Sample Capstone — Empirical Study of the Calibration Engine & the Golden Thread

A complete, worked doctoral capstone: a research design a doctoral candidate could execute to test the discipline's own empirical machinery — is the win-probability estimate calibrated, does the versioned DNA corpus change outcomes, and what explains confidence and success across agents, methods, DNA versions, and opportunity types?

> **What this is.** The assignment briefs for the doctoral tier live in `course/doctoral-program-overview.md` (DL 901–DL 903 and DL 990), and the dissertation sequence is mapped in `course/doctoral-dissertation-sequence.md`. This page is a **model answer** — a sample research design, the deliverable the DL 903 seminar asks for and that grows into the DL 990 dissertation. It is written to be read by a candidate who must design *their own* study and by a committee that must judge it. It is deliberately concept-first: the machinery is studied as *concepts* — what each piece claims, and how you would test the claim — with no lesson requiring the operation of any product (the doctoral tier's stack-agnostic pledge, doctrine/08).
>
> **The research object.** The pursuit discipline has built an empirical apparatus that *makes predictions that can be checked against reality.* That apparatus — and the pairing of *predicted* with *observed* — is the single most underused asset in the field. This sample design shows a candidate how to turn one claim of that apparatus into a dissertation.

Tier: doctoral (research doctorate) · **Course anchor:** DL 903 (seminar research design) → DL 990 (dissertation) · **Format:** individual; supervised by a dissertation committee · **Duration:** the design is the DL 903 deliverable; the executed study is a 2–3½ year dissertation · **Tools:** none required — the design is expressed in concepts and standard research methods.

The discipline's empirical machinery, at concept level

Before the design, the object of study — the machinery this capstone tests. The discipline encodes its doctrine in a **versioned package** — rules, playbooks, formulas, and a research corpus (the **DNA corpus**) — and around that package sits an empirical apparatus:

- **The win-probability composite.** A weighted composite of factor scores — past performance, technical capability, price competitiveness, incumbent/relationship advantage, team strength, customer knowledge, strategic alignment — normalized to a 0-to-1 **probability of win (pWin)**. It is the discipline's canonical instrument for its most important judgment.

- **The prediction-and-outcome record.** For each pursued opportunity, a stored **predicted pWin at decision time**, joined eventually to a **terminal outcome** (won / lost / withdrawn). This pairing — *predicted* and *observed* — is the field's empirical goldmine.
- **The calibration monitor.** The statistics that check whether estimates stay honest: **reliability diagrams** (observed win rate vs. predicted, by band), **Brier score** (mean squared error of the probability), **log-loss** (penalizing confident errors), **expected calibration error (ECE)** (the sample-weighted average reliability gap), and **population stability index (PSI)** (whether the distribution of predictions has drifted from baseline). The monitor bands drift into *watch / warn / critical* and recalibrates with a **monotone remap** — a curve that raises low estimates and lowers high ones while never reversing the ordering of opportunities.
- **The golden-thread verification.** A staged, reproducible end-to-end walk of the whole pursuit pipeline — prerequisites, seeding, sensing, scoring, gating, designing, analyzing, and the human-facing portal — executed against one canonical **canary company**, with each stage reporting pass / warn / fail and the run concluding in an overall green / amber / red status. It is a *reproducible research pipeline* for a whole discipline.
- **The DNA version lineage.** The corpus itself is versioned (semver-style). Every rule, playbook, and formula change is a **version event** — and each version event is a natural experiment in whether the doctrine's own evolution improves its predictions.

The hidden assumption under all of it, stated once: **the apparatus assumes its estimates mean something — that a 0.7 means roughly 70%.** This capstone designs the study that checks.

The research question bank

A candidate may build a dissertation on any question here. Each is falsifiable, feasible, and (in this young field) a real contribution. The bank is organized into the three families the title names — **calibration**, **DNA-version evolution**, and **attribution** — plus the golden-thread methodology family.

#	Family	Research question	The claim it tests
Q1	Calibration	Is the win-probability composite *calibrated* — does a 0.7 estimate win roughly 70% of the time?	The apparatus's central claim

#	Family	Research question	The claim it tests
Q2	Calibration	In which prediction bands is the composite over- or under-confident?	Reliability is not uniform
Q3	Calibration	Does calibration hold across **funder families** and **agencies**, or is a composite calibrated in one population miscalibrated in another?	The weights generalize
Q4	Calibration	Does calibration **drift over time**, and does drift predictably precede a market or estimator change?	The monitor's early-warning claim
Q5	Calibration	Do **structured composites** calibrate better than **expert judgment**?	The composite adds value beyond gut
Q6	Calibration	What is the right **observation floor** before a calibration claim is trustworthy?	The monitor's band thresholds mean something
Q7	DNA-version evolution	Does a **DNA version change** move predicted pWin, observed outcomes, or both?	The corpus's edits change behavior

#	Family	Research question	The claim it tests
Q8	DNA-version evolution	Does **recalibration after observed outcomes** actually improve calibration (a before/after of Brier/ECE around a version event)?	The learning loop learns
Q9	DNA-version evolution	Do organizations **learn from losses** — does a post-loss version adjustment predict improvement?	The cosmology loop compounds
Q10	Attribution	How much of **confidence** (predicted pWin) is explained by **agent × method × DNA version × opportunity type**?	Attribution is measurable
Q11	Attribution	How much of **outcome** (won/lost) is explained by the same interaction, once confidence is controlled?	Confidence and success are different objects
Q12	Attribution	Do the factors that predict wins at one **opportunity type** (SBIR, set-aside, IDIQ) hold at another?	Factor weights are not universal

#	Family	Research question	The claim it tests
Q13	Golden thread	What does the staged **golden-thread verification** certify — and what does it not?	A pass means the check ran, not the outcome was right
Q14	Golden thread	Is the verification **reproducible** — does a re-run agree with the original run?	The replication benchmark
Q15	Golden thread	What would a *rigorous* evaluation of a pursuit pipeline look like — outcome-based stage criteria instead of health checks?	A better method exists

---

The flagship dissertation design (one study, in depth)

A dissertation normally goes deep on **one** question. This section designs **Q1/Q2/Q3 — the calibration study** — the single most important claim the discipline makes, fully designed. Two alternative designs (DNA-version evolution and attribution) follow in compressed form.

Research question and hypotheses

RQ: *Is the win-probability composite calibrated, and is calibration stable across funder families?*

H1 (calibration): The composite's reliability diagram is flat — observed win rate equals predicted pWin in every band, within a pre-registered tolerance (ECE below a pre-registered threshold). *Falsifiable:* the reliability diagram shows systematic over-confidence (low predictions win more than predicted) or under-confidence (high predictions win less).

H2 (band-specificity): Calibration error is concentrated in the extreme bands (very low and very high predictions), not the middle. *Falsifiable:* ECE is uniformly distributed across bands.

H3 (stability): Calibration does not differ materially across funder families (research grants vs. service contracts vs. set-asides). *Falsifiable:* family-stratified reliability diagrams separate; a family × predicted-pWin interaction is significant.

H4 (drift): The calibration distribution is stable over the study window (PSI below a pre-registered threshold). *Falsifiable:* PSI crosses the warn band, indicating drift.

Design and measurement

- **Design.** Observational cohort of **pursued opportunities** over a defined window (e.g., 24–36 months), drawn from the prediction-and-outcome record. Every opportunity has a **predicted pWin at decision time** (the exposure) and a **terminal outcome** (the outcome). The design is not experimental — the discipline's decisions select the sample — which is why the survivorship hazard is handled explicitly below.
- **Constructs operationalized.**
  - *Predicted pWin:* the stored composite value at decision time, 0 to 1, with the **DNA version** that produced it recorded.
  - *Terminal outcome:* won / lost / withdrawn, with **won** defined by the published award record and **withdrawn** defined as a non-win under a pre-registered policy.
  - *Funder family:* a classification of the solicitation type (research grant, service contract, set-aside, commercial-adjacent), coded from the solicitation.
  - *Opportunity type:* the NAICS-style or program-family classification, recorded per opportunity.
- **Sampling and observation floor.** The design pre-registers the **observation floor** — the minimum number of terminal outcomes (e.g., N 300 resolved opportunities, with a minimum per funder family) before any calibration claim is trustworthy. The floor is decided *before* the data are analyzed, not after (the gate discipline applied to research).
- **Survivorship mitigation.** Only pursued bids resolve — the sample is chosen by the decisions under study. The design mitigates with: (1) a stated **withdrawn-as-non-win policy**; (2) family stratification so no single population dominates; (3) a bounded claim — the study estimates calibration *conditional on pursued bids*, and the dissertation says so plainly; and (4) a sensitivity analysis re-running the statistics with withdrawn excluded and with withdrawn-as-win, to show the conclusion does not rest on the choice.

Data needs

Data element	Source	Regime
Predicted pWin at decision time	The prediction-and-outcome record	Organizational (ethical)
DNA version per prediction	The version lineage (which corpus version produced each estimate)	Organizational (ethical)
Terminal outcome (won / lost / withdrawn)	The public award record, joined to each opportunity	Public
Funder family / opportunity type	The solicitation record	Public + organizational
Agency win-pattern corpus (for the secondary analysis)	The discipline's curated corpora	Organizational (as coding foundation)

The data environment has one asymmetry that shapes the whole study: the **public record is rich** (awards, dollar values, winners, review criteria) and the **decision side is hidden** (estimates, gates, deliberations). This design deliberately joins the two: the outcome side is public ground truth; the prediction side is ethically gathered organizational data.

Analysis plan (pre-registered)

1. **Reliability diagrams** for the full sample and each funder family: plot observed win rate against predicted pWin in bands (e.g., 0.1-wide), with the diagonal as the perfectly-calibrated reference.
2. **Headline statistics: Brier score**, **log-loss**, and **expected calibration error (ECE)**, computed overall and per band.
3. **Stability: population stability index (PSI)** comparing the prediction distribution across time-halves; a pre-registered threshold for the *watch / warn* band.
4. **Drift test (H4):** a time-split comparison of ECE with a pre-registered decision rule for "the composite is drifting."
5. **Family comparison (H3):** logistic regression of won/lost on predicted pWin with a funder-family interaction; a pre-registered test of whether the interaction improves fit.
6. **Robustness checks (pre-registered):** withdrawn excluded vs. counted-as-win; a winsorized extreme-band sensitivity; a random-time-half replication.
7. **Honest reporting:** the null results reported, the survivorship caveat stated, and the claims bounded to *conditional on pursued bids.*

Ethics and human-subjects care

The study touches **identifiable decision-makers** — the people whose estimates, gates, and budgets produced the prediction-and-outcome record — and the ethics plan is graded as seriously as the analysis plan:

- **IRB review** before data collection, per the host institution — the field's own gate discipline applied to research (doctrine/04).
- **Informed consent** for any organizational data that touches individuals: what is being studied, how confidentiality is protected, and the right to withdraw. Estimates are typically an organizational artifact, but the *decision-makers* who produced them are people, and the design treats them that way.
- **Confidentiality and anonymization.** The analysis uses opportunity-level records, not named individuals; the dissertation reports aggregated statistics and never identifies a specific person's estimate. Where the candidate studies their own firm, a clear boundary is drawn between the researcher hat and the practitioner hat.
- **Provenance discipline** for the public side: every award record, solicitation, and coding decision logged, so another researcher can verify the data is what the dissertation says it is.
- **The non-fabrication standard is absolute.** The discipline's own research corpora model the rule — documented facts and clearly-labeled illustrative material are kept apart, and invented specifics are never attributed to real organizations. A dissertation carries the same discipline.

What the study can honestly claim

If H1–H4 hold, the dissertation can claim: **the discipline's central instrument is calibrated, within a stated tolerance, across the funder families and window studied, conditional on pursued bids.** If they fail, the dissertation can claim the more valuable finding: **the instrument is overconfident in a specific band, or unstable across families — and here is where.** Either outcome is a contribution, because the field has almost no published calibration evidence for its most important number. The pre-registered decision rule is what makes both outcomes publishable: the analysis plan says *in advance* what the data must show, so the findings cannot be post-hoc justified.

Two alternative designs (compressed)

Design B — DNA-version evolution (Q7/Q8)

A **difference-in-differences** study around a DNA version event: for opportunities pursued before vs. after a corpus recalibration, compare (a) the distribution of predicted pWin and (b) the calibration quality (Brier/ECE). The identifying assumption — no concurrent market or organizational shock — is argued and

stress-tested with a placebo version event and a time-trend control. **Contribution:** the first evidence on whether the doctrine's own learning loop improves its estimates, or merely changes them.

Design C — Attribution (Q10/Q11)

A variance-decomposition study: for a sample of pursued opportunities, regress predicted pWin on agent (who ran it), method (structured composite vs. expert judgment vs. hybrid), DNA version, and opportunity type, with interactions; then regress outcome on the same interaction, controlling for predicted pWin. **Contribution:** the field's attribution question — how much of confidence and success is explained by each component — answered with variance explained (e.g., ANOVA or a hierarchical model), bounded by the same survivorship and selection caveats. This is the machine-side attribution claim (agent × method × DNA × opportunity type) turned into a testable research question.

Dissertation-chapter mapping

The dissertation sequence (`course/doctoral-dissertation-sequence.md`) maps onto the chapters of the final document. The candidate writes to the sequence, not around it:

Dissertation chapter	Content	Dissertation-sequence stage it satisfies
1 · Introduction	The research question, the gap, the intended contribution	Proposal (Stage 1)
2 · Literature review	The practice canon (Shipley/APMP claims), the scholarly neighbors (procurement economics, calibration/decision science, organizational learning), the field's thin empirical record	Literature review (Stage 2)
3 · Methodology	The frozen design: constructs, sampling, observation floor, ethics plan, pre-registered analysis plan	Methodology approval (Stage 3)
4 · Data	The dataset, its provenance log and codebook, the ethics record	Data collection (Stage 4)

Dissertation chapter	Content	Dissertation-sequence stage it satisfies
5 · Results	The reliability diagrams, Brier/ECE/PSI, family comparisons, robustness checks — executed as pre-registered	Analysis (Stage 5)
6 · Discussion + contribution	What the findings do and do not mean for the doctrine; the bounded claims; the next study on the research agenda	Defense (Stage 6)

The publication pathway follows: the literature review becomes a review article; the empirical core (Chapter 5) becomes the flagship article; a robustness extension (e.g., the cross-agency replication) becomes a second article; and if the candidate designs the golden-thread validity study (Q13–Q15), the methodology contribution becomes a methods paper.

Grading rubric for the doctoral capstone

The four-level scale — **Not Ready · Developing · Proficient · Distinguished** — applied to a research design. At the doctoral tier, *Distinguished* means the design is not just defensible but **publishable**: a committee would send the candidate to collect the data.

Dimension	Not Ready	Proficient	Distinguished
Question	Vague or unfalsifiable; answers an assertion, not a question	Specific, falsifiable, researchable — a committee can see the answer taking shape	The question is one the field needs answered, with the closest prior work named and the gap argued
Design	Design cannot support the claim	Design matches the question; the identifying assumption is stated	The identifying assumption is argued and stress-tested; a placebo or robustness check guards it

Dimension	Not Ready	Proficient	Distinguished
Feasibility	Data assumed that the candidate does not have	Data sources named; access and ethics plan realistic	The observation floor is pre-registered; a pilot proves the measurement protocol
Measurement	Constructs undefined	Constructs operationalized (pWin, outcome, funder family, DNA version)	Reliability (inter-rater, band resolution) is addressed before the main analysis
Ethics	No ethics plan; human-subjects care absent	IRB, consent, confidentiality, and provenance named	The ethics trail is complete and the boundary for studying one's own firm is drawn
Analysis plan	"The data will tell us"	Pre-registered statistics (Brier, ECE, PSI, reliability diagrams) and robustness checks	The pre-registered decision rule makes a null result as publishable as a positive one
Survivorship	Hazard unnamed	Hazard named and mitigated	Mitigations are bounded in the claim — the dissertation says exactly what it can and cannot conclude
Contribution	Repeats what practitioners assert	A defensible answer to an open agenda question	Either outcome is a contribution; the next study on the agenda is opened, not closed

The doctoral trap being tested: the "the data will tell us" plan — an analysis so vague that any finding can be post-hoc justified. A design that pre-registers its decision rule *before* seeing the data passes the test; a design that leaves itself room to rationalize fails it, whatever the data later say.

Where this maps to Shipley + Dream DNA

The industry canon. The doctoral capstone sits deliberately outside the Shipley *practitioner* canon — it studies the canon. The practitioner literature (Shipley, APMP, the proposal-professional bodies) is *prescriptive*: capture plans, win themes, color teams, four volumes, price-to-win. Almost none of it is *empirical*. The scholarly neighbors — procurement and auction economics, public administration, decision science and judgment/calibration research, organizational learning — supply the methods: reliability diagrams, Brier score, ECE, PSI, logistic regression, difference-in-differences. The dissertation's literature review (Chapter 2) is precisely the act of mapping what the practitioners claim against what the scholars have measured — and finding the empty space in between that this study fills.

The Dream DNA. The machine side of the doctrine is not the study's *subject matter* in the tool sense — it is the study's *object*, read at concept level. The **win-probability composite** is the instrument under test; the **prediction-and-outcome record** is the data; the **calibration monitor's statistics** (Brier, ECE, PSI, the monotone remap) are the measurement toolkit; the **DNA version lineage** is the natural-experiment lever for the evolution family; the **agency win-pattern corpora** are the coding foundation for a pattern-discrimination replication; and the **golden-thread verification** is itself a research method — a staged, reproducible pipeline that a doctoral student can study, critique, and improve. The canary company that the machine side runs its golden-thread on is the same Ravonics the human curriculum teaches on — which means the doctoral student's finding about the calibration engine is a finding about the same doctrine the B.S. student learns and the M.S. student leads. **Test the doctrine so the doctrine can be trusted** — that is the doctoral contribution, and it is what this capstone designs.

The capstone in one sentence

The discipline's most important number — probability of win — has almost no published calibration evidence, and this capstone designs the study that produces it: a pre-registered, ethically-grounded, survivorship-aware test of whether a 0.7 means roughly 70%, with the DNA version and the funder family recorded so the answer is a contribution whether the estimate is honest or not.