The ORBITAL Outcome-Data Research Agenda

The doctoral contribution thread. This is the program's shared map of the open questions the field needs answered — the questions a dissertation (or a career of dissertations) can be built on. It is organized around the field's single most underused asset: the ***outcome record*** *— the pairing of what organizations predicted with what actually happened.*

Level: doctoral · **Use:** the spine of DL 903 and the source of dissertation questions · **Companions:** course/doctoral-program-overview.md, course/doctoral-empirical-design.md.

---

1. Why a research agenda, and why "outcome data"

A young discipline needs more than individual dissertations; it needs a **map of its own ignorance** — a shared statement of the questions that matter, so a student's original research is also *useful* research, and so a field's scarce research effort does not all land on the same few questions.

The organizing insight is that the pursuit discipline generates, as a byproduct of its own operations, a scientific asset it barely uses. For every pursued opportunity, an organization:

1. **predicted** — it estimated a probability of winning, scored factors, set thresholds;
2. **acted** — it bid or declined, gated or pursued, teamed or soloed;
3. **the market produced an outcome** — won, lost, or withdrawn, with a public award record when won.

The pairing of *prediction* and *outcome* is the discipline's empirical goldmine. It is the raw material of calibration research, factor validation, governance evaluation, and learning research. The agenda below is organized into five threads, each a family of open questions that this pairing can answer.

2. The data environment (what a researcher actually has)

Before the threads, the data reality — because every agenda question is bounded by it:

- **The public record is rich.** Awards with dollar values and winners; solicitations with published review criteria; agency program data; congressional spending records; the paper trail of who won what. This is the ground truth, freely available, lawfully public.
- **The decision side is hidden.** What was estimated, scored, gated, and deliberated is not published. It lives in organizational records and in the discipline's own prediction-and-outcome store (the machine-side record

that logs, for each pursued opportunity, the predicted win probability at decision time and the eventual terminal outcome).

- **The central asymmetry.** Public outcomes + hidden decisions. Every agenda thread is shaped by this: the public record anchors the outcome side, and the hidden decision side must be studied through organizational data, ethically gathered, or through the machine-side record where it exists.
- **The central hazard.** Only pursued bids resolve. The sample of resolved outcomes is chosen by the very decisions under study. Every agenda thread must name this and mitigate it (observation floors, stratification by funder family, a stated withdrawn-as-non-win policy, and bounded claims).

3. The five threads

Thread A — Prediction and calibration

The question the field argues about. Are our probability-of-win estimates any good? Does a 0.7 mean roughly 70%?

The open questions.

- Is the discipline's win-probability estimate *calibrated* — and in which bands is it over- or under-confident? (Reliability diagrams localize the answer.)
- Does calibration hold across **funder families** and **agencies**, or is a composite calibrated in one population miscalibrated in another?
- Does calibration **drift over time** — and does the drift predictably precede a change in the market, in the estimator, or in what is being scored?
- Do **structured composites** calibrate better than **expert judgment**? (A field-experiment question — randomize the estimation method, compare calibration.)
- What is the right **observation floor** — how many terminal outcomes before a calibration claim is trustworthy?

Designs that fit. Calibration evaluation (predicted pWin joined to terminal outcome, reliability bins, Brier score, ECE, PSI); field experiments on estimation method; longitudinal drift studies.

Data needs. The prediction-and-outcome record (predicted pWin at decision time + terminal outcome + funder family + the version of the estimator); for the expert-vs-composite question, a purpose-built field experiment.

Expected contribution. The first systematic calibration evidence for the discipline — the studies that answer "is the doctrine's arithmetic honest?" with data instead of assertion.

Thread B — Factor validation

The question the field argues about. Of the factors that compose a win-probability estimate — past performance, technical capability, price competi-

tiveness, incumbent/relationship advantage, team strength, customer knowledge, strategic alignment — which actually predict wins, and are the weights right?

The open questions.

- **Construct validity:** Does the seven-factor composite predict awards better than a simple gut estimate?
- **Factor weights:** Which factors actually move the probability of winning — and do the discipline's hand-set weights match what the data show? (A logistic regression of won/lost on the factors is the natural instrument.)
- **Reliability:** Do two experienced scorers score the same opportunity similarly? If not, the composite is measuring the scorer, not the opportunity.
- **Cross-agency stability:** Do the factors that predict wins at one agency or funder family hold at another?
- **Thresholds:** Are the discipline's recommendation thresholds (strong bid / bid / conditional / likely no-bid / no-bid) set at the right points?

Designs that fit. Factor-weight validation (factor scores + outcomes, modeled with logistic regression); inter-rater reliability studies; cross-agency comparisons; threshold sensitivity analysis.

Data needs. Predicted factor scores and outcomes for a sample of pursued bids (organizational or machine-side record), plus public outcomes; scoring sessions for reliability studies.

Expected contribution. The empirical answer to the field's oldest internal argument — "which factor matters most?" — replacing seniority-based answers with evidence.

Thread C — Agency win patterns

The question the field argues about. Do agencies' published review criteria predict what they fund — and do recognizable winning patterns vary by agency?

The open questions.

- Do the **published review criteria** of a funding agency predict which proposals it funds? (A content-analysis study: code proposals against the criteria, model funding on the codes.)
- Do the discipline's curated **win patterns** — a named anchor partnership, a quantified outcome target, a sustainability/continuation mechanism, positioning coherence — actually discriminate winners from losers?
- Do the patterns **vary by agency** in the way the field's agency-level corpora claim (DOE, DOL, NSF, NIH, EDA-PWEA and beyond)?
- How much of a winning proposal is **pattern** and how much is **fit** — the unmeasurable match between a specific proposal and a specific agency's mission at a specific moment?

Designs that fit. Pattern-discrimination studies (won vs. lost proposals from published, closed competitions, coded against a codebook with inter-rater relia-

bility, modeled with logistic regression clustered by competition); cross-agency comparisons; qualitative comparative case work on outlier wins.

Data needs. Won and lost proposals (lawfully public for closed competitions), agency review criteria, the discipline's pattern corpora as the coding foundation.

Expected contribution. The first evidence that the field's strategy lore — the pattern cards — is real and not just plausible; and a measurement of how much of a win is attributable to pattern versus fit.

Thread D — Decision and governance

The question the field argues about. Does disciplined decision-making actually pay? Do organizations that gate honestly — decline below-threshold pursuits, write criteria in advance, allocate portfolios deliberately — outperform organizations that chase everything?

The open questions.

- Does **threshold discipline** improve portfolio outcomes? (Firms with honest bid/no-bid gates vs. firms that pursue everything, on conversion and win rate.)
- Does **gate governance** — criteria written before the work, named decision owners — improve the quality of pursuit decisions?
- Does **certification posture** (HUBZone, set-aside eligibility, small-business programs) move win rates in the pools it opens — and is the advantage real or selection?
- Does **portfolio allocation** across risk levels improve expected value — or is the field's risk appetite folklore?
- Does the **bid/no-bid call itself** discriminate — do declined opportunities really resolve worse than pursued ones? (The survivorship hazard is the whole ballgame here.)

Designs that fit. Comparative case studies of firms' governance; difference-in-differences around certification events or governance adoption; regression discontinuity at eligibility thresholds; portfolio-outcome panel studies.

Data needs. Public outcomes + organizational governance records (gate charters, decision logs, scored pursuit sheets); certification and eligibility data.

Expected contribution. The evidence base for the discipline's governance doctrine — "gates protect attention, money, and reputation" — tested against outcomes instead of asserted.

Thread E — Learning and evolution

The question the field argues about. Does the organization actually learn? Does measuring outcomes, feeding them back, and adjusting actually improve performance over time?

The open questions.

- Does **recalibration improve calibration**? (A before/after study of the estimate's Brier score and ECE around a recalibration event.)
- Does the **learning loop** improve organizational outcomes — do firms that measure and feed back win more over time than firms that do not?
- Do organizations **learn from losses** — and what distinguishes the firms that do from the firms that repeat the same mistake?
- Does the **cosmology loop** compound — does past performance predictably raise the next probability of win, in the way the doctrine claims?
- What is the **right boundary between automatic and human-reviewed learning** — when does adjusting a rule from outcomes become dangerous (e.g., a price floor pushed below a lawful level)?

Designs that fit. Before/after and difference-in-differences designs around recalibration or policy events; panel studies of organizational win rates; longitudinal case studies of learning behavior; meta-analytic reviews of the growing evaluation record.

Data needs. The prediction-and-outcome record over time; organizational outcome logs; the recalibration history (what changed, when, and why).

Expected contribution. The empirical test of the discipline's central promise — that the pipeline compounds, that the loop learns — and the first evidence on how fast, and under what conditions, a pursuit organization actually improves.

4. How the threads connect

The five threads are one story. The discipline *estimates* (Thread A: are the estimates honest?), using *factors* (Thread B: are the factors right?), against *agencies* (Thread C: do the criteria predict?), under *governance* (Thread D: does the discipline pay?), and it claims to *learn* (Thread E: does it improve?). A dissertation normally goes deep on one thread; a career can circle the whole map, with each study feeding the next — which is exactly how a young field builds an empirical base.

The threads also connect to the machine side of the discipline: the prediction-and-outcome record, the calibration monitor, the factor composite, and the agency corpora are not just infrastructure — they are **research instruments waiting for researchers**. Each piece of machinery embodies a claim, and each claim is an agenda question.

5. Picking a dissertation from the agenda

A good dissertation question has three properties:

1. **It is researchable.** A specific, falsifiable question a design can answer (DL 901).

2. **It is feasible.** The data are obtainable, the sample is big enough, the ethics are clear (DL 902's power discipline).
3. **It is a contribution.** The field needs the answer, and the answer has not been published (the literature review).

The agenda marks the *open* questions, but the student's own situation picks the one to pursue: access to organizational data points toward the decision and governance threads; comfort with the public record and large-N analysis points toward calibration and agency-pattern research; a professional's own firm is a natural (and ethical) site for a learning-loop case study. The agenda is a menu, not an assignment.

6. The agenda as a living document

This agenda is meant to change. As dissertations are completed, published, and absorbed, questions on this map get answered — they move from *open* to *answered*, and the answers open new questions. The program treats the agenda as a living artifact: every dissertation defense is also a review of the agenda, and every graduate leaves one question answered and — ideally — one new question written onto the map.

The agenda in two sentences

The discipline predicts, and the market produces outcomes — and the pairing is the field's most underused asset. This agenda maps the questions that pairing can answer: whether our estimates are honest, our factors are right, our agency lore is real, our governance pays, and our learning loop actually learns. **Answer one question well, and you have a dissertation. Answer a thread, and you have a career.**

Reflection questions

1. Which thread is closest to your own experience — and what question from it would you most want to answer?
2. For that question, what data could you plausibly obtain, and what is the ethical path to it?
3. What would the field be able to do — differently, better — if your study were published and believed?