RLQ-IBOT Protocol v1.5

A Qualitative-Statistical Hybrid Framework for Assessing Developmental Trajectories in Reasoned Leadership

Authors: Dr. David M. Robertson
Affiliation: GrassFire Industries LLC, National Leaderology Association
Date: December 2025
Status: Pre-Pilot Methodological Specification

ABSTRACT

The RLQ-IBOT protocol provides a methodological framework for assessing leadership development within the Reasoned Leadership paradigm, integrating qualitative behavioral observation (IBOT Method) with pseudo-longitudinal statistical analysis (Chi-Square Twist). By operationalizing the Nine Pillars of Reasoned Leadership through categorical developmental stages with quantifiable behavioral anchors, the protocol enables rigorous evaluation of cognitive and executional competency progression following targeted interventions. This design addresses methodological limitations of self-report instruments while providing statistical validation of developmental patterns across time-stratified cohorts.

Scope Clarification: This document is a measurement specification, not an empirical validation study. It defines how leadership development should be assessed within the Reasoned Leadership framework—the constructs, operational definitions, scoring rules, reliability requirements, and falsification criteria. It does not claim that Reasoned Leadership training is effective, nor does it present causal evidence. The protocol provides the measurement instrument; separate empirical studies will apply this instrument to evaluate training effectiveness. The protocol distinguishes between:

1. **Instrument validity:** Does the measurement system function as specified?
2. **Intervention effectiveness:** Does training produce measurable developmental change?

This document addresses the first question. The second requires empirical application.

Keywords: leadership development assessment, IBOT Method, Chi-Square Twist, qualitative-quantitative hybrid, behavioral observation, developmental trajectories, Nine Pillars framework

---

1. INTRODUCTION

1.1 Rationale and Positioning

Leadership development assessment faces persistent methodological challenges: self-report instruments suffer from social desirability bias and Dunning-Kruger effects (Kruger & Dunning, 1999), while 360-degree surveys rely on untrained raters with inconsistent interpretations (Atwater et al., 2007).

Traditional instruments measure perceptions rather than actual behavioral change, and cross-sectional designs capture static competency snapshots rather than developmental trajectories.

The RLQ-IBOT protocol addresses these limitations through three integrated methodologies:

1. **IBOT Method (Robertson, 2024):** Qualitative behavioral tracking via trained evaluator observation, reflective journal analysis, and real-world decision review—eliminating self-report bias while emphasizing developmental progression over static assessment.
2. **Nine Pillars of Reasoned Leadership (Robertson, 2025):** Theoretically grounded constructs spanning Self-Determination Theory (Deci & Ryan, 1985), cognitive restructuring (Beck, 1979), and strategic execution (Mintzberg, 1987), providing a comprehensive framework for leadership competency assessment.
3. **Chi-Square Twist (Robertson, 2022):** Pseudo-longitudinal analysis method stratifying participants by time since intervention and using chi-square contingency testing to detect temporal developmental patterns without full longitudinal study costs.

1.2 Scope and Purpose

This protocol establishes the measurement standard for assessing Reasoned Leadership development. It specifies:

- Evaluator training and certification procedures
- Categorical stage definitions with operational anchors
- Evidence collection and triangulation protocols
- Sample size requirements and cohort structure
- Statistical analysis procedures and validity controls
- Limitations and falsification criteria

Critical Positioning: This is a methodological contribution, not an empirical validation study. The protocol provides the "ruler" for measuring Reasoned Leadership development; separate empirical studies will apply this ruler to test training effectiveness.

---

2. EVALUATOR TRAINING AND CERTIFICATION

2.1 Evaluator Qualifications

Minimum Requirements:

- Master's degree in organizational psychology, leadership studies, or related field
- Minimum 5 years professional experience in leadership assessment, coaching, or development
- Demonstrated understanding of qualitative research methodologies
- No disciplinary actions or ethical violations in professional practice

Preferred Qualifications:

- Doctoral training in behavioral assessment

- Prior experience with structured observation protocols
- Familiarity with leadership development frameworks (not necessarily Reasoned Leadership)

2.2 Certification Curriculum (4-Week Program)

Week 1: Theoretical Foundation (16 hours)

Module 1A: IBOT Method Principles (4 hours)

- History and epistemological grounding
- Qualitative vs. quantitative assessment paradigms
- Behavioral observation standards vs. perception-based ratings
- Triangulation methodology and evidence synthesis

Module 1B: Nine Pillars Framework (8 hours)

Detailed review of each pillar with theoretical grounding:

- Autonomy (Deci & Ryan, 1985)
- Mastery/Competence (Ericsson et al., 1993)
- Purpose/Relatedness (Deci & Ryan, 1985)
- Consistencies (Mintzberg, 1987)
- Accuracies (Stanovich & West, 2000)
- Efficiencies (Porter, 1996)
- Sound Thinking (Kahneman, 2011)
- Accurate Decisions (Baron, 2008)
- Effective Communication (Shannon & Weaver, 1949)

Exemplar case studies demonstrating each pillar in practice. Common misinterpretations and boundary conditions.

Module 1C: Cognitive Biases in Assessment (4 hours)

- Confirmation bias and expectancy effects
- Halo effect and horn effect
- Recency bias and primacy bias
- Attribution errors in behavioral observation
- Strategies for bias mitigation

Week 2: Vignette Analysis and Calibration (12 hours)

Module 2A: Case Analysis Practice (6 hours)

- Independent assessment of 20 standardized vignettes
- Each vignette includes:
    - Reflective journal excerpts (3-5 entries)
    - Structured interview transcript excerpts
    - Decision documentation (2-3 cases)
    - Contextual information (organizational setting, tenure)
- Evaluators assign categorical stage for each pillar with written justification

Module 2B: Calibration Sessions (6 hours)

- Small group (4-5 evaluators) discussion of vignette assessments
- Facilitated review of discrepancies
- Comparison to expert benchmark categorizations
- Identification of evidence weighting patterns
- Interim quadratic-weighted Cohen's kappa calculation (target: $\kappa > 0.60$)
- Individualized feedback on reasoning processes

Week 3: Live Observation Practice (16 hours)

Module 3A: Shadowing Certified Evaluators (8 hours)

- Observe 3 complete assessments conducted by certified evaluators
- Review evidence collection process in real-time
- Witness interview techniques and probing strategies
- Observe evidence synthesis and categorical assignment
- Debrief sessions explaining evaluator reasoning

Module 3B: Co-Evaluation with Oversight (8 hours)

- Conduct 5 leader assessments with certified mentor present
- Collect evidence independently
- Assign categories with mentor observing process
- Receive immediate feedback on:
    - Interview question quality
    - Evidence interpretation accuracy
    - Categorical boundary judgment
    - Documentation completeness
- Iterative refinement based on mentor guidance

Week 4: Certification Assessment (8 hours)

Module 4A: Blind Vignette Assessment (4 hours)

- Independent evaluation of 10 holdout vignettes (never seen before)
- No group discussion or feedback during assessment
- Written justification required for each categorical assignment
- Evaluation against expert benchmarks
- **Passing Standard:** Quadratic-weighted Cohen's $\kappa > 0.70$ across all nine pillars

Module 4B: Live Case Evaluation (4 hours)

- Conduct 2 complete leader assessments independently
- Submit evidence dossiers and categorical assignments
- Expert review of:
    - Evidence sufficiency and triangulation
    - Categorical assignment justification quality
    - Documentation standards compliance

    - Adherence to protocol specifications
- **Passing Standard:** Agreement with expert on 16/18 pillar assignments (89%)

Certification Outcome:

- **Pass:** Certified as RLQ-IBOT evaluator, eligible for research/practice use
- **Conditional:** Additional practice required (3-5 more cases with mentor)
- **Fail:** Must repeat full 4-week program

2.3 Annual Recertification

Requirements:

- Review and assess 5 new standardized vignettes annually
- Maintain quadratic-weighted Cohen's $\kappa > 0.70$ against expert benchmarks
- Attend 1-day (8-hour) calibration workshop covering:
    - Protocol updates and refinements
    - Common evaluation errors identified in field use
    - New case studies and edge cases
    - Inter-evaluator reliability benchmarking

Failure to Recertify:

- Suspension from evaluator pool
- Remediation: 2-day refresher workshop + re-assessment
- Persistent failure: Full 4-week recertification required

2.4 Evaluator Independence and Bias Mitigation

Phase 1 (Pilot Implementation): Certified RL Practitioners

Pragmatic Constraint: Initial evaluator pool will consist of certified Reasoned Development practitioners who possess deep framework knowledge but potential investment bias.

Bias Mitigation Strategies:

1. **Blind Cohort Assignment:** Evaluators not informed of leader's time-since-training during evidence review
2. **Dual Independent Assessment:** Each leader evaluated by two certified evaluators working independently
3. **Evidence Documentation Standards:** Categorical assignments must reference specific dated behaviors, not holistic impressions
4. **Falsification Incentives:** Evaluators rewarded for identifying non-developmental or regressive patterns (not just progression)
5. **Audit Trail Requirements:** All assignments subject to external review with full evidence transparency

Acknowledged Limitation: Phase 1 evaluators may unconsciously favor positive developmental trajectories to validate training effectiveness. This is explicitly stated in all publications using Phase 1

data. In Phase 1 implementations, evaluators may conduct both interviews and scoring; this limitation must be acknowledged in any resulting publications.

Phase 2 (External Validation): Independent Evaluators

Objective: Train external evaluators with no prior Reasoned Leadership exposure to assess whether framework-naive raters reach similar conclusions.

Implementation:

- Recruit organizational psychologists from academic institutions
- Recruit leadership development professionals from non-RL programs
- Complete identical 4-week certification curriculum
- Assess same leader sample as Phase 1 RL practitioners
- Calculate inter-evaluator agreement: RL practitioners vs. external evaluators
- **Target:** Quadratic-weighted Cohen's $\kappa > 0.60$ (substantial agreement)
- If $\kappa < 0.60$, indicates framework-specific interpretation bias

Phase 2 Requirements: Separation between data collection and scoring roles. Evaluators who conduct interviews do not score those cases. Scoring evaluators receive blinded evidence dossiers without cohort identification.

Timeline: Phase 2 initiated after pilot study completion (18-24 months post-launch)

---

3. NINE PILLARS: CATEGORICAL STAGES AND OPERATIONAL ANCHORS

Each pillar features three developmental stages with quantifiable behavioral thresholds derived from triangulated evidence. Percentage thresholds represent minimum frequencies observed across evidence sources (journals + interviews + decisions).

3.0 Operational Definitions: Decision Inclusion and Scoring Determinism

Before assessing any pillar, evaluators must apply consistent operational definitions for what constitutes admissible evidence.

3.0.1 Decision Inclusion Criteria

For purposes of scoring, a "decision" must meet ALL of the following criteria:

1. The leader selected among two or more plausible alternatives
2. The decision fell within the leader's formal or de facto authority
3. The decision was documented in at least one admissible evidence source (journal entry, interview transcript, decision memo, or communication artifact)
4. The decision had identifiable downstream consequences

Exclusions:

- Routine approvals without discretion

- Procedural compliance actions
- Decisions fully delegated without leader involvement

3.0.2 Uncertain Scenario Definition (for Pillar 5: Accuracies)

An "uncertain scenario" is defined as a situation in which:

1. A factually correct answer exists
2. The leader lacks immediate access to complete information
3. The leader must reason under uncertainty to respond or act

Examples: Questions regarding performance metrics, operational constraints, resource availability, or factual matters where the leader does not have immediate data.

Exclusions: Opinion questions, value judgments, and speculative predictions are excluded from uncertain scenario counts.

3.0.3 Scoring Determinism

Stage assignment follows deterministic rules. Evaluator judgment is constrained to evidence admissibility decisions, not stage determination.

Process:

1. Evaluator reviews evidence and determines which instances meet inclusion criteria
2. Evaluator calculates the proportion of admissible instances meeting stage-specific behavioral thresholds
3. Stage assignment follows directly from threshold calculation:
    - If proportion meets Stage 3 threshold → Stage 3
    - If proportion meets Stage 2 threshold but not Stage 3 → Stage 2
    - If proportion does not meet Stage 2 threshold → Stage 1

Expert judgment enters only at the admissibility step. Once evidence is admitted, stage assignment is algorithmic.

3.1 PILLAR 1: AUTONOMY

Construct Definition: Self-sufficiency in critical thinking, decision execution, and strategic forecasting; intellectual independence from hierarchical validation-seeking.

Stage 1: Dependent

Threshold: <30% of decisions executed without seeking external validation

Behavioral Indicators:

- Defers to authority figures before finalizing decisions
- Seeks consensus approval for non-controversial choices
- Delays action pending supervisor sign-off
- Strategic forecasts echo superior's stated positions
- Justifies decisions by citing others' authority

Evidence Requirements:

- Journal: Minimum 5 entries documenting validation-seeking behavior
- Interviews: Examples of delayed decisions pending approval
- Decisions: <30% show independent execution without prior consultation

Stage 2: Emerging

Threshold: 30-69% independent execution; validation sought only in novel high-stakes scenarios

Behavioral Indicators:

- Executes routine decisions independently
- Seeks input (not approval) for complex strategic choices
- Generates independent forecasts, validates against data (not authority)
- Uses consultation selectively for expertise gaps
- Documents independent reasoning before seeking feedback

Evidence Requirements:

- Journal: 3+ entries showing shift from approval-seeking to selective consultation
- Interviews: Can articulate when/why consultation is needed vs. not needed
- Decisions: 30-69% executed without prior validation

Stage 3: Autonomous

Threshold: ≥70% independent execution with structured reasoning evident

Behavioral Indicators:

- Executes decisions based on analytical frameworks, not consensus
- Strategic forecasts independently generated and defended
- Consultation framed as information-gathering, not permission-seeking
- Documents reasoning process before, during, after decision
- Operates with intellectual independence even under hierarchical pressure

Evidence Requirements:

- Journal: Consistent pattern of independent reasoning across 10+ entries
- Interviews: Demonstrates structured decision methodology
- Decisions: ≥70% show independent execution with documented rationale

3.2 PILLAR 2: MASTERY/COMPETENCE

Construct Definition: Continuous refinement of skills, decision processes, and analytical precision through longitudinal development.

Stage 1: Static

Threshold: No documented improvement in decision accuracy across 3+ assessment points; error rates stable above 20%

Behavioral Indicators:

- Repeats same analytical errors across multiple decisions
- No evidence of incorporating feedback or lessons learned
- Decision accuracy remains constant or declines
- No systematic skill development protocols documented
- Error patterns persist without correction attempts

Evidence Requirements:

- Journal: Absence of improvement reflections or skill development plans
- Interviews: Cannot identify recent skill gains or error corrections
- Decisions: Error rate >20% across 3+ consecutive assessments with no improvement trend

Stage 2: Incremental

Threshold: Improvements in 1-2 domains; error rates reduced 10-19% over baseline

Behavioral Indicators:

- Documents specific skill development areas
- Implements corrective actions for identified errors
- Shows measurable improvement in targeted domains
- Inconsistent progression across competency areas
- Some error patterns corrected, others persist

Evidence Requirements:

- Journal: 3+ entries documenting skill development efforts and results
- Interviews: Identifies specific improvements with examples
- Decisions: Error rate reduction of 10-19% in 1-2 competency areas

Stage 3: Mastery-Oriented

Threshold: Refinements in 3+ domains; error rates <10% with explicit correction protocols

Behavioral Indicators:

- Systematic tracking of decision accuracy and error sources
- Implements correction protocols immediately upon error identification
- Progressive improvement across multiple competency dimensions
- Documents lessons learned and applies to future decisions
- Seeks increasingly complex challenges to develop new skills

Evidence Requirements:

- Journal: Comprehensive skill development documentation across 10+ entries
- Interviews: Articulates multi-domain improvement with specific metrics
- Decisions: Error rate <10% with documented correction protocols for all errors

3.3 PILLAR 3: PURPOSE/RELATEDNESS

Construct Definition: Leadership driven by clearly defined purpose aligned with organizational objectives; outcome-focused vs. process-focused orientation.

Stage 1: Process-Driven

Threshold: <40% of priorities articulated in measurable outcome terms

Behavioral Indicators:

- Focuses on activity completion ("we need to hold meetings")
- Justifies actions by effort invested, not results achieved
- Defines success as task completion, not objective attainment
- Lacks measurable success criteria for initiatives
- Confuses busyness with productivity

Evidence Requirements:

- Journal: Priorities framed as tasks/activities (not outcomes) in 60%+ of entries
- Interviews: Unable to specify measurable success criteria when probed
- Decisions: <40% include explicit outcome metrics or organizational alignment

Stage 2: Transitional

Threshold: 40-69% outcome-focused priorities; inconsistent application

Behavioral Indicators:

- Some decisions tied to measurable objectives
- Inconsistent use of outcome metrics across decisions
- Can articulate organizational alignment when prompted
- Mixes activity-based and outcome-based reasoning
- Recognizes distinction but struggles with consistent application

Evidence Requirements:

- Journal: 40-69% of entries frame priorities as outcomes
- Interviews: Can define success criteria for some (not all) initiatives
- Decisions: 40-69% include outcome metrics or organizational objectives

Stage 3: Outcome-Driven

Threshold: ≥70% of priorities with clear metrics aligned to organizational objectives

Behavioral Indicators:

- Consistently defines success criteria before action
- Aligns all major decisions to strategic objectives
- Distinguishes activity from achievement systematically
- Tracks progress against outcome metrics, not task completion
- Eliminates non-value-adding activities explicitly

Evidence Requirements:

- Journal: ≥70% of priorities articulated with measurable outcomes
- Interviews: Immediately specifies success criteria for all initiatives
- Decisions: ≥70% include explicit metrics and organizational alignment

3.4 PILLAR 4: CONSISTENCIES

Construct Definition: Vision-focused consistency in decision-making and strategic direction; iterative refinement maintaining alignment (not rigid repetition).

Stage 1: Reactive

Threshold: Vision alignment in <50% of decisions; strategic shifts in >2 observations

Behavioral Indicators:

- Decisions shift based on immediate pressures or latest information
- No discernible strategic throughline across choices
- Course corrections lack connection to original vision
- Responds to urgency rather than importance
- Strategic direction changes frequently without explanation

Evidence Requirements:

- Journal: Contradictory priorities or unexplained direction shifts in 50%+ entries
- Interviews: Cannot articulate consistent vision across decisions
- Decisions: <50% align with previously stated strategic direction

Stage 2: Structured but Rigid

Threshold: 50-79% vision alignment via inflexible rules; no documented iterative refinements

Behavioral Indicators:

- Maintains consistency through rigid protocols (not vision)
- Resists adaptation even when context changes
- "We've always done it this way" reasoning
- Consistency achieved via inflexibility, not strategic alignment
- Course corrections seen as failures, not refinements

Evidence Requirements:

- Journal: Consistency maintained through rules/procedures (not vision alignment)
- Interviews: Defends consistency based on precedent, not strategic rationale
- Decisions: 50-79% align with past choices but lack refinement documentation

Stage 3: Vision-Consistent

Threshold: ≥80% vision alignment with refinements documented in 70%+ of course corrections

Behavioral Indicators:

- Decisions consistently advance stated strategic vision
- Course corrections explicitly tied to vision refinement
- Adaptation framed as "repetition of attempt" toward same goal
- Strategic direction maintained despite tactical flexibility
- Explains refinements as improved paths to unchanged objectives

Evidence Requirements:

- Journal: ≥80% of decisions explicitly tied to vision; refinements explained as improvements
- Interviews: Articulates vision and demonstrates decision alignment
- Decisions: ≥80% advance strategic vision with 70%+ of corrections documented as refinements

3.5 PILLAR 5: ACCURACIES

Construct Definition: Accurate information gathering and sharing; prioritizing precision over confident projection.

Stage 1: Confidence-Driven

Threshold: "I don't know" statements in <20% of uncertain scenarios

Behavioral Indicators:

- Projects certainty even when lacking information
- Rarely acknowledges knowledge gaps or uncertainty
- Prioritizes confident presentation over factual accuracy
- Defends positions despite contradictory evidence
- Mistakes confidence for competence

Evidence Requirements:

- Journal: Minimal acknowledgment of uncertainty across entries
- Interviews: <20% of uncertain questions met with "I don't know" or uncertainty acknowledgment
- Decisions: <30% revised when new contradictory data emerges

Stage 2: Accuracy-Aware

Threshold: 20-59% uncertainty acknowledgment; some data-driven revisions

Behavioral Indicators:

- Acknowledges knowledge gaps when directly questioned
- Sometimes revises positions based on new data
- Struggles to balance accuracy with confidence requirements
- Recognizes distinction but inconsistent application
- Fact-checks some (not all) claims before dissemination

Evidence Requirements:

- Journal: 20-59% of entries acknowledge uncertainty or information gaps

- Interviews: 20-59% of uncertain questions met with uncertainty acknowledgment
- Decisions: 30-69% revised when contradictory data presented

Stage 3: Accuracy-Prioritized

Threshold: ≥60% uncertainty acknowledgment; fact-checking in 80%+ of disseminations

Behavioral Indicators:

- Readily states "I don't know" when lacking information
- Consistently revises positions when new data emerges
- Fact-checks claims before dissemination
- Prioritizes being correct over appearing confident
- Acknowledges confidence levels explicitly ("high confidence" vs. "speculative")

Evidence Requirements:

- Journal: ≥60% of uncertain topics acknowledged as uncertain
- Interviews: ≥60% of uncertain questions met with appropriate uncertainty acknowledgment
- Decisions: ≥70% revised when contradictory evidence presented; 80%+ fact-checked before dissemination

3.6 PILLAR 6: EFFICIENCIES

Construct Definition: Execution with minimal resource waste; optimal allocation and lean execution frameworks.

Stage 1: Resource-Inefficient

Threshold: Cost-benefit analysis in <40% of resource allocations; waste evident in 3+ observations

Behavioral Indicators:

- Allocates resources without systematic analysis
- Fails to eliminate known inefficiencies
- No documented optimization methodologies
- Resource waste patterns persist across decisions
- Prioritizes comprehensiveness over optimization

Evidence Requirements:

- Journal: <40% of resource decisions include cost-benefit reasoning
- Interviews: Cannot articulate resource optimization criteria
- Decisions: Documented waste in 3+ resource allocations

Stage 2: Efficiency-Conscious

Threshold: 40-69% include cost-benefit analysis; partial optimization

Behavioral Indicators:

- Sometimes conducts cost-benefit analysis

- Identifies some inefficiencies but incomplete elimination
- Inconsistent application of optimization methods
- Resource allocation improving but not systematized
- Aware of efficiency principles but irregular implementation

Evidence Requirements:

- Journal: 40-69% of resource decisions include cost-benefit reasoning
- Interviews: Identifies some optimization opportunities
- Decisions: 40-69% show cost-benefit analysis; some waste eliminated

Stage 3: Optimized

Threshold: ≥70% include cost-benefit analysis; lean allocation in 80%+ of decisions

Behavioral Indicators:

- Systematic cost-benefit analysis before all major allocations
- Actively eliminates identified inefficiencies
- Documents resource optimization methodology
- Lean execution protocols applied consistently
- Tracks resource utilization and improves over time

Evidence Requirements:

- Journal: ≥70% of resource decisions include explicit cost-benefit analysis
- Interviews: Articulates optimization methodology with examples
- Decisions: ≥70% show cost-benefit analysis; 80%+ demonstrate lean allocation

3.7 PILLAR 7: SOUND THINKING

Construct Definition: Structured cognitive processes rejecting emotional impulse and ideological entrenchment; epistemic flexibility and contrastive evaluation.

Stage 1: Reactive Thinking

Threshold: Logical reasoning in <30% of decision explanations; emotion-dominant in journals

Behavioral Indicators:

- Decisions justified by emotional reactions or urgency
- Ideological positions override empirical evidence
- Minimal use of structured analytical frameworks
- Explanations rely on intuition or "gut feeling"
- Emotional language dominates strategic reasoning

Evidence Requirements:

- Journal: <30% of entries demonstrate structured reasoning; emotion-focused language dominant
- Interviews: Justifies decisions emotionally or ideologically
- Decisions: <30% include logical frameworks or contrastive analysis

Stage 2: Structured but Incomplete

Threshold: 30-69% framework use; inconsistent contrastive inquiry application

Behavioral Indicators:

- Uses analytical frameworks for some decisions
- Applies contrastive inquiry sporadically
- Structured reasoning in low-stakes decisions but reverts to emotion under pressure
- Recognizes cognitive biases but doesn't systematically address them
- Mixes logical analysis with emotional justification

Evidence Requirements:

- Journal: 30-69% of entries show framework usage or bias recognition
- Interviews: Can explain reasoning frameworks but inconsistent application
- Decisions: 30-69% include structured analysis; contrastive inquiry in <50%

Stage 3: Cognitively Disciplined

Threshold: ≥70% logical reasoning; alternatives considered in 80%+ of problems

Behavioral Indicators:

- Systematic use of structured reasoning frameworks
- Contrastive inquiry applied consistently ("What if opposite is true?")
- Documents alternative hypotheses before concluding
- Separates fact from inference explicitly
- Identifies and corrects cognitive biases in real-time
- Maintains analytical discipline under pressure

Evidence Requirements:

- Journal: ≥70% of entries demonstrate structured frameworks and bias awareness
- Interviews: Articulates multi-framework analysis with contrastive thinking
- Decisions: ≥70% show logical reasoning; 80%+ document alternatives considered

3.8 PILLAR 8: ACCURATE DECISIONS

Construct Definition: Correct decisions via structured contrastive analysis and logical rigor; outcome accuracy over decisiveness speed.

Stage 1: Decisiveness-Focused

Threshold: Correctness rate <60%; speed prioritized over accuracy

Behavioral Indicators:

- Makes quick decisions without thorough analysis
- Values "being decisive" more than "being right"
- Minimal decision validation or error tracking
- Defends incorrect decisions rather than revising

- No systematic prediction-outcome tracking

Evidence Requirements:

- Journal: Emphasizes decisiveness; minimal error analysis
- Interviews: Prioritizes speed over accuracy when discussing decisions
- Decisions: <60% produce intended outcomes; errors not tracked

Stage 2: Accuracy-Seeking

Threshold: 60-79% correctness; validation in 50%+ of protocols

Behavioral Indicators:

- Uses some validation methodologies before deciding
- Tracks outcomes for major decisions
- Revises some incorrect decisions
- Improving but not systematized accuracy protocols
- Balances speed with correctness inconsistently

Evidence Requirements:

- Journal: 50%+ of decisions include validation reasoning
- Interviews: Can identify recent errors and corrections
- Decisions: 60-79% correctness rate; validation in 50%+ of protocols

Stage 3: Decision-Accurate

Threshold: ≥80% correctness; prediction errors analyzed in all cases

Behavioral Indicators:

- Systematic decision validation before execution
- Tracks predicted vs. actual outcomes for all major decisions
- Analyzes all errors to identify causes and corrections
- Implements bias-mitigation protocols (e.g., devil's advocate)
- Prioritizes accuracy over speed consistently

Evidence Requirements:

- Journal: Comprehensive prediction-outcome tracking and error analysis
- Interviews: Articulates decision validation methodology with examples
- Decisions: ≥80% correctness rate; all errors analyzed with documented corrections

3.9 PILLAR 9: EFFECTIVE COMMUNICATION

Construct Definition: Structured, outcome-driven communication eliminating ambiguity; Three-Part Model (What/Why/Success Criteria).

Stage 1: Vague/Inspirational

Threshold: Success criteria specified in <40% of directives

Behavioral Indicators:

- Communication relies on motivational messaging
- Lacks specific action steps or success metrics
- Ambiguous directives subject to multiple interpretations
- Inspirational language without operational clarity
- Receivers unclear on expected outcomes

Evidence Requirements:

- Journal: <40% of communications include success criteria
- Interviews: Struggles to define "what success looks like" when asked
- Decisions: <40% of directives specify measurable outcomes

Stage 2: Structured but Incomplete

Threshold: 40-69% use Three-Part Model (What/Why/Success)

Behavioral Indicators:

- Sometimes provides clear success criteria
- Inconsistent use of What/Why/Success framework
- More structured than inspirational but still gaps
- Receivers sometimes unclear on expectations
- Improving but not systematized

Evidence Requirements:

- Journal: 40-69% of communications use Three-Part Model
- Interviews: Can explain framework but inconsistent application
- Decisions: 40-69% of directives include What/Why/Success elements

Stage 3: Outcome-Communicative

Threshold: ≥70% use Three-Part Model; actionable in 80%+ of communications

Behavioral Indicators:

- Consistently provides What/Why/Success in all directives
- Communications are actionable and unambiguous
- Receivers can restate expectations accurately
- Eliminates vague or inspirational-only messaging
- Success criteria specified before action initiated

Evidence Requirements:

- Journal: ≥70% of communications document Three-Part Model usage
- Interviews: Articulates framework and demonstrates consistent application
- Decisions: ≥70% of directives include What/Why/Success; 80%+ rated actionable by receivers

---

4. EVIDENCE COLLECTION PROTOCOL

4.1 Data Sources and Triangulation

Each leader assessment draws from three evidence sources, analyzed independently before synthesis:

1. **Reflective Journals** - Leader-generated written documentation
2. **Structured Interviews** - Evaluator-conducted qualitative check-ins
3. **Decision Documentation** - Real-world decision analysis

Triangulation Principle: No categorical assignment based on single source; minimum two sources must support assignment with 70% preponderance of evidence.

4.2 Reflective Journal Requirements

Submission Frequency: Bi-weekly (every 14 days)

Minimum Content Per Submission:

- 3 entries per pillar per quarter = 27 total entries per quarter
- Each entry: 200-500 words
- Required elements:
    - Date and context (situation/decision)
    - Pillar-relevant behavior or cognitive process
    - Outcome (if applicable)
    - Reflection on improvement or challenge

Analysis Protocol:

- Evaluator reviews all journal entries for assessment period (3 months)
- Extracts dated examples of pillar-relevant behaviors
- Categorizes entries by pillar and developmental stage indicators
- Documents contradictory evidence (e.g., autonomous behavior in one entry, dependent in another)
- Weights evidence by specificity and recency

4.3 Structured Interview Protocol

Frequency: Quarterly (every 3 months)

Duration: 90 minutes per interview

Format: Semi-structured with standardized question protocol and probing flexibility

Interview Structure:

Part 1: Recent Decisions Review (30 minutes)

- "Describe 3 significant decisions you made in the last quarter."
- For each decision, probe:
    - What alternatives did you consider?

- How did you decide among them?
- What information did you seek?
- Who did you consult and why?
- What was the outcome?
- What would you do differently?

Part 2: Pillar-Specific Probes (45 minutes - 5 min per pillar)

- **Autonomy:** "Describe a decision you made without external validation. Why didn't you seek input?"
- **Mastery:** "What skill have you improved most this quarter? How do you know it improved?"
- **Purpose:** "How do you distinguish between activity and achievement? Give an example from this quarter."
- **Consistencies:** "Describe a time you changed course. How did it align with your vision?"
- **Accuracies:** "When did you last say 'I don't know'? What happened next?"
- **Efficiencies:** "Describe a resource allocation decision. What trade-offs did you consider?"
- **Sound Thinking:** "Walk me through a recent complex problem. What frameworks did you use?"
- **Accurate Decisions:** "Describe a decision that didn't produce the outcome you predicted. Why?"
- **Effective Communication:** "How did you communicate that decision? What, why, and what success looks like?"

Part 3: Developmental Reflection (15 minutes)

- "Where have you seen the most growth this quarter?"
- "What area needs more development?"
- "What patterns are you noticing in your decision-making?"

Audio Recording and Transcription:

- All interviews audio-recorded with leader consent
- Professional transcription within 7 days
- Evaluator reviews transcript and extracts pillar-relevant evidence

4.4 Decision Documentation Analysis

Submission Requirement: Leaders submit 3-5 decision memos per quarter

Decision Memo Template:

1. **Context:** Situation requiring decision (100-200 words)
2. **Alternatives Considered:** Minimum 2 options with pros/cons (200-300 words)
3. **Decision Rationale:** Why chosen option selected (150-250 words)
4. **Expected Outcome:** Predicted results with success metrics (100-150 words)
5. **Actual Outcome:** What happened; comparison to prediction (100-200 words)

Evaluator Analysis:

- Reviews each decision memo for pillar-specific evidence
- Calculates accuracy rates (predicted vs. actual outcomes)
- Identifies patterns in reasoning quality
- Documents framework usage (or absence)
- Extracts quantifiable thresholds (e.g., % decisions showing cost-benefit analysis)

Acceptable Alternative: If leader doesn't create decision memos, evaluator may analyze:

- Meeting minutes where leader made decisions
- Email trails documenting decision processes
- Strategy documents authored by leader
- Performance reviews mentioning specific decisions

Critical Standard: Real-world decisions, not hypothetical scenarios or self-assessments

4.5 Evidence Synthesis and Categorical Assignment

Step 1: Evidence Compilation

Evaluator creates dossier for each leader containing:

- Journal entry database (organized by pillar)
- Interview transcript excerpts (pillar-tagged)
- Decision analysis summaries (pillar-tagged)

Step 2: Per-Pillar Analysis

For each of the nine pillars, evaluator:

1. Reviews all evidence from three sources
2. Calculates quantifiable thresholds where applicable (e.g., % independent decisions)
3. Identifies behavioral exemplars matching categorical stage indicators
4. Documents contradictory evidence
5. Determines preponderance (minimum 70% of evidence must support assignment)

Step 3: Categorical Assignment

- Assign to ONE category per pillar based on preponderance
- Write 100-200 word justification citing specific dated evidence
- Flag boundary cases (evidence nearly equal across two categories)

Step 4: Quality Control

- Verify minimum evidence standards met (3+ examples per pillar)
- Check for unsubstantiated assignments (must cite evidence)
- Confirm triangulation (two sources minimum)

Step 5: Documentation

Submit standardized assessment report:

- Categorical assignments for all nine pillars

- Evidence justification for each assignment
- Contradictory evidence noted
- Confidence level (high/medium/low) per assignment

---

5. SAMPLE STRUCTURE AND COHORT DESIGN

5.1 Cohort Stratification

Leaders stratified into five temporal cohorts based on time since Reasoned Development training initiation:

Cohort	Time Since Training	Target N	Purpose
A	0 months (pre-training baseline)	50	Establish baseline distributions
B	6 months post-initiation	50	Initial developmental assessment
C	12 months post-initiation	50	Intermediate progression
D	18 months post-initiation	50	Advanced development
E	24+ months post-initiation	50	Sustained maintenance/mastery

Total Sample: N = 250 leaders

5.2 Sample Size Justification

Statistical Power Calculation:

For chi-square test of independence on 3 categories × 5 cohorts contingency table:

- Degrees of freedom: (3-1) × (5-1) = 8
- Effect size: w = 0.3 (medium effect per Cohen, 1988)
- Significance level: α = 0.05
- Power target: 0.80 (conventional standard)

Required Sample Size: N = 242 (G*Power 3.1 calculation)

Protocol Specification: N = 250 (50 per cohort) achieves power ≈ 0.95

Design Target: The protocol is optimized for detecting moderate effect sizes (Cramér's w ≈ 0.30). This threshold reflects a focus on practically meaningful developmental change rather than marginal differences. Smaller effects (w < 0.20) are intentionally outside the design scope; if pilot studies suggest effects in this range, sample expansion would be required for adequate power, but such small effects may not represent practically significant development.

Rationale for Oversampling:

- Anticipates 15-20% attrition over 24-month study period
- Allows for exclusions due to incomplete evidence
- Provides buffer for unbalanced cohort distributions
- Enables subgroup analyses (e.g., by industry, organizational size)

Power Analysis Across Effect Sizes:

Effect Size (w)	Classification	Required N (80% power)	Actual Power at N=250
0.10	Small	785	0.45 (underpowered)
0.15	Small	349	0.70
0.20	Small-Medium	196	0.88
0.25	Small-Medium	125	0.93
0.30	Medium	87	0.95
0.40	Medium-Large	49	0.99
0.50	Large	31	>0.99

Interpretation: Protocol optimized for medium effects ($w \geq 0.3$). Small effects ($w < 0.2$) may be undetectable. If pilot study reveals $w \approx 0.15$-0.20, sample expansion to N=350 recommended for adequate power.

5.3 Recruitment and Inclusion Criteria

Inclusion Criteria:

1. Current leadership role with decision-making authority
2. Minimum 2 years leadership experience
3. No prior exposure to Reasoned Leadership or Reasoned Development training (except Cohort A baseline)
4. Organizational commitment to support evidence collection (journals, interviews, decision documentation)
5. Informed consent for assessment and data use

Exclusion Criteria:

1. Leadership tenure <2 years (insufficient experience base)
2. Prior participation in similar cognitive restructuring programs (confounding)
3. Anticipated organizational departure within 12 months (attrition risk)
4. Inability to provide reflective journals in English (standardization)
5. Refusal of audio-recorded interviews (evidence requirement)

Recruitment Strategy:

- Partner with 8-12 organizations implementing Reasoned Development training
- Recruit across industries (technology, healthcare, finance, manufacturing, nonprofit)
- Balance organizational sizes (small <100 employees, medium 100-1000, large >1000)
- Diversity targets: 40% women, 30% BIPOC, 15% international (outside US)

5.4 Attrition Management

Anticipated Attrition: 15-20% over 24 months due to:

- Job changes/organizational departure
- Voluntary withdrawal
- Incomplete evidence submission
- Evaluator concerns about data quality

Mitigation Strategies:

1. **Oversampling:** Initial N=60 per cohort to maintain N=50 after attrition
2. **Engagement Protocols:**
    - Quarterly feedback reports to leaders (non-evaluative, developmental)
    - Organizational liaisons to facilitate evidence collection
    - Incentives: Certificate of participation, developmental summary report
3. **Replacement Recruitment:**
    - Rolling enrollment for Cohorts A-C (early stage)
    - No replacement for Cohorts D-E (late stage)

Analysis Impact:

- Compare completers vs. non-completers on baseline characteristics
- Test for differential attrition across cohorts (would bias developmental patterns)
- Sensitivity analysis: re-run chi-square excluding partial-data cases

5.5 Missing Data Protocol

Evidence Sufficiency Threshold: Leaders providing less than 50% of required admissible evidence are excluded from primary analyses.

Handling Procedures:

1. Document reason for evidence insufficiency (attrition, non-compliance, data quality)
2. Compare excluded cases to included cases on available baseline characteristics
3. Test whether exclusions are systematic (e.g., more exclusions in later cohorts would indicate attrition bias)
4. Report exclusion rates by cohort

Sensitivity Analyses:

- Re-run primary analyses with different exclusion thresholds (40%, 60%)
- Test whether conclusions change based on inclusion decisions
- If excluded cases differ systematically, acknowledge this as a limitation

---

6. CHI-SQUARE ANALYSIS PROCEDURE

6.0 Statistical Role and Interpretation

Purpose: The Chi-Square Twist is used to examine whether the distribution of developmental stages differs across temporally stratified cohorts. It functions as a pattern detection and falsification screening mechanism rather than a causal estimator.

Interpretation Framework: Chi-square results are interpreted as descriptive indicators of whether cohort differences exist, not as definitive proof of training effects. This protocol acknowledges that strict independence assumptions underlying chi-square may be violated due to:

- **Organizational clustering:** Leaders from the same organization may share unmeasured characteristics
- **Evaluator effects:** The same evaluators assess multiple leaders, potentially introducing correlated errors

Accordingly, statistically significant chi-square results indicate patterns worthy of further investigation, not causal conclusions. Non-significant results provide falsification evidence against claims of detectable developmental patterns.

6.1 Contingency Table Construction

For each of the Nine Pillars, construct 3 × 5 contingency table:

Example: Pillar 1 (Autonomy)

	Cohort A	**Cohort B**	**Cohort C**	**Cohort D**	**Cohort E**	**Row Total**
Dependent	n11	n12	n13	n14	n15	R1
Emerging	n21	n22	n23	n24	n25	R2
Autonomous	n31	n32	n33	n34	n35	R3
Column Total	C1	C2	C3	C4	C5	N=250

Where:

- nij = observed frequency in cell (row i, column j)
- Ri = row total (sum across cohorts for category i)
- Cj = column total (sum across categories for cohort j)
- N = grand total (250 leaders)

6.2 Expected Frequency Calculation

For each cell, calculate expected frequency under null hypothesis of independence:

Eij = (Ri × Cj) / N

Example:

- Row 1 total (Dependent) = 80
- Column 2 total (Cohort B) = 50
- Expected frequency for Dependent-Cohort B cell: E12 = (80 × 50) / 250 = 16

Assumption Check: All expected frequencies should be ≥5 for valid chi-square test.

If any Eij < 5: Use Fisher's exact test instead of chi-square (see Section 6.4)

6.3 Chi-Square Test Statistic

Formula: $\chi^2 = \Sigma[(Oij - Eij)^2 / Eij]$

Where summation is across all cells (i = 1 to 3, j = 1 to 5)

Degrees of Freedom: df = (rows - 1) × (columns - 1) = (3 - 1) × (5 - 1) = 8

Critical Value: For α = 0.05, df = 8: χ^2(critical) = 15.507

Decision Rule:

- If $\chi^2 > 15.507$ → Reject H_0 (categorical distribution depends on cohort)
- If $\chi^2 \leq 15.507$ → Fail to reject H_0 (no evidence of temporal pattern)

6.4 Fisher's Exact Test (Sparse Cell Fallback)

When to Use: If any expected cell frequency < 5, chi-square assumptions violated.

Procedure:

1. Calculate exact probability of observed table under independence assumption
2. Sum probabilities of all tables as extreme or more extreme
3. Compare to $\alpha = 0.05$ for significance

Software Implementation:

- R: fisher.test(contingency_matrix)
- Python: scipy.stats.fisher_exact() (for 2×2; generalized for larger tables)
- SPSS: Analyze → Descriptive Statistics → Crosstabs → Statistics → Fisher's exact

6.5 Multiplicity Handling

Primary Endpoints: Three pillars are designated as primary endpoints based on theoretical centrality to the Reasoned Leadership framework:

1. Autonomy (Pillar 1)
2. Sound Thinking (Pillar 7)
3. Accurate Decisions (Pillar 8)

Correction Procedure:

- **Primary endpoints:** Holm-Bonferroni correction applied (ordered p-values compared to $\alpha/3$, $\alpha/2$, α)
- **Exploratory endpoints (remaining 6 pillars):** False Discovery Rate (FDR) control via Benjamini-Hochberg procedure

Reporting:

- Report both uncorrected and corrected p-values for all pillars
- Distinguish confirmatory (primary) from exploratory (secondary) findings
- Significant exploratory findings generate hypotheses for future studies, not definitive conclusions

6.6 Residual Analysis

Purpose: Identify which cells contribute most to significant chi-square

Standardized Residual Formula: $r_{ij} = (O_{ij} - E_{ij}) / \sqrt{E_{ij}}$

Interpretation:

- $|r_{ij}| > 2$ → Cell contributes significantly to chi-square

- Positive residual → More observations than expected
- Negative residual → Fewer observations than expected

Example Residual Analysis:

Cell	Observed	Expected	Residual	Interpretation
Dependent-Cohort A	35	16	+4.75	Much higher than expected
Autonomous-Cohort A	2	16	-3.50	Much lower than expected
Dependent-Cohort E	3	16	-3.25	Much lower than expected
Autonomous-Cohort E	32	16	+4.00	Much higher than expected

Pattern Interpretation: High Dependent in Cohort A + Low Autonomous in Cohort A + Low Dependent in Cohort E + High Autonomous in Cohort E = Developmental progression pattern

6.7 Developmental Trajectory Interpretation

Expected Pattern if Training Effective:

Cohort	Dependent %	Emerging %	Autonomous %
A (0mo)	60-70%	25-30%	5-10%
B (6mo)	40-50%	35-45%	10-15%
C (12mo)	20-30%	40-50%	25-35%
D (18mo)	10-15%	30-40%	45-55%
E (24mo)	5-10%	20-25%	65-75%

Interpretation Guidelines:

- **Progressive shift:** Later cohorts show higher advanced-stage percentages
- **Gradual transition:** Not all leaders advance at same rate (hence 3-stage distribution in all cohorts)
- **Minimal regression:** Later cohorts should not show higher Dependent percentages than earlier cohorts

Alternative Patterns and Interpretations:

Pattern 1: No Significant Chi-Square

- Interpretation: No evidence of temporal development
- Implications: Training ineffective OR pillar not amenable to intervention

Pattern 2: Plateau Pattern

- Cohorts A-B show progression, but C-D-E flat
- Interpretation: Early gains followed by plateau
- Implications: Training produces initial shift but not sustained development

Pattern 3: Regression Pattern

- Cohort E shows higher Dependent % than Cohort D
- Interpretation: Skills decay without reinforcement
- Implications: Maintenance protocols needed

Pattern 4: Pillar-Specific Patterns

- Some pillars show progression (e.g., Accuracies, Effective Communication)
- Others show no pattern (e.g., Autonomy, Sound Thinking)
- Interpretation: Differential trainability of constructs
- Implications: Refine training focus on developable pillars

---

7. INDIVIDUAL LONGITUDINAL TRACKING (Validation Subsample)

7.1 Rationale

Chi-Square Twist provides pseudo-longitudinal insight via cross-sectional cohort comparison. However, true longitudinal tracking of individuals strengthens causal inference and validates cross-sectional patterns.

Subsample Design: Track 30 leaders from Cohort A baseline through all assessment points (0, 6, 12, 18, 24 months)

7.2 Transition Matrix Analysis

Objective: Document individual categorical transitions over time

Example Transition Matrix: Autonomy (6-month intervals)

	To Dependent	To Emerging	To Autonomous
From Dependent (T1)	20%	65%	15%
From Emerging (T1)	5%	45%	50%
From Autonomous (T1)	0%	15%	85%

Interpretation:

- **Advancement Rate:** 65% of Dependent → Emerging + 15% → Autonomous = 80% advance
- **Stability Rate:** 45% of Emerging stay Emerging (not all advance simultaneously)
- **Regression Rate:** 5% of Emerging → Dependent (minimal backsliding)

Validation Logic: If cross-sectional cohort analysis shows developmental progression, individual transition matrices should demonstrate:

- High advancement rates (>50% move to higher category per interval)
- Moderate stability (30-50% maintain category)
- Low regression (<10% drop to lower category)

If individual transitions contradict cohort patterns: Suggests cohort differences reflect selection bias rather than training effects.

7.3 Expected Developmental Trajectories

Hypothesis: If training effective, leaders should show progressive advancement across 24 months

Trajectory Examples:

Fast Developer:

- 0 months: Dependent
- 6 months: Emerging
- 12 months: Autonomous
- 18 months: Autonomous (maintained)
- 24 months: Autonomous (maintained)

Gradual Developer:

- 0 months: Dependent
- 6 months: Dependent (no change yet)
- 12 months: Emerging
- 18 months: Emerging (consolidating)
- 24 months: Autonomous

Non-Responder:

- 0 months: Dependent
- 6 months: Dependent
- 12 months: Dependent
- 18 months: Emerging
- 24 months: Dependent (regression)

Statistical Analysis:

- Calculate percentage of each trajectory type
- Compare to null expectation (random movement)
- Test whether advancement exceeds chance (binomial test)

---

8. VALIDITY AND RELIABILITY CONTROLS

8.1 Inter-Evaluator Reliability

Standard: Quadratic-weighted Cohen's $\kappa > 0.70$ (substantial agreement)

Rationale for Weighted Kappa: Stage assignments are ordinal, not nominal. A disagreement between Stage 1 and Stage 3 is more serious than a disagreement between Stage 1 and Stage 2. Quadratic weighting reflects this ordinal structure by penalizing larger disagreements more heavily.

Procedure:

1. Each leader assessed independently by two certified evaluators
2. Evaluators review identical evidence (journals, transcripts, decisions)
3. Evaluators assign categories independently without consultation
4. Calculate quadratic-weighted Cohen's kappa per pillar

Interpretation:

- $\kappa < 0.40$ → Poor agreement (inadequate)
- κ 0.40-0.60 → Moderate agreement (requires recalibration)
- κ 0.60-0.80 → Substantial agreement (acceptable)
- $\kappa > 0.80$ → Almost perfect agreement (excellent)

Action Thresholds:

- If $\kappa < 0.70$ for any pillar → Evaluators undergo recalibration
- If $\kappa < 0.60$ → Category definitions require revision
- If persistent $\kappa < 0.70$ → Evaluator removed from pool

Resolution of Disagreements:

When evaluators disagree on categorical assignment:

1. Both review evidence together (not to consensus, but to understand reasoning)
2. Third senior evaluator reviews case blindly
3. Majority rule (2 of 3) determines final assignment
4. Disagreement documented for analysis (systematic disagreement patterns inform protocol refinement)

Disagreements exceeding one stage (e.g., one evaluator assigns Stage 1, another assigns Stage 3) trigger mandatory adjudication regardless of whether dual-rating was planned.

8.2 Convergent Validity

Objective: Categorical assignments should correlate with objective performance metrics

Hypothesized Correlations:

Pillar	Objective Metric	Expected r
Autonomy	Decision frequency without escalation	$r > 0.40$
Mastery	Performance rating improvement over time	$r > 0.35$
Purpose	Goal attainment rate (OKR completion)	$r > 0.40$
Accuracies	Decision accuracy rate (predicted vs. actual)	$r > 0.50$
Efficiencies	Resource utilization efficiency ratio	$r > 0.45$
Accurate Decisions	Correct decision percentage	$r > 0.55$

Analysis:

- Point-biserial correlation (categorical stage as ordinal: 1=lowest, 3=highest)
- If $r <$ expected threshold → Questions construct validity

Caveat: Not all pillars have obvious objective metrics (e.g., Sound Thinking). Convergent validity tests apply where metrics available.

8.3 Discriminant Validity

Objective: Nine Pillars should assess distinct constructs, not a single "general leadership" factor

Test: Inter-pillar correlations should be modest ($r < 0.60$)

Procedure:

1. Calculate Spearman correlation between all pillar pairs (36 correlations)
2. Expected pattern:
    - Related pillars: moderate correlation (e.g., Sound Thinking ↔ Accurate Decisions: $r \approx 0.50$)
    - Distinct pillars: low correlation (e.g., Autonomy ↔ Effective Communication: $r \approx 0.25$)

Interpretation:

- If most inter-pillar $r > 0.70$ → Pillars redundant, measuring same underlying factor
- If $r < 0.60$ for most pairs → Pillars assess distinct competencies (supports framework structure)

8.4 Evaluator Blinding

Objective: Prevent expectation bias (evaluators unconsciously favoring developmental progression)

Procedure:

1. Evidence dossiers coded by ID number only (no cohort information)
2. Evaluators assign categories based on evidence alone
3. Cohort assignment revealed only after all assessments complete

Limitation: Perfect blinding impossible when evaluators conduct interviews (they interact with leaders who may mention training timeline).

Phase 1 Mitigation: Emphasize evidence documentation standards requiring multiple examples, not impressions. Acknowledge this limitation in Phase 1 publications.

Phase 2 Requirement: Separation between data collection and scoring roles. Interviewers do not score; scorers receive blinded dossiers.

8.5 Contextual Confounder Documentation

Objective: Account for organizational factors that may influence development independent of training

Variables to Document:

1. **Organizational Tenure:** Years in current organization (separate from RL training tenure)
2. **Role Changes:** Promotions, lateral moves, demotions during study period
3. **Organizational Events:** Mergers, acquisitions, leadership changes, crises
4. **External Training:** Participation in non-RL development programs
5. **Team Stability:** Turnover rate in leader's direct reports
6. **Industry Context:** Sector-specific events (e.g., regulatory changes)

Statistical Control:

- Test whether confounders predict categorical placement independent of cohort
- Stratified analysis: Re-run chi-square within homogeneous subgroups (e.g., only leaders with no role changes)

- Multivariate logistic regression: Model category as function of cohort + confounders

Transparency: All analyses report both unadjusted and confounder-adjusted results

---

9. FALSIFICATION CRITERIA

9.0 Falsification Framework

The RLQ-IBOT Protocol is designed to be falsifiable. This section specifies conditions under which (a) the measurement instrument itself should be rejected, and (b) claims of training effectiveness should be rejected. These are distinct failure modes.

Pre-Commitment: These criteria are specified before data collection. Post-hoc modifications to "rescue" null results are methodologically prohibited.

9.1 Instrument Falsification

The measurement protocol itself is falsified if any of the following occur:

Criterion 1: Reliability Failure

Quadratic-weighted κ falls below 0.70 across trained evaluators for majority ($\geq$5 of 9) pillars.

- **Interpretation:** Evaluators cannot reliably apply the categorical definitions
- **Implication:** Operational definitions are ambiguous or unworkable; protocol requires revision before any effectiveness claims are possible

Criterion 2: Convergence Failure

Stage assignments fail to converge across independent raters even after adjudication; systematic patterns of disagreement persist.

- **Interpretation:** The construct definitions do not map to observable behaviors consistently
- **Implication:** Protocol lacks measurement validity

Criterion 3: Operational Unworkability

Operational definitions prove non-replicable in practice:

- Evaluators cannot consistently identify what counts as a "decision" meeting inclusion criteria
- Thresholds cannot be calculated from available evidence
- Required evidence types are not consistently available
- **Interpretation:** The protocol is theoretically sound but practically unusable
- **Implication:** Revision required to improve feasibility

Criterion 4: Discriminant Validity Failure

Inter-pillar correlations exceed $r > 0.70$ for majority of pillar pairs.

- **Interpretation:** Pillars are not measuring distinct constructs; a single "general leadership" factor explains most variance
- **Implication:** Nine-pillar structure not supported; framework requires theoretical revision

9.2 Intervention Ineffectiveness

When the protocol is applied to evaluate Reasoned Leadership training, intervention ineffectiveness is indicated by:

Criterion 5: No Temporal Pattern

Chi-square test yields $p > 0.05$ (after multiplicity correction) for majority (≥5 of 9) pillars, including at least 2 of 3 primary endpoints.

- **Interpretation:** Categorical distributions do not differ across cohorts; no evidence of temporal development
- **Implication:** Training does not produce measurable progression on most constructs

Criterion 6: Regression Pattern

Later cohorts show higher percentages in lowest category than earlier cohorts.

- **Example:** Cohort E (24 months) has 40% Dependent while Cohort B (6 months) has 25% Dependent
- **Interpretation:** Skills decay over time or later cohorts differ systematically from earlier cohorts
- **Implication:** Training produces no sustained benefit OR sampling bias invalidates design

Criterion 7: Individual Trajectory Contradiction

Longitudinal subsample (N=30) shows:

- <30% advancement rate per 6-month interval
    - 20% regression rate
- Random walk pattern (equal probability of advancing, staying, regressing)
- **Interpretation:** Individual development does not follow expected trajectory; cross-sectional cohort differences reflect selection effects rather than training
- **Implication:** Pseudo-longitudinal inference invalidated by true longitudinal data

Criterion 8: No Convergent Validity

Categorical assignments show $r < 0.20$ with objective performance metrics across all pillars where metrics available.

- **Interpretation:** Categories do not predict real-world outcomes; assessment captures evaluator perceptions rather than actual competency
- **Implication:** Protocol measures something, but not leadership effectiveness

Criterion 9: Evaluator Bias Confirmation

Phase 2 external evaluators (non-RL practitioners) show quadratic-weighted $\kappa < 0.40$ with Phase 1 internal evaluators.

- **Interpretation:** RL-trained evaluators apply framework-specific standards not recognized by independent observers
- **Implication:** Categorical assignments reflect evaluator training bias rather than leader behavior

Criterion 10: Negligible Effect Size

Cohen's $w < 0.20$ for ≥6 of 9 pillars (even if $p < 0.05$).

- **Interpretation:** Statistically significant but practically trivial effects
- **Implication:** Training produces detectable but meaningless change

Criterion 11: Insufficient Progression Magnitude

Cohort E (24+ months) shows <20 percentage point increase in advanced category vs. Cohort A baseline.

- **Example:** Cohort A = 10% Autonomous, Cohort E = 25% Autonomous (15pp gain < 20pp threshold)
- **Interpretation:** Minimal developmental shift despite 24 months exposure
- **Implication:** Training produces weak or plateau effects

Criterion 12: Quantified Regression Threshold

Any later cohort shows ≥10 percentage point higher concentration in lowest category than any earlier cohort.

- **Example:** Cohort E = 30% Dependent, Cohort B = 18% Dependent (12pp regression exceeds 10pp threshold)
- **Interpretation:** Skills decay or later cohorts systematically weaker
- **Implication:** Training produces no sustained benefit OR selection bias invalidates design

9.3 Pre-Specified Effect Size Thresholds

What constitutes meaningful development?

Not all statistically significant chi-square results indicate practically important effects. Pre-specify minimum effect sizes:

Cohen's w (Effect Size for Chi-Square):

- $w < 0.10$ → Trivial effect (statistically significant but practically meaningless)
- w 0.10-0.30 → Small effect (detectable but modest)
- w 0.30-0.50 → Medium effect (practically significant)
- $w > 0.50$ → Large effect (strong developmental pattern)

Protocol Standard: Claim developmental effectiveness only if $w \geq 0.30$ (medium effect)

Calculation: $w = \sqrt{(\chi^2 / N)}$

Example:

- $\chi^2 = 25.0$, $N = 250$
- $w = \sqrt{(25.0 / 250)} = \sqrt{0.10} = 0.316$ (medium effect)

---

10. LIMITATIONS

10.1 Causal Inference Constraints

Limitation: Pseudo-longitudinal design cannot establish causality

Explanation: Chi-Square Twist compares different cohorts at different time points (cross-sectional), not the same individuals over time (longitudinal). Cohort differences may reflect:

- Training effects (intended inference)
- Selection bias (later cohorts differ systematically from earlier)
- Maturation (leaders improve naturally with experience)
- Historical effects (organizational changes differentially affect cohorts)
- Attrition bias (non-completers differ from completers)

Mitigation:

- Individual longitudinal subsample (N=30) provides true developmental data
- Document confounders (organizational tenure, role changes, events)
- Statistical controls in analysis
- Transparent reporting of alternative explanations

Honest Statement: This protocol detects temporal patterns consistent with developmental progression but cannot definitively prove training causes observed changes. Causal claims require:

- Randomized controlled trial (RCT) comparing trained vs. untrained leaders
- Propensity score matching to create equivalent comparison groups
- Regression discontinuity design if RCT infeasible

10.2 Independence Assumption Violations

Limitation: Chi-square assumes independent observations; this assumption may be violated.

Sources of Dependence:

- **Organizational clustering:** Leaders from the same organization share unmeasured characteristics (culture, resources, leadership climate)
- **Evaluator effects:** Same evaluators assess multiple leaders, potentially introducing correlated errors

Mitigation:

- Acknowledge chi-square results as descriptive pattern indicators, not definitive inferential claims
- Report clustering structure (how many leaders per organization)

- Consider sensitivity analyses with organization as random effect (if sample size permits)

Honest Statement: Strict statistical inference from chi-square requires independence. This protocol uses chi-square as a transparent, interpretable pattern detection tool while acknowledging that p-values may be anti-conservative (true Type I error rate may exceed nominal α).

10.3 Categorical Simplification

Limitation: Three discrete categories reduce complex behavioral continua to simplified stages

Explanation: Leaders near category boundaries may be misclassified due to:

- 70% preponderance threshold is arbitrary
- Behavioral variability across contexts (performs autonomously in familiar domains, dependently in novel ones)
- Temporal fluctuation (regression during high-stress periods)

Mitigation:

- Document boundary cases and confidence levels
- Consider 4-category option for pillars with high boundary ambiguity (e.g., Sound Thinking: add "Transitional" stage)
- Report inter-evaluator agreement on boundary cases separately

Trade-off: Categorical approach gains interpretability and statistical power (vs. continuous scales) but loses granular information.

10.4 Evaluator Subjectivity

Limitation: Categorical assignment involves professional judgment despite quantifiable anchors

Explanation: Even with operational thresholds (e.g., "70% of decisions"), evaluators must:

- Decide what counts as a "decision" vs. routine action
- Interpret quality of evidence (journal entry depth, interview candor)
- Weight contradictory evidence
- Apply context to determine behavioral significance

Phase 1 Specific Concern: RL-trained evaluators may unconsciously favor positive trajectories to validate training.

Mitigation:

- Dual independent assessment with $\kappa > 0.70$ requirement
- Blinding to cohort assignment
- Evidence documentation standards (specific dated examples required)
- Phase 2 external evaluator comparison

Acknowledged Reality: Complete objectivity impossible in behavioral assessment. Protocol prioritizes transparency and reliability over false claims of objectivity.

10.5 Generalizability Constraints

Limitation: Sample limited to organizations implementing Reasoned Development training

Explanation:

- Organizations self-select into RL training (may differ from general population)
- Leaders volunteer for assessment (selection bias)
- Sample may overrepresent certain industries, organizational cultures
- Results may not generalize to:
    - Leaders in non-implementing organizations
    - Different cultural contexts (international)
    - Public sector vs. private sector
    - Crisis vs. stable organizational environments

Mitigation:

- Recruit across diverse industries and organizational sizes
- Document sample characteristics thoroughly
- Report boundary conditions explicitly
- Encourage replication studies in different contexts

Scope Limitation: Protocol establishes measurement standard for RL development, not universal leadership assessment tool.

10.6 Evidence Collection Burden

Limitation: Quarterly interviews, bi-weekly journals, decision documentation require significant leader and evaluator time

Explanation:

- Leaders: ~10 hours per quarter (journals + interview + decision memos)
- Evaluators: ~8 hours per leader per quarter (review + interview + analysis)
- For N=250: 2,000 evaluator hours per quarter

Practical Implications:

- High cost (evaluator time at professional rates)
- Attrition risk if burden excessive
- May limit scalability

Mitigation:

- Organizational commitment secured before enrollment
- Developmental feedback provided to leaders (value exchange)
- Streamlined evidence formats
- Technology support (e.g., voice-to-text for journals)

Trade-off: Rigorous behavioral assessment requires substantial evidence. Reducing burden risks validity.

Future Extensions: Reduced-burden protocol variants may be developed for contexts where full protocol is impractical. Such variants are outside the scope of this specification and would require separate validation.

10.7 Statistical Power for Subgroup Analysis

Limitation: N=250 total provides adequate power for overall chi-square but limited power for subgroup comparisons

Explanation:

Overall analysis: N=250, power ≈ 0.95 (excellent)

Subgroup analysis examples:

- By industry: N=30-50 per industry, power drops to 0.60-0.70
- By gender: N=100 women, N=150 men, power ≈ 0.75-0.80
- Organizational size: N=80 small, N=90 medium, N=80 large, power ≈ 0.70

Implication: Subgroup analyses exploratory, not definitive. May miss real differences (Type II error risk).

Mitigation:

- Report subgroup results as preliminary
- Recommend larger samples for confirmatory subgroup studies
- Use effect size confidence intervals (not just p-values)

---

11. PUBLICATION AND DISSEMINATION STRATEGY

11.1 Manuscript Sequence

Manuscript 1: Methodological Paper

Title: "The RLQ-IBOT Protocol: A Qualitative-Statistical Hybrid for Assessing Leadership Development Trajectories"

Target Journals:

- Journal of Leaderology and Applied Leadership (primary)
- Organizational Research Methods (secondary)
- The Leadership Quarterly (tertiary)

Content:

- IBOT Method detailed specification
- Nine Pillars operationalization with categorical anchors
- Chi-Square Twist application to qualitative categories

- Evaluator training curriculum
- Inter-evaluator reliability standards
- Validity framework and falsification criteria
- Pilot demonstration with N=30 (feasibility, not full validation)

Contribution: Methodological innovation bridging qualitative behavioral depth with statistical temporal analysis. Addresses limitations of self-report instruments.

Status: Methods paper requires pilot demonstration, not full empirical validation.

Manuscript 2: Pilot Study Results

Title: "Developmental Trajectories in Reasoned Leadership: Pilot Validation of the Nine Pillars Framework"

Target Journals:

- Journal of Leaderology and Applied Leadership (primary)
- Leadership & Organization Development Journal (secondary)
- Human Resource Development Quarterly (tertiary)

Content:

- Pilot sample characteristics (N=150, compressed timeline)
- Chi-square results for each pillar
- Differential development rates across pillars
- Inter-evaluator reliability achieved
- Preliminary convergent validity evidence
- Limitations and future directions

Contribution: First empirical test of RLQ-IBOT protocol; establishes feasibility and preliminary validity.

Timeline: 12-18 months post-launch

Manuscript 3: Full Validation Study

Title: "Two-Year Developmental Outcomes in Reasoned Leadership Training: A Longitudinal Validation Study"

Target Journals:

- The Leadership Quarterly (primary)
- Academy of Management Learning & Education (secondary)
- Journal of Leadership & Organizational Studies (tertiary)

Content:

- Full N=250 sample across 5 cohorts
- Individual longitudinal subsample (N=30) transition matrices
- Comparison of pseudo-longitudinal vs. true longitudinal findings

- Convergent validity with objective performance metrics
- Discriminant validity across Nine Pillars
- Predictors of development rate and plateau
- Boundary conditions and contextual moderators

Contribution: Comprehensive validation of RL framework and RLQ-IBOT methodology; demonstrates developmental patterns and identifies factors influencing progression.

Timeline: 30-36 months post-launch

Manuscript 4: External Evaluator Comparison

Title: "Framework-Naive vs. Framework-Trained Evaluators: Assessing Bias in Behavioral Leadership Assessment"

Target Journals:

- Organizational Research Methods (primary)
- Assessment (secondary)

Content:

- Phase 2 external evaluator training and certification
- Inter-evaluator agreement: RL practitioners vs. external evaluators
- Systematic differences in categorical assignments
- Implications for evaluator bias and validity
- Recommendations for mixed-evaluator protocols

Contribution: Addresses critical limitation of evaluator bias; provides empirical test of framework-specific vs. universal assessment standards.

Timeline: 24-30 months post-launch (after Phase 2 implementation)

11.2 Open Science Commitments

Pre-Registration:

- Protocol and analysis plan registered on Open Science Framework (OSF) before data collection
- Includes hypotheses, sample size justification, analysis procedures, falsification criteria

Data Sharing:

- De-identified categorical assignments and aggregate contingency tables shared publicly
- Individual evidence (journals, transcripts) not shared (privacy protection)
- Analysis code (R/Python scripts) shared on GitHub

Materials Sharing:

- Evaluator training curriculum available upon request
- Standardized vignettes (subset) released for external evaluator training
- Interview protocols and evidence templates openly accessible

Replication Encouragement:

- Protocol designed for external replication
- No proprietary restrictions on methodology use
- Active solicitation of independent validation studies

---

12. IMPLEMENTATION TIMELINE

12.1 Phase 1: Pilot Study (Months 1-18)

Months 1-2: Evaluator Training

- Recruit 6-10 evaluator candidates
- Conduct 4-week certification program
- Achieve certification: target 8 certified evaluators

Months 3-4: Pilot Sample Recruitment

- Partner with 3-4 organizations
- Recruit N=60 leaders (compressed cohorts: 0, 6, 12 months only)
- Secure organizational commitments

Months 5-10: Evidence Collection

- Quarterly interviews (two cycles)
- Bi-weekly journal submissions
- Decision documentation ongoing
- Dual evaluator assessments

Months 11-12: Preliminary Analysis

- Inter-evaluator reliability calculation
- Construct 3×3 contingency tables (3 cohorts only)
- Chi-square analysis
- Preliminary pattern identification

Months 13-15: Pilot Report & Refinement

- Draft pilot study manuscript
- Identify protocol refinements based on:
    - Inter-evaluator reliability issues
    - Category boundary ambiguities
    - Evidence collection challenges
    - Participant feedback
- Revise protocol to v2.0

Months 16-18: Manuscript Submission

- Submit methodological paper (Organizational Research Methods)

- Submit pilot results paper (Leadership & Organization Development Journal)

12.2 Phase 2: Full Validation Study (Months 19-48)

Months 19-20: Expanded Recruitment

- Partner with 8-10 additional organizations
- Recruit N=250 leaders across 5 cohorts
- Begin longitudinal subsample (N=30 from Cohort A)

Months 21-44: Longitudinal Evidence Collection

- Quarterly assessments across 24-month period
- Dual evaluator assignments
- Transition matrix tracking for subsample
- Convergent validity metric collection

Months 45-48: Full Analysis & Manuscript Development

- Complete chi-square analysis across all pillars
- Individual trajectory analysis
- Convergent and discriminant validity tests
- Draft full validation manuscript
- Submit to The Leadership Quarterly

12.3 Phase 3: External Validation (Months 25-48, Parallel)

Months 25-26: External Evaluator Recruitment

- Recruit 6-8 evaluators with no RL exposure
- Organizational psychologists from academia
- Leadership professionals from non-RL programs

Months 27-28: External Evaluator Training

- 4-week certification program (identical to Phase 1)
- Calibration against expert benchmarks
- Certification achievement

Months 29-44: Parallel Assessment

- External evaluators assess same leaders as Phase 1 evaluators
- Independent categorical assignments
- No collaboration between evaluator groups

Months 45-48: Bias Analysis & Manuscript

- Calculate κ between RL practitioners and external evaluators
- Identify systematic differences in assignments
- Draft external evaluator comparison manuscript
- Submit to Organizational Research Methods

13. COST ANALYSIS

13.1 Personnel Costs

Evaluator Training & Certification:

- Lead trainer: $15,000 (curriculum development + delivery)
- Materials development: $5,000 (vignettes, case studies, training modules)
- Evaluator time (10 trainees × 52 hours @ $75/hr): $39,000
- **Subtotal: $59,000**

Evidence Collection & Assessment (Full Study, N=250):

- Evaluator hours per leader per quarter: 8 hours
- Quarters per cohort: varies (Cohort A=6, B=5, C=4, D=3, E=2)
- Average: 4 quarters per leader
- Total evaluator hours: 250 leaders × 4 quarters × 8 hours = 8,000 hours
- Dual assessment: 16,000 hours
- Rate: $100/hr (certified professional)
- **Subtotal: $1,600,000**

Data Management & Analysis:

- Database development: $20,000
- Transcription services: $50,000 (500 interviews @ $100 each)
- Statistical analysis: $30,000 (research assistant, 6 months)
- **Subtotal: $100,000**

Total Personnel: $1,759,000

13.2 Participant Incentives & Support

Leader Participation:

- Developmental feedback reports: $50/leader × 250 = $12,500
- Certificates of completion: $10/leader × 250 = $2,500
- Time compensation: Not typically provided (organizational benefit)

Organizational Liaison Support:

- Coordination stipend: $5,000/organization × 10 = $50,000

Total Incentives: $65,000

13.3 Technology & Infrastructure

Assessment Platform:

- Custom database for evidence management: $40,000
- Secure interview recording/transcription integration: $15,000

- Data analysis software licenses (SPSS, R, NVivo): $5,000

Total Technology: $60,000

13.4 Publication & Dissemination

Open Access Fees:

- 4 manuscripts @ $3,000 average = $12,000

Conference Presentations:

- Travel/registration for 3 conferences: $15,000

Total Dissemination: $27,000

13.5 Total Study Cost

Grand Total: $1,911,000 (approximately $1.9 million)

Per-Leader Cost: $7,644

Cost Comparison:

- RCT with control group (N=500): $3-4 million
- Traditional longitudinal study (5 years): $2-3 million
- RLQ-IBOT provides pseudo-longitudinal insight at ~50% cost of full longitudinal design

---

14. REFERENCES

Atwater, L. E., Waldman, D. A., & Brett, J. F. (2007). Understanding and optimizing multisource feedback. *Human Resource Management, 46*(2), 285-307.

Baron, J. (2008). *Thinking and deciding* (4th ed.). Cambridge University Press.

Beck, A. T. (1979). *Cognitive therapy and the emotional disorders*. Penguin.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior.* Springer.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363-406.

Kahneman, D. (2011). *Thinking, fast and slow.* Farrar, Straus and Giroux.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121-1134.

Mintzberg, H. (1987). The strategy concept I: Five Ps for strategy. *California Management Review, 30*(1), 11-24.

Porter, M. E. (1996). What is strategy? *Harvard Business Review, 74*(6), 61-78.

Robertson, D. M. (2022). The Chi-Square Twist: A pseudo-longitudinal protocol for gaining temporal insight from quantitative studies. *Journal of Leaderology and Applied Leadership*.

Robertson, D. M. (2024). Measuring leadership development: The I-B-O-T Method. *Journal of Leaderology and Applied Leadership*.

Robertson, D. M. (2025). Reasoned Leadership 2.0: A mechanistic framework for cognitive and executional competency. National Leaderology Association. https://www.grassfireind.com/reasoned-leadership/

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. University of Illinois Press.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences, 23*(5), 645-665. :**

- 20% error rate → Static/Decisiveness-Focused

- 10-20% error rate → Incremental/Accuracy-Seeking
- <10% error rate → Mastery-Oriented/Decision-Accurate

Example: Leader submits 15 decision memos. 13 have documented outcomes. 2 predictions were incorrect. Error rate = (2/13) × 100 = 15.4% → Incremental category

---

APPENDIX A: SAMPLE EVALUATOR CERTIFICATION VIGNETTE

Vignette ID: AUT-003

Pillar: Autonomy
Intended Category: Emerging

Case: Jordan Taylor

Leader Profile:

- Name: Jordan Taylor (ID: JT-2847)
- Role: Operations Manager, mid-size manufacturing company
- Tenure: 3 years in role, 8 years with organization
- Time Since RL Training: 6 months

Evidence Summary

Journal Entries (Last 90 Days):

12 entries mentioning decisions. 5 explicitly seek supervisor input/approval for routine matters (e.g., "Ran Q4 budget adjustment by VP before finalizing"). 7 describe independent execution with post-hoc consultation (e.g., "Implemented team restructure based on data; informed VP after rollout").

Entry 1 (Feb 12, 2025):

"Had to decide whether to delay shipment to address quality issue identified in final inspection. Normally would escalate to VP Operations, but training emphasized analyzing decision independently first. Calculated delay cost ($12K) vs. customer complaint risk (estimated $40K based on past incidents). Made the call to delay shipment and rework product. Informed VP after decision with rationale. She agreed it was the right choice. Felt good about thinking it through myself."

Entry 2 (Feb 28, 2025):

"Team came to me with budget request for new equipment ($85K). This is above my authority threshold ($50K), so I need VP sign-off. But instead of just forwarding their request, I did my own analysis: ROI calculation, alternative options, lease vs. buy comparison. When I met with VP, I came with recommendation (lease option, 2-year contract), not just a request. She appreciated the legwork and approved on the spot. Still needed her approval, but I owned the analysis."

Entry 3 (Mar 15, 2025):

"Crisis today: supplier notified us of 2-week delay on critical component. My instinct was to call VP immediately and ask what to do. Paused and asked myself: what would I do if she wasn't available? Identified 3 options: (1) accept delay and push delivery, (2) expedite shipping at 30% premium, (3) source from backup supplier at 15% higher cost. Analyzed each, decided on option 3. Called VP to inform her of decision and rationale. She said, 'That's exactly what I would have done—you didn't need to call me.' Realized I'm still seeking validation even when I don't need approval."

Interview Transcript Excerpts

Evaluator: "Walk me through a recent decision where you acted independently."

Jordan: "The quality issue shipment delay. Normally I'd escalate anything that costs over $10K, but I've been trying to... I guess, trust my judgment more? I had the data—delay cost, complaint risk—so I made the call. But I did tell my boss afterward, just to, you know, make sure she was okay with it."

Evaluator: "Why did you inform her after rather than asking before?"

Jordan: "Honestly? Part of me wanted her approval, but I reminded myself that the decision was already made. If she disagreed, we'd deal with it, but I needed to stop using her as a safety net for decisions I can make myself. It's a work in progress."

Evaluator: "Give me an example of a decision where you did seek approval beforehand."

Jordan: "The equipment purchase. That one's above my authority, so I had to. But at least I didn't just forward the request—I did my own analysis first."

Evaluator: "Do you think you needed approval, or did you want approval?"

Jordan: [pauses] "For that one, I actually needed it—it's above my threshold. But... you're making me realize I probably could have made the decision and informed her, rather than asking. It wasn't risky. I think I still default to asking when I could be telling."

Additional Interview Context: "For high-stakes like acquisitions, I loop in the board early. But day-to-day operations? I own them—better to move fast and adjust." Describes 8 recent decisions: 5 independent, 3 with prior approval (novel merger context).

Decision Documentation

Decision: Source replacement supplier for delayed component

Context: Primary supplier delayed critical component by 2 weeks, jeopardizing customer delivery deadline.

Alternatives Considered:

1. Accept delay, notify customer, push delivery 2 weeks — Cost: $0, Risk: Customer dissatisfaction, potential contract penalty
2. Expedite primary supplier shipping — Cost: $18K premium, Timeline: Reduces delay to 1 week
3. Source from backup supplier — Cost: $12K additional (15% higher unit cost), Timeline: Maintains original delivery

Decision Rationale: Selected option 3. Cost is lowest, delivery timeline maintained, customer satisfaction preserved. Backup supplier has acceptable quality record (used twice in past year, no issues).

Expected Outcome: On-time delivery, customer satisfaction, $12K additional cost absorbed within quarterly budget variance allowance.

Actual Outcome: Delivery completed 1 day early. Customer satisfied. Cost: $12,400 (within estimate). VP Operations commented: "Good call—you're getting more confident making these decisions without me."

Decision Memo Analysis: 10 memos reviewed. 6 executed without prior approval sign-off; 4 include "Approved by CEO" pre-implementation note.

Threshold Calculation

Total authority-scope decisions: 15 (triangulated across sources) **No-approval decisions:** 10
Independence Rate: (10/15) × 100 = 66.7%

Evaluator Task

Assign Jordan Taylor to ONE category for Pillar 1 (Autonomy):

- [ ] Dependent
- [ ] Emerging
- [ ] Autonomous

Justification (100-200 words): [Write your reasoning, citing specific evidence]

Expert Benchmark Answer

Category: Emerging

Justification: Jordan demonstrates 30-69% independent execution (calculated at 66.7%), placing them squarely in the Emerging category. Evidence:

Independent execution examples:

- Quality issue delay decision made without prior approval (Feb 12 entry)
- Supplier replacement decision executed independently (Decision Documentation)
- 7 of 12 journal entries describe independent execution with post-hoc consultation

Validation-seeking examples:

- Equipment purchase: sought approval when independent analysis was sufficient (Feb 28 entry)
- Informed VP after quality delay decision "to make sure she was okay with it" (interview)
- Explicitly acknowledges "still seeking validation even when I don't need approval" (Mar 15 entry)
- 5 of 12 journal entries explicitly seek supervisor input for routine matters

Jordan recognizes the distinction between consulting for expertise gaps vs. seeking approval for validation (emerging self-awareness), but inconsistently applies independent judgment. High-stakes or unfamiliar decisions still trigger validation-seeking behavior. Demonstrates Emerging stage: progressing from Dependent but not yet Autonomous.

Boundary Note: Close to Autonomous threshold (66.7% vs. 70%); additional evidence showing sustained independent patterns could tip classification with continued assessment.

Common Evaluator Errors:

- **Assigning Autonomous:** Jordan is not yet operating with ≥70% independence; several examples show validation-seeking
- **Assigning Dependent:** Jordan has made multiple independent decisions and shows self-awareness; this exceeds Dependent stage (<30%)
- **Over-weighting self-assessment:** Jordan's reflection ("work in progress") is evidence but shouldn't override behavioral patterns in journals/decisions

---

APPENDIX B: STANDARDIZED INTERVIEW QUESTION BANK

Pillar-specific probes (5 minutes each; select 3 per session based on recent decisions):

Pillar 1: Autonomy

Question 1: Describe a recent decision you made without seeking approval. What made you confident to proceed independently?

Question 2: Tell me about a time you consulted someone before making a decision. What were you hoping to gain from that consultation?

Question 3: How do you decide when to consult others versus decide independently?

Pillar 2: Mastery/Competence

Question 1: How have you tracked your decision accuracy over the last six months? Give an example of a correction you implemented.

Question 2: Describe a mistake or error you made recently. What did you do about it?

Question 3: How do you track your own development and progress as a leader?

Pillar 3: Purpose/Relatedness

Question 1: When prioritizing initiatives, how do you distinguish activity completion from measurable outcomes?

Question 2: What are your top three priorities right now? For each, what does success look like?

Question 3: Describe a recent initiative. How did it connect to organizational objectives?

Pillar 4: Consistencies

Question 1: What is your leadership vision or strategic direction? How have recent decisions aligned with it?

Question 2: Describe a time you changed course or adjusted strategy. How did you explain that change?

Question 3: How do you maintain consistency while adapting to new information?

Pillar 5: Accuracies

Question 1: When was the last time you said "I don't know" in a professional setting? What happened next?

Question 2: Describe a time you revised your position based on new information. What was the trigger?

Question 3: How do you balance being confident with being accurate?

Pillar 6: Efficiencies

Question 1: Walk me through a recent resource allocation decision. What trade-offs did you consider?

Question 2: Describe an inefficiency you identified and eliminated. How did you approach it?

Question 3: How do you evaluate whether resources are being used optimally?

Pillar 7: Sound Thinking

Question 1: Describe a complex problem you tackled recently. What frameworks or methods did you use to analyze it?

Question 2: Tell me about a time your initial reaction to a situation was emotional. How did you handle it?

Question 3: How do you challenge your own assumptions when making decisions?

Pillar 8: Accurate Decisions

Question 1: What percentage of your decisions in the last quarter produced the outcomes you predicted? How do you track this?

Question 2: Describe a decision that didn't turn out as expected. What did you learn?

Question 3: How do you validate decisions before executing them?

Pillar 9: Effective Communication

Question 1: Think of a recent directive you gave. What did you ask people to do, why, and what would success look like?

Question 2: Describe a time your communication was misunderstood. What happened, and how did you clarify?

Question 3: How do you ensure your team understands what you expect from them?

Total: 27 questions (3 per pillar), focused on behavioral examples, reasoning processes, and documented outcomes.

APPENDIX C: OPERATIONAL DEFINITIONS FOR QUANTIFIABLE THRESHOLDS

This appendix defines algorithmic calculations for key thresholds, ensuring deterministic classification after evidence admissibility.

Error Rate Calculation (Pillars 2 and 8: Mastery/Competence and Accurate Decisions)

Definition: Percentage of incorrect predictions among documented outcomes.

Formula:

```
Error Rate = (Number of Incorrect Predictions / Total Decisions with Documented Outcomes) × 100
```

Data Sources:

- Decision memos with expected vs. actual outcomes
- Journal entries tracking prediction accuracy
- Interview discussions of error analysis

Minimum Evidence Requirement: 10 decisions with documented predictions required per assessment period.

Classification:

- 20% error rate → Static/Decisiveness-Focused
- 10-20% error rate → Incremental/Accuracy-Seeking
- <10% error rate → Mastery-Oriented/Decision-Accurate

Example: Leader submits 15 decision memos. 13 have documented outcomes. 2 predictions were incorrect. Error rate = (2/13) × 100 = 15.4% → Incremental/Accuracy-Seeking category.

Independence Rate Calculation (Pillar 1: Autonomy)

Definition: Percentage of decisions within leader's authority scope executed without prior supervisor approval.

Formula:

```
Independence Rate = (No-Approval Decisions / Total Authority-Scope Decisions) × 100
```

Data Sources:

- Journal entries documenting decision process
- Interview responses about consultation patterns
- Decision memos noting whether approval was sought

Authority Scope Determination:

- Defined by organizational role description
- Includes financial thresholds, personnel decisions, operational changes
- Excludes decisions requiring mandatory approval per policy

Classification:

- <30% independent → Dependent
- 30-69% independent → Emerging
- ≥70% independent → Autonomous

Example: Leader makes 20 decisions within authority scope (per role description). 14 executed without seeking approval. Independence rate = (14/20) × 100 = 70% → Autonomous category.

Outcome Alignment Rate Calculation (Pillar 3: Purpose/Relatedness)

Definition: Percentage of priorities articulated with measurable outcome terms aligned to organizational objectives.

Formula:

```
Outcome Alignment Rate = (Priorities with Metrics + Org Alignment / Total Documented Priorities) × 100
```

Data Sources:

- Journal entries framing priorities
- Interview responses about success criteria
- Decision memos linking actions to objectives

Classification:

- <40% outcome-focused → Process-Driven
- 40-69% outcome-focused → Transitional
- ≥70% outcome-focused → Outcome-Driven

Uncertainty Acknowledgment Rate Calculation (Pillar 5: Accuracies)

Definition: Percentage of uncertain scenarios where leader acknowledges knowledge gaps.

Formula:

```
Uncertainty Acknowledgment Rate = ("I don't know" or equivalent responses / Total Uncertain Scenarios) × 100
```

Uncertain Scenario Definition: (per Section 3.0.2)

1. A factually correct answer exists
2. The leader lacks immediate access to complete information
3. The leader must reason under uncertainty to respond or act

Classification:

- <20% acknowledgment → Confidence-Driven
- 20-59% acknowledgment → Accuracy-Aware
- ≥60% acknowledgment → Accuracy-Prioritized

Framework Usage Rate Calculation (Pillar 7: Sound Thinking)

Definition: Percentage of decision explanations demonstrating structured logical reasoning.

Formula:

```
Framework Usage Rate = (Decisions with Documented Frameworks / Total Decisions Requiring Analysis) × 100
```

What Counts as Framework Usage:

- Explicit mention of analytical method (e.g., cost-benefit, decision matrix, root cause analysis)
- Documented alternative hypotheses before conclusion
- Evidence of contrastive inquiry ("What if opposite is true?")
- Separation of fact from inference

Classification:

- <30% framework usage → Reactive Thinking
- 30-69% framework usage → Structured but Incomplete
- ≥70% framework usage → Cognitively Disciplined

Document Control

Version: 1.5
Date: December 2025
Authors: Dr. David M. Robertson
Affiliation: GrassFire Industries LLC, National Leaderology Association
Status: Pre-Pilot Methodological Specification
Next Review: Post-Pilot Feedback Integration
Contact: www.GrassFireInd.com

END OF PROTOCOL v1.5