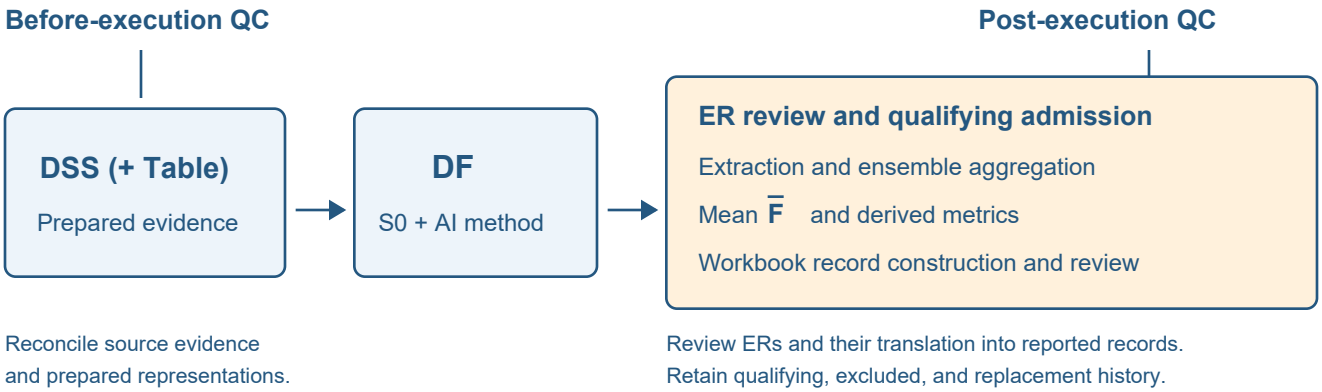
Quality Control and Execution Integrity

DJE | Working reader edition | October 28, 2026

Validity Precedes Performance

Quality Control (QC) supports the integrity of the evidence on which engineering conclusions and governance judgments depend. Errors in evidence preparation, execution, or result handling can undermine that basis even when reported results appear persuasive.

A Decision Function (DF) can produce a plausible, coherent, and even accurate result while departing from the procedure intended to be tested. It may receive incorrectly prepared evidence, use ineligible information, misidentify the decision subject, misinterpret an authorized input, or depart from the specified procedure. A favorable outcome does not cure these defects.



QC therefore asks two central questions:

Before execution: has the evidence supplied to the DF been checked for errors in selection, transcription, or transformation?

After execution: did the execution follow the specified procedure, and were its outputs correctly extracted, admitted, aggregated, and represented in the resulting workbook records?

A qualifying execution that produces an inaccurate estimate remains evidence of performance under the tested evidence state and procedure. An invalid execution does not become valid because its estimate agrees with the realized outcome. QC admission follows validity requirements rather than the attractiveness of the result. Qualification does not itself establish accuracy, Engineering Sufficiency (ES), or authority for autonomous execution.

Before DF Execution: Evidence and Representation

Front-end QC asks whether the intended evidence is correctly bounded and faithfully conveyed before the DF is engaged. In the Real Property Valuation Decision Node (RPVDN), the Decision Surface Snapshot (DSS) supplies property evidence drawn from MLS property detail reports, including property fields, remarks, photographs, and private comments. The DSS presents the evidence; the governing specification defines its authorized use.

The Table Intervention: Preparation and Observed Results

The Table (T) condition additionally supplies a structured quantitative table organizing selected information already contained in the DSS. The No Table (NT) condition supplies the DSS alone. Adding the table is an input-representation intervention. Checking its fidelity to the source is a separate front-end QC responsibility. An accessible representation can still contain a transcription error, assign a value to the wrong property, or otherwise transform information incorrectly.

In the laboratory, ChatGPT constructed and reviewed the structured table, and the investigator visually compared it with MLS-generated tables. Those source tables were generated separately from the AI used in the DF; they were not another AI reconstruction of the same evidence. Those visual checks were part of the engineering process from the beginning of table use. Their purpose was to check the represented values and property associations against the source material before using the table as experimental input.

OPVAL supplies the most relevant NT/T comparison because both conditions were run with a comparable QC process. It is the retrospective prediction setting: realized subject outcomes were concealed from the DF. The following results summarize recorded exclusions under the two evidence representations.

Condition	Evidence supplied	Qualifying executions	Excluded attempts	Total attempts	Exclusion rate
NT	DSS	90	32	122	26.2%
T	DSS + Table	90	2	92	2.2%

The Table condition was associated with fewer detected invalid executions and fewer replacement attempts. These exclusions were identified through **completed-execution QC**. The comparison does not isolate the effectiveness of front-end checking or establish that the table eliminates failures or improves price accuracy.

Understanding the result also requires examining what completed-ER review detected and how exclusion judgments were reached. The later recorded examples make those failure mechanisms and judgments inspectable. The table intervention and the review that detected its associated exclusion pattern address connected but different parts of QC.

The **Decision Function Engineering** companion develops the experimental design: the same S0 specification governed both conditions, the selected ChatGPT-based method was used, and NT/T varied the evidence representation presented to it. Here the focus is how that representation was checked and how subsequent executions were reviewed. The two responsibilities are connected: a reliable input intervention requires both faithful preparation and examination of what the DF actually did with it.

How the Front-End Checks Evolved

Beginning with LOOCV1, the investigator also supplied MLS tables for ChatGPT-assisted comparison with the structured tables, using an independently generated MLS Property List report as a separate reference. This added an AI-assisted source-comparison step to the investigator's existing visual checks.

The MLS Property List report was never supplied to the experimental DF. It was not used for AI-assisted front-end comparison in the P4A OPVAL worked example and is not an artifact in that retained case package. The P4A case guide describes the investigator's visual source check without presenting a reconstructed report as an original retained exhibit.

For a deployed AI-assisted QC process, providing a separately generated source table is a useful design option: it gives the reviewer a reference against which to check the AI-prepared representation. The automated retrieval or generation of the DSS and source table, property matching, and discrepancy handling would themselves need defined controls. A separate source does not by itself establish review independence or effective detection.

These front-end checks are distinct from using the structured table later as a reference when reviewing completed Experimental Records (ERs). The evolution of one does not establish when or how the other was applied.

Bound the Evidence for the Actual Task

Correct preparation requires the right subject, eligible comparison population, and outcome redactions. OPVAL and PROSP both concern sale-price predictions for individual subjects using subject-specific evidence cutoffs. OPVAL is retrospective evaluation with realized Y concealed from the DF; PROSP produces and freezes a prediction before Y is known.

LOOCV instead uses held-out sale-price estimates to characterize an identified evidence state. Each held-out task uses the state's other eligible completed sales, including those that closed after the held-out property. The LOOCV companion explains state construction and records the historical prediction-window anchors.

These task boundaries must be conveyed accurately. A requirement appropriate to a prediction can change the intended LOOCV task if applied indiscriminately.

The checks described here were human-led engineering and experimental activities. Autonomous execution under Standing cannot depend on a human performing them for every candidate. Evidence-preparation controls, discrepancy handling, and responses to unresolved errors would need to be engineered and assessed for that operating mode.

After DF Execution: Review the Experimental Record

Completed-ER review asks whether an execution followed the intended task and used its supplied evidence consistently. The laboratory review included human engineering judgment and AI assistance. In the table-assisted review used for OPVAL and LOOCV1, the structured table provided a reference for checking the ER's treatment of quantitative evidence. Such review could identify consequential misreadings, unsupported assertions, wrong-subject reconstruction, or specification departures.

The ER's reported reasoning makes claims and interpretations available for inspection. It is not an independently verified trace of internal model computation. A coherent explanation is review material, not proof of execution validity.

The following nine examples preserve recorded Registry findings and historical dispositions. They illustrate mechanisms detected during completed-ER review; they are not a representative sample, frequency evidence, or a new adjudication of the judgments.

Failure mechanism	Execution ID and condition	Recorded QC finding
Wrong held-out subject	LOOCV1-P5-S0-T-E004 · Table.	“Wrong experimental condition. Execution reconstructs and values P6 / 144 Crosstide Circle rather than the designated P5 / 161 Crossroad Lakes Drive held-out property, materially substituting the wrong Subject and evidence condition.”
Material size misreading	OPVAL-P2A-S0-NT-E002 · No Table.	“EXCLUDE / REPLACE — material Subject-state reconstruction error: Subject reconstructed at approximately 2,818 SF rather than 2,006 SF; erroneous additional size expressly used as a major positive valuation characteristic.”
Listing-history reconstruction	OPVAL-P3A-S0-NT-E013 · No Table.	“EXCLUDE / REPLACE — material Subject listing-state reconstruction error: represented as originally listed at $810,000 with no visible reduction, erasing the actual $885,000→$810,000 repricing history used in transaction reasoning.”
Market-exposure reconstruction	OPVAL-P3A-S0-NT-E012 · No Table.	“EXCLUDE / REPLACE — material Subject transaction-state reconstruction error: approximately 23 days exposure reconstructed instead of approximately 237 days and used in transaction reasoning.”
A chronology rule applied to the wrong task	LOOCV0-P1-S0-T-E007 · Table.	“Excluded because the Decision Function removed P2/P3 from the primary valuation based on transaction chronology, materially altering the supplied LOOCV0 evidence surface. Exclusion is process/state based, not numerical-result based.”
Outcome-redaction protocol misinterpreted	OPVAL-P4A-S0-NT-E006 · No Table.	“EXCLUDE / REPLACE — protocol-interpretation non-execution. DF incorrectly treated visible completed-sale outcomes for comparables as failure of the Subject outcome-redaction control and produced no valuation output.”
Required execution isolation not followed	LOOCV1-P2-S0-T-E008 · Table.	“Execution conducted in non-temporary chat; replacement execution generated under required isolation condition. Substantive QC identified no independent reasoning/evidence defect.”
Misstated roof and HVAC dates	LOOCV1-P8-S0-T-E010 · Table.	“EXCLUDE — material Subject evidence-assimilation error. ER reverses the reported roof and HVAC vintages, reconstructing a 2022 roof and 2016 HVAC where the authorized evidence reports a 2015 roof and 2022 HVAC, materially altering major-system condition evidence.”
Disputed garage interpretation	OPVAL-P4A-S0-T-E005 · Table.	“EXCLUDE — Subject reconstructed as having a 1.5-car garage and that fact was admitted into comparative reasoning; DSS establishes 2 garage spaces. No causal effect on I/F is asserted.”

Different Failures Require Different Checks

The recorded examples span several kinds of validity requirements. Evidence-reading and reconstruction failures include wrong-subject valuation, size and listing-history errors, misread exposure, and misstated roof and HVAC dates. Task and protocol failures include inappropriate chronology restrictions and misinterpretation of subject-outcome redaction. Execution-condition failures can arise even without an independently detected reasoning defect, as the isolation example shows.

No single arithmetic check addresses that range. Controls need to test evidence fidelity, task compliance, required execution conditions, and interpretive claims. The chronology example also exposes a specification issue: S0 refers to an applicable cutoff without explicitly distinguishing the task-specific rules. Findings can therefore call for clearer instructions as well as correction or replacement of an individual ER. These are engineering requirements surfaced by the laboratory, not a completed automated detection system.

QC Judgments Can Also Be Challenged

The final example makes QC judgment itself inspectable. E005 accepted the reported 1.5-car garage and used it in comparative reasoning. ChatGPT challenged that interpretation during post-execution review and proposed exclusion; the investigator reviewed and agreed with that rationale. E005 therefore records an AI-proposed exclusion accepted through human review. The DSS reports 1.5 garage spaces, so the recorded claim that it establishes two is not established as factually correct. There is no evidence that the challenge was based on a visual comparison of garage photographs. Whether photographs support comparable garage sizes despite differing agent-entered capacity fields remains an interpretive question, not a newly adjudicated finding here.

Such conflicts show why the steward must assess the acceptability of the DF's interpretive discretion before authorizing autonomous execution. Collaborative use permits a human to challenge an interpretation before proceeding. Autonomous processing cannot depend on that intervention, although sufficiently developed post-ER, pre-release controls could detect a conflict and require HOLD. This laboratory does not establish the effectiveness of such an automated review protocol. Human review after autonomous release is retrospective.

Check Extraction and Aggregation

After admission, result-handling checks must ensure that values are extracted correctly, associated with the right property and state, and aggregated over the intended qualifying population. RPVDN uses the mean of ten qualifying executions. Excluded attempts require replacement to complete that population; their values do not enter the qualifying ensemble calculations.

In the laboratory, review of the translation from ERs into workbook records and the checking of formulas and aggregates were collaborative activities involving the investigator and AI assistance. This work did not establish a standing automated procedure for reconciling workbook records against their originating ERs.

A deployed process would need controls for that full lineage: extracted values and identifiers, qualifying dispositions, ensemble membership, mean construction, derived metrics, and the records presented in the workbook or other reporting destination. Independent recalculation can check arithmetic, but cannot alone establish that the right values were faithfully extracted from valid executions. Correct workbook arithmetic and faithful DF execution are separate requirements.

Where a requirement can be checked by a defined deterministic procedure, that is the preferred control: for example, reconciling identifiers and extracted values, checking ensemble membership, or

recalculating formulas. AI review can address interpretations that resist such direct checks. If AI is used, a separate, isolated review context can limit carryover from the producing conversation; it does not by itself establish independent errors or reliable detection. A temporary chat is one way to isolate conversational context in collaborative use; an automated workflow would need to implement and assess the corresponding separation.

Retain Exclusions and Replacement History

Excluded ERs remain part of the experimental record, with their dispositions and replacement histories. Retaining only qualifying outputs would conceal detected defects, contested judgments, and the extra attempts required to construct the ensemble. The P4A case guide makes that lineage inspectable, including the excluded record and its replacement.

Traceability supports scrutiny of both the DF and QC. It does not establish that every admitted execution was valid or that every exclusion was justified.

Use QC Findings to Improve the Decision Node

QC findings can expose weaknesses in evidence preparation, the configured DF, result handling, or the controls themselves. The structured-table intervention illustrates a consequential engineering response: change how evidence is presented and examine its effect on detected execution failures. Its observed benefit helped motivate the project's attention to QC and the procedures developed and proposed here. Engineering can address the source of recurring failures rather than only reject and replace individual executions.

Disposition and retention remain essential. A local defect can require correction, replacement, or candidate HOLD; a recurring pattern or shared weakness can require broader engineering reassessment. The Governance companion explains the implications for Standing and authority. QC supplies findings for those judgments rather than substituting for them. Keeping the evidence, procedure, and review history identifiable also helps avoid mistaking better detection for worse DF behavior.

The laboratory developed human-led, AI-assisted checks and exposed further needs for autonomous evidence preparation, ER review, and translation into reported records. Developing and assessing those controls remains a critical deployment-engineering task. A reduction in detected exclusions does not by itself establish better price accuracy or an effective autonomous QC system.

Scope and Limitations

Comparability. *The historical record does not establish that every review step, execution setting, and model/version detail was identical across conditions. The comparison is interpreted as one conducted with comparable QC, not a demonstration that all potential sources of variation were controlled.*

Comparison coverage. *LOOCV1 and PROSP had no No Table runs, so neither supports a direct NT versus T comparison. LOOCV0 included both conditions, but its QC process was less consistent across them: the No Table runs preceded development of the structured tables and did not receive table-assisted ER review. That additional review was not retrospectively applied. Table-assisted ER review was used for OPVAL and LOOCV1. These differences explain why the presentation emphasizes OPVAL rather than combining unlike populations.*

Detection and judgment. *Admission means that an ER qualified under the applicable review process, not that every possible defect was absent. QC can miss defects or reject an execution that should qualify. Historical judgments remain reviewable against the specification and evidence. Traceability supports scrutiny without proving that the controls or their judgments were sufficient.*

Shared errors and review independence. *An ensemble can reduce variation among qualifying outputs while preserving a common mistake. A second review can share the assumptions or blind spots of the first. Independence should be assessed through the failure mechanisms a control can detect, rather than inferred from a separate execution or reviewer label. Where direct non-AI checks are unavailable, a different AI model may provide a useful challenge to an interpretation. That possibility requires evaluation: model diversity alone does not establish independent errors or reliable detection. The use of an independently generated MLS report likewise does not by itself make the AI-assisted check an independent reviewer.*

Demonstrated practice and proposed operation. *The laboratory demonstrates human-led, AI-assisted experimental QC and its evolution. Workbook-to-ER reconciliation was collaborative; no automated end-to-end record-reconciliation QC capability was established. Providing independently generated source tables to AI reviewers is a proposed design option, not a validated autonomous control. It does not validate a complete automated deployment protocol, control-effectiveness claim, retry budget, release procedure, or monitoring regime. These require further engineering for the intended use. Human–AI collaborative routes likewise require their own defined responsibilities and permissions; developing them is beyond this companion's scope.*