S1 File: Comprehensive Methodology Documentation for SPUR Framework

Table of Contents

---

1. Detailed Scoring Protocols

1.1 Dimension 1: Methodological Innovation (20% Weight)

Detailed Scoring Rubric:

90-100 Points: Novel Method Development

- Introduces entirely new analytical techniques, measurement approaches, or research methods
- Enables investigations previously impossible with existing methods
- Demonstrates clear methodological advancement over current practices
- Shows evidence of rigorous development and testing
- Examples: New statistical models, novel data collection techniques, innovative experimental designs

Assessment Criteria:

- Novelty verification through comprehensive literature review (minimum 200 papers)
- Technical feasibility assessment
- Methodological rigor evaluation
- Innovation impact potential analysis

70-89 Points: Significant Method Modification

- Substantially improves existing methods in accuracy, efficiency, or applicability
- Demonstrates clear advantages over standard approaches
- Shows systematic validation of improvements
- Enables new applications of established methods
- Examples: Enhanced algorithms, improved measurement protocols, modified experimental procedures

Assessment Criteria:

- Comparison with baseline methods
- Improvement quantification
- Validation completeness
- Practical applicability assessment

50-69 Points: Creative Method Combination

- Innovatively integrates established methods from different fields
- Applies methods to novel domains or research questions
- Demonstrates synergistic benefits from combination
- Shows methodological creativity within established frameworks
- Examples: Cross-disciplinary method integration, novel application contexts

Assessment Criteria:

- Integration quality evaluation
- Cross-field synthesis assessment
- Novel application validation
- Methodological coherence analysis

30-49 Points: Standard Method Application

- Competent use of established methods without significant innovation
- Follows standard protocols and procedures
- Demonstrates methodological competence
- May include minor modifications or adaptations
- Examples: Standard surveys, conventional experiments, routine analyses

Assessment Criteria:

- Protocol adherence verification
- Methodological competence assessment
- Minor innovation evaluation
- Standard practice comparison

0-29 Points: Routine Method Use

- Uses established methods in conventional ways
- No methodological innovation or creativity
- May have methodological limitations or errors
- Limited methodological contribution
- Examples: Basic descriptive studies, standard replications

Assessment Criteria:

- Methodological adequacy verification
- Innovation absence confirmation
- Error identification
- Contribution limitation assessment

1.2 Dimension 2: Conceptual Originality (18% Weight)

Detailed Scoring Rubric:

90-100 Points: Paradigm-Shifting Concepts

- Introduces fundamentally new ways of understanding phenomena
- Challenges existing theoretical paradigms
- Provides novel conceptual frameworks with broad implications
- Demonstrates potential for field transformation
- Examples: New theoretical models, paradigm-changing hypotheses, revolutionary concepts

Assessment Criteria:

- Paradigm shift potential evaluation
- Conceptual novelty verification
- Theoretical coherence assessment
- Field impact potential analysis

70-89 Points: Novel Theoretical Frameworks

- Develops new theories or substantially extends existing frameworks
- Provides original theoretical contributions
- Demonstrates conceptual innovation within established paradigms
- Shows theoretical advancement potential
- Examples: New theoretical models, extended frameworks, novel hypotheses

Assessment Criteria:

- Theoretical novelty verification
- Framework development quality
- Conceptual coherence evaluation
- Innovation significance assessment

50-69 Points: Creative Conceptual Connections

- Links previously unconnected concepts or domains
- Generates innovative hypotheses or propositions
- Demonstrates conceptual creativity
- Provides novel insights within established frameworks
- Examples: Cross-domain connections, creative hypotheses, novel applications

Assessment Criteria:

- Connection novelty verification
- Conceptual creativity evaluation
- Insight significance assessment
- Innovation quality analysis

30-49 Points: Incremental Conceptual Advances

- Makes modest theoretical contributions
- Refines existing understanding
- Provides limited conceptual innovation
- Demonstrates competent theoretical application
- Examples: Theory refinements, modest extensions, incremental insights

Assessment Criteria:

- Contribution significance evaluation
- Innovation limitation assessment
- Theoretical competence verification
- Advancement measurement

0-29 Points: Standard Conceptual Application

- Applies established concepts without novel insights
- No significant theoretical contribution
- Limited conceptual innovation

- May contain conceptual errors or limitations
- Examples: Standard applications, routine theoretical use

Assessment Criteria:

- Conceptual adequacy verification
- Innovation absence confirmation
- Error identification
- Contribution limitation assessment

1.3 Dimension 3: Empirical Scope & Scale (15% Weight)

Normalization Procedures by Discipline:

Natural Sciences Baselines:

- Sample Size: Median values from recent publications in specific subfields
- Temporal Coverage: Standard observation periods by research type
- Geographic Range: Typical study locations and coverage areas
- Data Comprehensiveness: Standard measurement variables and protocols

Social Sciences Baselines:

- Sample Size: Survey and experimental standards by methodology type
- Temporal Coverage: Longitudinal study norms by research area
- Geographic Coverage: Cross-cultural and cross-national study standards
- Data Sources: Multiple source integration expectations

Applied Sciences Baselines:

- Sample Size: Clinical trial and field study standards
- Temporal Coverage: Follow-up period norms by application area
- Practical Coverage: Real-world application scope standards
- Implementation Scale: Pilot to full-scale deployment ranges

Scoring Adjustments:

- Resource constraint multipliers for underfunded research contexts
- Quality weighting factors to prevent quantity-over-quality optimization
- Multi-domain integration bonuses for cross-contextual studies
- Methodological rigor adjustments for scope-quality tradeoffs

1.4 Dimension 4: Societal Impact Potential (15% Weight)

Impact Assessment Framework:

UN Sustainable Development Goals Alignment Analysis:

- Direct contribution assessment to specific SDGs
- Indirect benefit evaluation across multiple goals
- Implementation pathway identification
- Timeline and feasibility analysis

Policy Implementation Feasibility Evaluation:

- Stakeholder analysis and engagement potential
- Regulatory framework compatibility
- Implementation barrier identification
- Cost-benefit analysis considerations

Stakeholder Benefit Assessment:

- Primary beneficiary identification
- Benefit magnitude estimation
- Distribution equity analysis
- Accessibility and inclusion considerations

Time-to-Implementation Analysis:

- Research-to-application timeline estimation
- Intermediate milestone identification
- Scaling pathway analysis
- Adoption barrier assessment

1.5 Dimension 5: Cross-Disciplinary Integration (12% Weight)

Disciplinary Distance Measurement:

- Academic classification system analysis
- Methodological tradition comparison
- Theoretical framework distance assessment
- Publication pattern analysis

Integration Quality Indicators:

- Synthesis depth evaluation (surface vs. deep integration)
- Novel insight generation from cross-field combination
- Methodological integration assessment
- Theoretical coherence across disciplines

Integration Types:

- Methodological Integration: Cross-field method application
- Theoretical Integration: Cross-domain concept synthesis
- Empirical Integration: Multi-field data combination
- Applied Integration: Cross-sector solution development

---

2. Baseline Sample Generation Procedures

2.1 Systematic Literature Search Protocol

Database Selection:

- Web of Science (primary)
- Scopus (secondary)
- PubMed (life sciences)
- IEEE Xplore (engineering/computer science)
- JSTOR (social sciences/humanities)
- arXiv (preprints)

Search Strategy:

- Discipline-specific keyword sets
- Journal classification filtering
- Publication date range specification
- Language filtering (English primary, multilingual secondary)
- Quality filtering by impact factor percentiles

Sample Size Requirements:

- Minimum 200 papers per discipline-methodology combination
- Power analysis for statistical significance (minimum 80% power)
- Stratified sampling across subdisciplines
- Temporal stratification (60% recent, 40% historical)

2.2 Quality Control Procedures

Inclusion Criteria:

- Peer-reviewed publications only
- Minimum methodological standards verification
- Complete methodology documentation requirements
- Reproducibility criteria assessment

Exclusion Criteria:

- Predatory journal publications
- Incomplete methodological documentation
- Obvious methodological errors
- Duplicate publications or self-plagiarism

Quality Assessment:

- Impact factor verification
- Citation pattern analysis
- Peer review process validation
- Methodological rigor evaluation

2.3 Baseline Calculation Methods

Statistical Procedures:

- Distribution normality testing (Kolmogorov-Smirnov)
- Outlier detection and handling (modified Z-score method)
- Percentile calculation with confidence intervals
- Bootstrap resampling for robust estimates

Temporal Adjustments:

- Historical contextualization for changing standards
- Field evolution accounting
- Technology advancement adjustments
- Publication practice normalization

---

3. Inter-Rater Reliability Protocols

3.1 Expert Selection Criteria

Minimum Qualifications:

- PhD in relevant discipline
- Minimum 10 years research experience
- Demonstrated publication record in field
- Experience in research evaluation or peer review

Expertise Verification:

- Publication history analysis
- Citation impact assessment
- Professional recognition verification
- Evaluation experience confirmation

Panel Composition:

- Minimum 3 experts per evaluation
- Disciplinary expertise balance
- Geographic diversity requirements
- Institution type diversity (academic, government, industry)

3.2 Training and Calibration Procedures

Initial Training Phase:

- SPUR framework comprehensive overview
- Dimension-specific scoring guidelines
- Practice evaluations with known benchmarks
- Calibration exercises with expert consensus

Ongoing Calibration:

- Regular benchmark paper evaluations
- Inter-rater agreement monitoring
- Consensus-building discussions
- Scoring guideline refinements

3.3 Reliability Assessment Methods

Statistical Measures:

- Intraclass Correlation Coefficient (ICC)
- Pearson correlation coefficients
- Cronbach's alpha for internal consistency
- Cohen's kappa for categorical agreements

Quality Thresholds:

- Minimum ICC > 0.80 for overall scores
- Minimum r > 0.75 for dimensional correlations
- Maximum disagreement thresholds by dimension
- Consensus requirements for problematic cases

Composite Scoring Algorithm (Finalized)

This section provides the complete mathematical specification of SPUR score calculation, including conditional weighting logic, impact multiplier, and field normalization procedures.

Step 1: Determine Conditional Weights

SPUR implements dynamic weighting to reward breakthrough methodological innovation while acknowledging that frontier methods are inherently harder to replicate.

Conditional Logic:
```
IF Methodological_Innovation_Score ≥ 80:
   w_method = 0.25  (25%)
   w_replicability = 0.05  (5%)
ELSE:
   w_method = 0.20  (20%)
   w_replicability = 0.10  (10%)
```

Rationale:
- Highly innovative methods (score ≥ 80) often require:
  - Specialized equipment not yet widely available
  - Novel techniques requiring expert training
  - Unique observational opportunities
  - Computational resources beyond standard infrastructure

- The framework acknowledges these legitimate barriers by:
  - Increasing method innovation weight from 20% to 25%
  - Reducing replicability weight from 10% to 5%
  - Maintaining emphasis on transparency (what CAN be shared)

Threshold Justification:
- Score of 80/100 represents top ~15% of methodological innovations
- Empirically validated through expert panel calibration
- Aligns with "truly novel method" vs. "incremental improvement" distinction

Step 2: Calculate Base Score

The base score represents the weighted average of all seven dimensions, incorporating conditional weights from Step 1.

Formula:
```
Base_Score = (M × w_M) + (C × 0.18) + (E × 0.15) + (S × 0.15) +
             (X × 0.12) + (R × w_R) + (T × 0.10)
```

Where:
- M = Methodological Innovation score (0-100)
- C = Conceptual Originality score (0-100)
- E = Empirical Scope & Scale score (0-100)
- S = Societal Impact Potential score (0-100)
- X = Cross-Disciplinary Integration score (0-100)
- R = Replicability & Transparency score (0-100)
- T = Theoretical Advancement score (0-100)
- w_M = Method weight (0.20 or 0.25, from Step 1)
- w_R = Replicability weight (0.10 or 0.05, from Step 1)

Weight Verification: All weights sum to 1.00 under both conditions
- Standard: 0.20 + 0.18 + 0.15 + 0.15 + 0.12 + 0.10 + 0.10 = 1.00 ✓
- Conditional: 0.25 + 0.18 + 0.15 + 0.15 + 0.12 + 0.05 + 0.10 = 1.00 ✓

Step 3: Calculate Impact Multiplier

The impact multiplier amplifies the base score for research with exceptional societal relevance.

Formula:
```
```

Impact_Multiplier = 1 + (0.3 × Societal_Impact_Score / 100)
```

**Multiplier Range**:
- Minimum: 1.00 (when Societal Impact = 0)
- Maximum: 1.30 (when Societal Impact = 100)
- Typical: 1.15-1.25 for moderate-high impact research

**Design Philosophy**:
- Maximum 30% bonus preserves primacy of core research quality
- Linear scaling ensures proportional impact recognition
- Prevents societal impact from overwhelming technical merit

**Step 4: Calculate Final SPUR Score**

**Formula**:
```
Final_SPUR = Base_Score × Impact_Multiplier
```

**Score Range**:
- Theoretical minimum: 0 (all dimensions score 0)
- Theoretical maximum: 130 (all dimensions score 100, including Societal Impact)
- Typical range: 40-90 for published research
- Exceptional papers: 85-100

**Note on Unbounded Scale**:
- SPUR does not cap the final score at 100
- Scores >100 are theoretically possible but empirically rare
- A score >100 would indicate perfect research (100 on all dimensions) with transformative societal impact
- This design choice allows the scale to distinguish between "excellent research" and "excellent research with massive impact"
- To date, no paper in validation datasets has exceeded 100

**Step 5: Field-Normalized Score (Optional)**

For cross-field comparisons, raw SPUR scores can be normalized using field-specific baselines.

**Formula**:
```
SPUR_z = (Final_SPUR − μ_field) / σ_field
```
```

Where:
- SPUR_z = Field-normalized z-score
- Final_SPUR = Raw SPUR score from Step 4
- μ_field = Mean SPUR score for the field (from baseline calibration)
- σ_field = Standard deviation for the field (from baseline calibration)

Field Baseline Values (Established from calibration samples):

Field	Baseline Mean (μ)	Baseline SD (σ)	Typical Range
Biomedical Sciences	47.5	11.2	36-59
Physical Sciences	45.0	10.8	34-56
Social Sciences	50.0	13.4	37-63
Mathematics	42.5	9.6	33-52
Humanities	41.5	11.8	30-53

Interpretation:
- SPUR_z = 0: Average for field
- SPUR_z = +1: One standard deviation above field average (≈84th percentile)
- SPUR_z = +2: Two standard deviations above field average (≈98th percentile)
- SPUR_z = +3: Three standard deviations above field average (≈99.9th percentile)

When to Use Field Normalization:
- Comparing papers across different disciplines
- Identifying "most unique" papers in a multi-field database
- Grant evaluation across diverse applicant pools
- NOT recommended for within-field ranking (use raw scores)

Worked Example: Complete Calculation

Hypothetical Paper: Novel neuroimaging method for studying consciousness

Step 1: Dimensional Scores
- Methodological Innovation: 85
- Conceptual Originality: 78
- Empirical Scope & Scale: 72
- Societal Impact Potential: 88
- Cross-Disciplinary Integration: 80
- Replicability & Transparency: 65
- Theoretical Advancement: 75

Step 2: Determine Weights

```
Methodological Innovation (85) ≥ 80? YES
Therefore:
  w_method = 0.25
  w_replicability = 0.05
```

Step 3: Calculate Base Score

```
Base = (85 × 0.25) + (78 × 0.18) + (72 × 0.15) + (88 × 0.15) +
       (80 × 0.12) + (65 × 0.05) + (75 × 0.10)

Base = 21.25 + 14.04 + 10.80 + 13.20 + 9.60 + 3.25 + 7.50
Base = 79.64
```

Step 4: Calculate Impact Multiplier

```
Multiplier = 1 + (0.3 × 88/100)
Multiplier = 1 + 0.264
Multiplier = 1.264
```

Step 5: Calculate Final SPUR

```
Final_SPUR = 79.64 × 1.264
Final_SPUR = 100.62
```

Step 6: Field Normalization (if comparing to other fields)

```
Field: Biomedical Sciences
μ_field = 47.5
σ_field = 11.2

SPUR_z = (100.62 − 47.5) / 11.2
SPUR_z = 53.12 / 11.2
SPUR_z = 4.74
```

Interpretation:

- **Raw SPUR**: 100.62 (Exceptional - top tier research)
- **Field-Normalized**: z = 4.74 (>99.99th percentile for biomedical sciences)
- **Classification**: Breakthrough research with transformative potential
- **Note**: This paper scores >100 due to perfect combination of high innovation (triggering conditional weighting) and exceptional societal impact (high multiplier)

Comparison Example: Standard vs. Conditional Weighting

To illustrate the conditional weighting impact, consider two papers with identical dimensional scores except for Methodological Innovation:

Paper A: Incremental Method
- Methodological Innovation: 70 (below threshold)
- All other dimensions: 75

```
Weights: w_method = 0.20, w_replicability = 0.10
Base = (70 × 0.20) + (75 × 0.18) + (75 × 0.15) + (75 × 0.15) +
      (75 × 0.12) + (75 × 0.10) + (75 × 0.10)
Base = 14.00 + 13.50 + 11.25 + 11.25 + 9.00 + 7.50 + 7.50 = 74.00
Multiplier = 1 + (0.3 × 75/100) = 1.225
Final = 74.00 × 1.225 = 90.65
```

Paper B: Breakthrough Method
- Methodological Innovation: 85 (above threshold)
- All other dimensions: 75

```
Weights: w_method = 0.25, w_replicability = 0.05
Base = (85 × 0.25) + (75 × 0.18) + (75 × 0.15) + (75 × 0.15) +
      (75 × 0.12) + (75 × 0.05) + (75 × 0.10)
Base = 21.25 + 13.50 + 11.25 + 11.25 + 9.00 + 3.75 + 7.50 = 77.50
Multiplier = 1 + (0.3 × 75/100) = 1.225
Final = 77.50 × 1.225 = 94.94
```

Impact of Conditional Weighting:
- Paper B receives additional +4.29 points for breakthrough methodology
- This represents appropriate recognition of higher innovation value
- The difference (94.94 vs. 90.65 = 4.29 points) reflects the premium placed on genuine

methodological advancement

Sensitivity to Dimensional Changes

Understanding how each dimension influences the final score helps interpret SPUR values:

Dimensional Influence (holding all other dimensions constant at 75):

Dimension	Weight	10-point increase → SPUR change
Methodological Innovation*	20-25%	+2.0 to +2.5 points
Conceptual Originality	18%	+1.8 points
Empirical Scope & Scale	15%	+1.5 points
Societal Impact**	15% + multiplier	+1.5 base + multiplier effect
Cross-Disciplinary Integration	12%	+1.2 points
Replicability & Transparency*	5-10%	+0.5 to +1.0 points
Theoretical Advancement	10%	+1.0 points

*Varies based on conditional weighting
**Has dual effect: base score + multiplier

Key Insights:
1. Methodological Innovation has largest direct effect
2. Societal Impact has unique amplifying effect via multiplier
3. No single dimension can drive SPUR >90 alone; excellence requires multi-dimensional strength
4. Conditional weighting creates appropriate incentive for genuine innovation

---

Common Scoring Scenarios

Scenario 1: Methodologically Conservative but High Impact
- Method: 60, Concept: 75, Empirical: 80, Impact: 95, Cross-Disc: 70, Replic: 85, Theory: 72
- Base: 75.15 (standard weights)
- Multiplier: 1.285
- Final: 96.57 (High uniqueness driven by societal impact)

Scenario 2: Methodologically Brilliant but Limited Impact
- Method: 92, Concept: 88, Empirical: 75, Impact: 45, Cross-Disc: 70, Replic: 70, Theory: 80
- Base: 80.45 (conditional weights trigger)

- Multiplier: 1.135
- Final: 91.31 (High uniqueness driven by technical excellence)

Scenario 3: Balanced Excellence
- All dimensions: 85
- Base: 85.00 (conditional weights trigger)
- Multiplier: 1.255
- Final: 106.68 (Exceptional - extremely rare)

Scenario 4: Solid but Unremarkable
- All dimensions: 55
- Base: 55.00 (standard weights)
- Multiplier: 1.165
- Final: 64.08 (Above average, publishable)

Implementation Notes

For Authors:
- Use the complete SPUR assessment prompt (in main manuscript)
- Ensure all dimensional scores are justified with specific examples
- Document conditional weighting trigger (Method ≥ 80) if applicable
- Calculate all steps manually to verify AI assessment accuracy

For Peer Reviewers:
- Verify conditional weighting was correctly applied
- Check arithmetic at each step
- Assess whether Methodological Innovation ≥ 80 is genuinely warranted
- Confirm Societal Impact score supports claimed multiplier effect

For Journals:
- Request authors show complete calculation (not just final score)
- Flag any score >100 for additional scrutiny (should be extremely rare)
- Consider field normalization for cross-disciplinary special issues
- Track SPUR scores over time to monitor grade inflation

Mathematical Verification: All formulas have been validated through:
- 250 benchmark calculations (50 landmark + 200 recent papers)
- Cross-validation with independent AI systems
- Expert review by statisticians and measurement specialists

- Sensitivity analyses confirming stable behavior across score ranges

Software Implementation: Reference R code available in S2 File

---

4. Gaming Resistance Testing Procedures

4.1 Artificial Enhancement Scenarios

Vocabulary Manipulation:

- Unique term injection without conceptual contribution
- Jargon complexity inflation
- Superficial language sophistication
- Technical term misuse or overuse

Methodological Gaming:

- Superficial method combination without integration
- Complexity inflation without added value
- Novel terminology for established procedures
- Artificial innovation claims

Interdisciplinary Gaming:

- Citation diversity without genuine integration
- Surface-level cross-field references
- Terminology borrowing without synthesis
- Artificial boundary spanning claims

4.2 Detection Mechanisms

Semantic Depth Analysis:

- Natural language processing for conceptual coherence
- Argument structure analysis
- Evidence-conclusion connection assessment
- Theoretical consistency evaluation

Historical Precedent Verification:

- Comprehensive literature search for genuine precedents

- Novelty claim verification against existing work
- Innovation assessment relative to field knowledge
- Temporal contextualization for fair comparison

Expert Validation Loops:

- Domain expert assessment for genuine contribution
- Gaming attempt identification protocols
- Artificial enhancement flagging procedures
- Quality assessment override mechanisms

4.3 Resistance Validation Testing

Controlled Gaming Attempts:

- Systematic manipulation of test papers
- Gaming strategy effectiveness measurement
- Detection rate calculation
- False positive assessment

Robustness Testing:

- Multiple gaming strategy combinations
- Sophisticated manipulation attempts
- Expert gaming resistance evaluation
- System reliability assessment

---

5. Discipline Classification System

5.1 Primary Discipline Categories

Natural Sciences:

- Physics and Astronomy
- Chemistry and Materials Science
- Biology and Life Sciences
- Earth and Environmental Sciences
- Mathematics and Statistics

Social Sciences:

- Psychology and Cognitive Sciences
- Sociology and Anthropology
- Political Science and International Relations
- Economics and Business
- Education and Communication

Applied Sciences:

- Medicine and Health Sciences
- Engineering and Technology
- Computer Science and Information Technology
- Agriculture and Food Sciences
- Architecture and Planning

Interdisciplinary Fields:

- Science and Technology Studies
- Environmental Studies
- Public Policy and Administration
- Biomedical Engineering
- Computational Social Science

5.2 Secondary Classification Protocols

Methodology Family Classification:

- Experimental Research
- Observational Studies
- Theoretical/Computational Work
- Mixed Methods Research
- Systematic Reviews/Meta-analyses

Temporal Period Classification:

- Historical Era (pre-1990)
- Modern Era (1990-2010)
- Contemporary Era (2010-present)
- Emerging Era (2020-present)

5.3 Classification Procedures

Automated Classification:

- Journal classification database matching

- Keyword analysis and mapping
- Citation network analysis
- Abstract content analysis

Manual Verification:

- Expert review of automated classifications
- Interdisciplinary work special handling
- Edge case resolution procedures
- Quality control verification

---

6. Statistical Distribution Analysis Methods

6.1 Distribution Fitting Procedures

Normality Testing:

- Kolmogorov-Smirnov test for distribution normality
- Shapiro-Wilk test for smaller samples
- Q-Q plot visual inspection
- Skewness and kurtosis analysis

Alternative Distributions:

- Log-normal transformation for right-skewed data
- Beta distribution for bounded data
- Gamma distribution for continuous positive data
- Non-parametric methods for irregular distributions

6.2 Percentile Calculation Methods

Bootstrap Resampling:

- 1000+ resampling iterations for robust estimates
- Bias-corrected and accelerated (BCa) intervals
- Confidence interval calculation for percentiles
- Stability assessment across resampling runs

Outlier Handling:

- Modified Z-score method for outlier identification

- Robust percentile calculation methods
- Influence analysis for extreme values
- Sensitivity analysis for outlier impact

6.3 Cross-Validation Procedures

Sample Splitting:

- Training/validation sample division (70/30)
- Cross-validation for parameter stability
- Out-of-sample prediction accuracy
- Generalization performance assessment

Temporal Validation:

- Historical sample predictions
- Forward validation on new publications
- Stability assessment across time periods
- Trend adjustment for evolving fields

Implementation Notes

This comprehensive methodology documentation provides detailed protocols for implementing the SPUR framework across different contexts and disciplines. All procedures are designed to maintain objectivity, resist gaming, and ensure reproducible results while accommodating the diverse nature of scientific research across fields.

Regular updates to these protocols will be necessary as fields evolve and new gaming strategies emerge. The framework's strength lies in its systematic approach to addressing known limitations while remaining adaptable to future challenges in research evaluation.