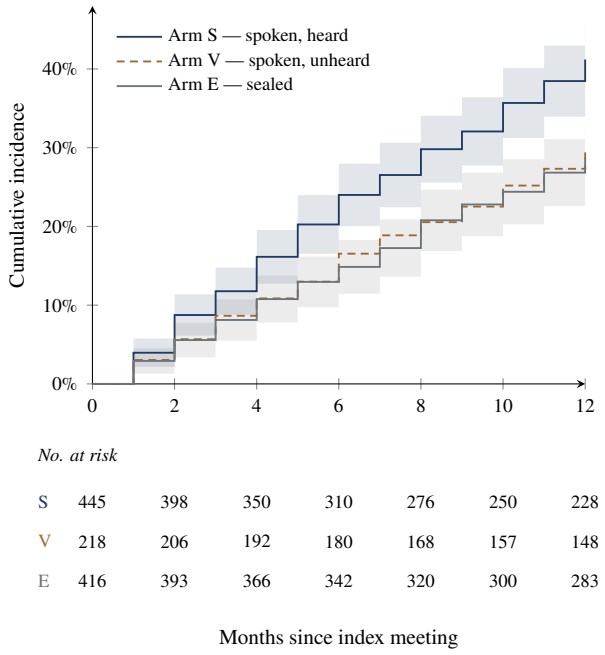

Figure 3: Cumulative incidence of first realisation of an enumerated failure mode, by arm (Aalen–Johansen, cancellation treated as a competing event). Shaded bands are pointwise 95% intervals for arms S and E. Arm S separates by month three. Arms V and E overlap throughout.

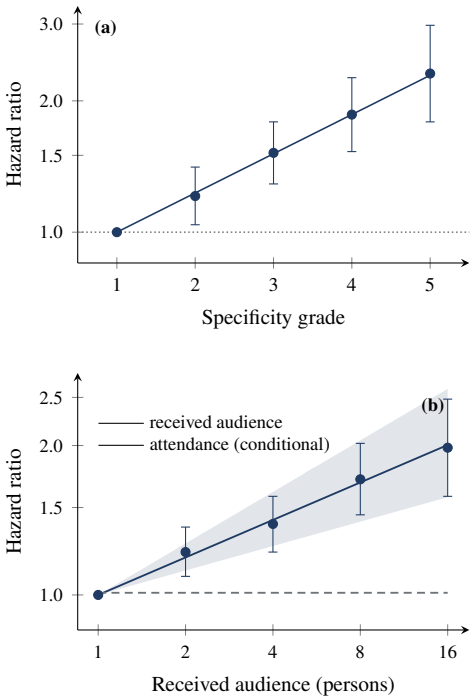

Figure 4: Dose–response. (a) Hazard ratio by specificity grade, arm S, grade 1 as reference; the line is the fitted ordinal trend, 1.23 per grade. (b) Hazard ratio against the number of people who demonstrably received the utterance, on a $\log_2$ axis, with the fitted 1.19 per doubling and its 95% band; the dashed line is the effect of meeting attendance entered simultaneously.

balanced (Table 1). There were 366 first realisations over 10,154 project-months.

Failure modes spoken aloud in the presence of listeners were realised more often than modes written and sealed unread: HR 1.61 (1.34–1.94), $p < 0.001$. Twelve-month cumulative incidence was 41.2% in arm S against 28.7% in arm E. The subdistribution analysis was materially identical (sHR 1.56, (1.30–1.88)), as was the count outcome (IRR 1.54, (1.33–1.78)). Cancellation, the competing event, did not differ by arm (sHR 1.09, (0.88–1.35)).

6.2 Arm V, and what it removes

Modes read aloud in an empty booth were realised at the rate of modes never spoken: HR 1.02 (0.81–1.28), $p = 0.87$, against arm E. Twelve-month incidence was 29.4% and 28.7%. In Figure 3 the arm-V and arm-E curves are difficult to separate by eye, and that is the finding.

This is the comparison that decides what the primary result is about. Arm V speaks. Arm V produces the sentence, with the same wording, the same duration, the same breath. What arm V lacks is a listener. Whatever the exposure is, it is not production.

6.3 Dose–response

The hazard rose monotonically with specificity: HR 1.23 (1.17–1.30) per grade, with no detectable departure from linearity ($p = 0.44$), and 2.31 (1.79–2.98) for grade 5 against grade 1 (Figure 4a).

The hazard rose log-linearly with the number of people who heard the sentence: 1.19 (1.12–1.27) per doubling of received audience. Meeting attendance, entered simultaneously, was flat: 1.01 (0.94–1.09) (Figure 4b). The variable that acts is not the variable that is visible on the invitation.

6.4 Falsification

Failures matching no enumeration in any register did not differ between arms: HR 0.98 (0.86–1.12). Arm S projects did not fail more at the things nobody had named. They failed more at the things that had been said in them.

6.5 Within-project contrast

Within arm S alone, modes reached before the time-box expired were realised at 18.4%; modes written by the same people, in the same meeting, in the same register, but not reached, at 11.9% (Table 2). HR 1.58 (1.31–1.91). Reading order was centrally randomised, so no project-level characteristic can account for the contrast.

6.6 Subgroups

Eleven pre-specified hazard classes (Figure 5). The effect is present in ten, with point estimates between 1.44 and 1.83

Table 2: Within-project contrast, arm S only. Reading order was centrally randomised and meetings were time-boxed.

	Modes	Realised	12-mo %
Reached before time-box	2,612	480	18.4
Not reached	1,349	160	11.9
HR 1.58 (1.31–1.91), $p < 0.001$			

and no significant heterogeneity among them.

It is absent in the eleventh. Environmental and meteorological failure modes — flood, unseasonal ground conditions, wind-limit standdowns, weather-driven schedule loss — gave HR 1.04 (0.79–1.37), interaction $p = 0.008$. The base rate in that class is 0.31, which is mid-range for our taxonomy, so a ceiling effect does not account for it.

Effects did not differ by sector, budget band, jurisdiction, team size, or prior pre-mortem experience.

7 Discussion

Three arms, and the third one narrows the claim considerably. A two-arm trial would have shown that speaking is worse than writing, which is a loose result compatible with meeting fatigue, with attention, with a dozen things. Arm V removes them. The sentence was produced in arm V. It was produced with the same wording, by the same kind of person, in the same week of the project. What was missing was somebody to hear it, and without that the effect is gone entirely.

The observational literature, read back with this in hand, had the sign right and the arrow backwards. Register richness predicts failure [3]; it predicts failure more strongly for specific items than for vague ones [6]; and both of those hold under randomisation, where a troubled project cannot reach back and write its own register. The reverse-causation reading was reasonable and is now unnecessary.

One of the eleven hazard classes does not behave like the other ten. Environmental and meteorological modes gave 1.04, and the interaction is not attributable to sampling noise. We can offer no principled reason why an enumerated hazard should be exempt on the grounds of being meteorological. We report it because it is there.

The practical consequences run against every register-quality standard we are aware of. Those standards prize grade-5 wording and treat grade-1 vagueness as a defect to be coached out of participants. On our estimates grade-1 vagueness is protective, by a factor of 2.31 relative to the wording that scores best on current guidance. The better a register is by the criteria in use, the more of its contents arrive.

We propose no mechanism. The omission is deliberate. We have no candidate, we do not find the available ones credible, and we will not manufacture one to make the result easier to accept. Nothing in the analysis depends on having one. A procedure adopted to prevent an outcome, and shown under randomisation to produce it, does not get to continue while the mechanism is argued about. It gets suspended, and the argument continues without it.

What we would do, on this evidence, is narrow. Keep the enumeration: the anticipation is not the problem, and arms V and E anticipated as thoroughly as arm S. Keep the register: arm E received one and had the lowest incidence in the trial. Change the middle step only. Write the modes down. Seal them. Have them typeset offsite and returned as a document. Circulate the document, review it at gates, act on it, and do not at any point read it out.

8 Limitations

The registered protocol required an enumeration of anticipated threats to validity [12]. Three were named, and all three occurred.

Differential attrition. Cluster attrition was 9.4% in arm E against 3.1% in arm S, in the direction that would inflate our estimate if the departing arm-E clusters were failing ones. Inverse-probability-of-censoring weighting on the observed attrition predictors gave HR 1.57 (1.29–1.91), and a tipping-point analysis requires the departing clusters to have failed at 2.4 times the retained arm-E rate before the primary result crosses unity.

Contamination. In 14.2% of arm-E projects an audit identified a post-meeting event at which a sealed mode was spoken aloud, most often by a participant briefing a colleague who had not attended. Contamination of this kind biases toward the null, so it cannot manufacture the primary result. It can distort the magnitude, and the audit recovers only utterances that left a documented trace, so 14.2% is a floor and not an estimate.

Under-recruitment. Enrolment closed at 1,148 clusters against a target of 1,400, reducing power from 0.90 to 0.81. The primary comparison was adequately powered. The subgroup analyses were not, and are reported as exploratory.

Generalisability is limited to funded delivery projects with a formal steering board, in four jurisdictions, in organisations willing to be randomised. We have no data on unfunded work, on domestic plans, or on any setting in which the enumeration is not written down first.

The threats-to-validity enumeration was read aloud to the trial steering committee at its inaugural meeting, and therefore falls within the scope of the exposure under study. No valid adjustment is available and none was attempted.

9 Conclusion

The practice had never been tested against the alternative of not speaking. It has been now, and the register turns out to be a poor forecast of the year and a good description of it. The effect requires a listener, scales with how exactly the thing was put, and does not touch the failures nobody mentioned.

We are not in a position to tell any organisation how to

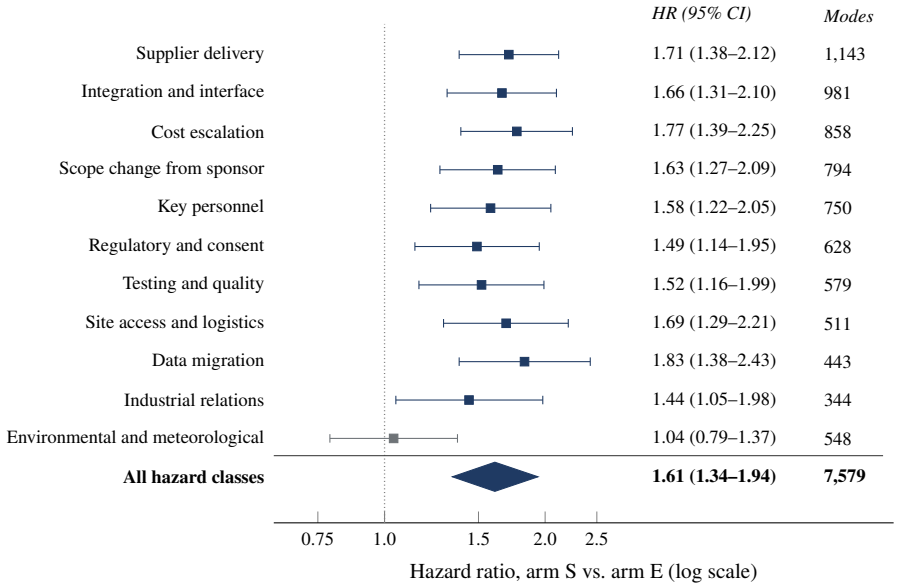

Figure 5: Mode-level hazard ratios by pre-specified hazard class, arm S against arm E. Ten of the eleven classes lie between 1.44 and 1.83 with no significant heterogeneity among them. The eleventh is environmental and meteorological hazard (interaction $p = 0.008$).

run its programmes, and we are not going to try. We are in a position to say what this trial measured and where it points. Enumerate in writing, seal the sheets, circulate the register as a document, and read it to nobody. A risk register is an order form.

Data availability

Registers, adjudication records and audio-reception logs are held by the Enumeration Registry under a restricted-access agreement. The analysis code is available on request.

Conflicts of interest

The Enumeration Registry is funded by subscription from eleven of the fourteen participating organisations. Neither author holds an interest in any facilitation practice.

References

[1] M. Aldreth. Apotropaic speech and the naming taboo: a comparative survey of forty-one societies. *Journal of Comparative Ethnography*, 12(3):211–244, 1983.

[2] R. Bewick and L. Ostrander. A facilitation protocol for prospective failure enumeration. *Organisational Methods Quarterly*, 19(2):88–107, 2004.

[3] Chelmsford Delivery Assurance Group. Register richness and delivery outcome: ten years of programme records. Technical Report CDAG-2016-04, 2016.

[4] D. Fennimore. Replication and extension of the prospective-hindsight effect under incentive. *Journal of Behavioural Decision Analysis*, 9(1):33–51, 1996.

[5] J. Hault. The pre-mortem: converting anxiety into inventory. *Management Practice Review*, 41(4):17–29, 1998.

[6] A. Nyström. Why do well-documented projects fail more often? Reverse causation in the risk-register literature. *International Journal of Project Studies*, 28(2):145–169, 2011.

[7] P. Ollerenshaw and T. Vane. Prospective hindsight and the enumeration of outcomes. *Journal of Behavioural Decision Analysis*, 4(2):101–118, 1991.

[8] S. Quaife. Adjudication of composite process outcomes under partial specification. *Trials Methodology*, 7(1):1–14, 2019.

[9] K. Reinholt and B. Sandoval. Cluster-randomised designs when the exposure is a meeting. *Statistics in Applied Delivery*, 11(3):240–262, 2020.

[10] Standing Committee on Trial Reporting. Extension of the reporting statement to cluster-randomised pragmatic trials. *Reporting Standards in Applied Research*, 6:1–22, 2018.

[11] G. Tewkes. Subdistribution hazards for competing terminations in programme cohorts. *Applied Survival Analysis*, 22(4):519–540, 2014.

[12] O. Vasseur and B. Haugsrud. Protocol for a three-arm cluster-randomised trial of enumerative practice in delivery projects. Registry of Delivery Trials, RDT-2023-0714, 2023.