

The Question-Refinement Script

PURPOSE: Use this 5-step script during initial research project discovery to clarify research objectives, establish measurable metrics, validate underlying data sources, identify change drivers, and determine an alternative plan where feasible.

1. What **decision** will this analysis inform?

Decision Alignment

Identify specific decision or choice contingent on the findings. If no decision hinges on the result, consider how this request fits within institutional priorities.

2. How do you **define** success (or metric of interest): what would you count, for whom, by when?

Criteria & Scope Definition

Establish precise metric(s), target cohort(s), timeframes, and metric thresholds.

3. What **data** do you have? Who enters it, and when? Is the data accurate (validation)?

Data Integrity

Evaluate data generation process, entry bias, and validation status to prevent bad inputs.

4. What **changed** recently that prompted this request?

Context

Uncover recent circumstances (i.e., performance drops) that prompted the inquiry or urgency. Understanding the root cause helps tailor the analysis.

5. If we can't answer this question, what's the **next most useful** one do they agree to move forward?

Fallback Strategy & Alignment

Identify fallback objectives and alternative proxy questions in case primary data is missing, incomplete, or inconclusive. This can also be done in a follow up meeting.

The Pre-Project Data Audit Checklist

PURPOSE: Use this 7-step data audit prior to beginning analytical modeling and preferably in the initial meeting. Evaluate the database you are working from.

1. Who enters each key field, and when? Has the process changed over the years?

Data Historical Entry | *Ex.:* Applicants self-report grades, evaluator validates; paper application era vs electronic application – entry gaps.

2. Where is the authoritative source of each key field?

Source of Truth | *Ex.:* The source of truth was the evaluator's copy of the application where grades were validated. Not all grades were entered into our SIS (student information system).

3. Is any field overwritten rather than versioned?

Data Changes and Historical Loss | *Ex.:* TEAS score was overwritten on retake, making this application admission criterion data unusable.

4. Can you replicate each derived value (GPA, totals) from other sources?

Derived Metric Reconciliation | *Ex.:* Because not all grades were in SIS, could not replicate the GPA calculation to validate results.

5. Count each field's usable records.

Completeness of Data | *Ex.:* A few metrics such as the veteran status were too sparse to enter the analysis. Document what can and cannot be tested and why.

6. Build a data dictionary as you go - one entry per field.

Living Documentation for Above Findings | *Ex.:* We noted which fields needed to be created, i.e., completer vs on-time completer vs non-completer; work experience overlaps between fields, how/why.

7. Who answers questions as they arise - and when?

Subject Matter Expert (SME) Support | *Ex.:* We collaborated with admin support who entered data into the database when new Qs arose.

The Pivot Decision Framework

PURPOSE: Use as decision protocol – when a data gap destabilizes the analysis, consider the three pivot pathways and document the choice for everyone’s awareness.

1. Drop the variable that cannot be used

State why it cannot be used, in writing, where stakeholders will see it. Be ready to present this to them when sharing the results.

Case Study Example *TEAS, veteran status, languages – could not be used, documented each with its reason (i.e., small sample size, overwritten scores).*

2. Change the method

Simplify, reweight, or re-specify to fit the data you have and name the “cost”.

Case Study Example *We pivoted from science GPA to reweighted science course grades, trading 556 records for a stronger 218-record model.*

3. Reconsider the scope

Defer questions that need data not currently available (nobody has collected it yet).

Case Study Example *The persistence-factors question was deferred to a future interview study; meanwhile provided basis descriptive data available as a starting point.*