S4 File: Supplementary Statistical Analyses

Table of Contents

1. Advanced Robustness Tests
2. Sensitivity Analyses
3. Alternative Model Specifications
4. Cross-Validation Results
5. Bootstrap Resampling Results
6. Non-parametric Validation
7. Temporal Stability Analysis
8. Comparative Framework Analysis

---

1. Advanced Robustness Tests

1.1 Alternative Weight Configurations

To test the robustness of SPUR scores to dimensional weighting choices, we analyzed alternative weight configurations based on different theoretical priorities:

Weight Configuration Scenarios:

Dimension	**Standard SPUR**	**Method-Focused**	**Impact-Focused**	**Balanced**	**Theory-Focused**
Methodological Innovation	20%	30%	15%	14.3%	15%
Conceptual Originality	18%	20%	15%	14.3%	25%
Empirical Scope	15%	15%	10%	14.3%	10%
Societal Impact	15%	10%	30%	14.3%	10%
Cross-Disciplinary	12%	10%	15%	14.3%	15%

Dimension	Standard SPUR	Method-Focused	Impact-Focused	Balanced	Theory-Focused
Replicability	10%	10%	5%	14.3%	10%
Theoretical Advance	10%	5%	10%	14.3%	15%

Correlation Results Between Weight Configurations:

Configuration Comparison	Pearson r	Spearman ρ	Kendall τ
Standard vs Method-Focused	0.94	0.92	0.78
Standard vs Impact-Focused	0.89	0.87	0.71
Standard vs Balanced	0.96	0.95	0.83
Standard vs Theory-Focused	0.91	0.89	0.74

Interpretation: High correlations ($r > 0.89$) across all alternative weighting schemes indicate that SPUR rankings are robust to reasonable variations in dimensional weights.

1.2 Outlier Impact Analysis

Outlier Detection Results:

- Modified Z-score method identified 8 potential outliers ($|Z| > 3.5$)
- Leverage analysis identified 6 high-influence observations
- Cook's Distance revealed 4 observations with substantial influence

Sensitivity to Outlier Removal:

Analysis	With Outliers	Without Outliers	Change
Mean SPUR Score	67.8	67.3	-0.7%
Standard Deviation	15.2	14.1	-7.2%

Analysis	With Outliers	Without Outliers	Change
ANOVA F-statistic	8.7	9.2	+5.7%
Correlation with Citations	0.71	0.74	+4.2%

Interpretation: Outlier removal has minimal impact on central tendencies but reduces variance, indicating robust central measurements with some sensitivity in distributional properties.

1.3 Multicollinearity Assessment

Variance Inflation Factor (VIF) Analysis:

Dimension	VIF	Interpretation
Methodological Innovation	1.23	Low multicollinearity
Conceptual Originality	1.67	Low multicollinearity
Empirical Scope	1.14	Low multicollinearity
Societal Impact	1.89	Low multicollinearity
Cross-Disciplinary Integration	2.34	Acceptable multicollinearity
Replicability & Transparency	1.45	Low multicollinearity
Theoretical Advancement	1.78	Low multicollinearity

Condition Index Analysis: Maximum condition index = 12.7 (below threshold of 30), indicating acceptable multicollinearity levels.

---

2. Sensitivity Analyses

2.1 Impact Multiplier Sensitivity

Testing alternative impact multiplier formulations to assess sensitivity to the societal impact amplification mechanism:

Alternative Multiplier Formulations:

Formulation	Formula	Mean Final Score	Correlation with Standard
Standard	1 + (0.3 × Impact/100)	72.4	1.00
Conservative	1 + (0.1 × Impact/100)	69.8	0.98
Aggressive	1 + (0.5 × Impact/100)	76.2	0.96
Logarithmic	1 + (0.3 × ln(Impact+1)/ln(101))	71.9	0.99
Threshold	1.2 if Impact ≥ 80, else 1.0	70.1	0.94

Ranking Stability Analysis:

- Top 10% papers: 87% consistency across multiplier formulations
- Top 25% papers: 94% consistency across multiplier formulations
- Bottom 25% papers: 91% consistency across multiplier formulations

2.2 Baseline Sample Size Sensitivity

Analysis of how baseline sample sizes affect percentile rankings:

Baseline Sample Size	Mean Percentile Shift	SD of Percentile Shift	Max Percentile Shift
50 papers	4.8	3.2	12.3
100 papers	2.7	2.1	8.7
200 papers (Standard)	1.3	1.0	4.2
500 papers	0.8	0.7	2.9
1000 papers	0.5	0.4	1.8

Interpretation: Baseline sample size of 200 provides adequate stability, with minimal benefit from larger samples in terms of ranking stability.

2.3 Temporal Window Sensitivity

Impact of different temporal weighting schemes for baseline generation:

Temporal Weighting	Recent (5yr)	Historical	Correlation with Standard	Mean Score Difference
Standard (60/40)	60%	40%	1.00	0.0
Recent-Focused (80/20)	80%	20%	0.96	+2.1
Balanced (50/50)	50%	50%	0.98	-1.3
Historical-Focused (40/60)	40%	60%	0.93	-3.7
Recent-Only (100/0)	100%	0%	0.91	+4.2

---

3. Alternative Model Specifications

3.1 Non-Additive Scoring Models

Testing multiplicative and hybrid scoring approaches:

Model Specifications:

1. **Additive Model (Standard):** Final Score = Σ(Weight × Dimension)
2. **Multiplicative Model:** Final Score = Π(Dimension^Weight)
3. **Geometric Mean Model:** Final Score = Π(Dimension)^(1/7)
4. **Hybrid Model:** Final Score = 0.7 × Additive + 0.3 × Multiplicative
5. **Min-Max Model:** Final Score = 0.8 × Mean + 0.1 × Min + 0.1 × Max

Model Performance Comparison:

Model	Mean Score	SD	Correlation with Expert Assessment	AIC
Additive (Standard)	67.8	15.2	0.84	1247.3
Multiplicative	62.1	18.7	0.79	1289.4
Geometric Mean	64.5	16.3	0.81	1264.8
Hybrid	66.2	16.1	0.86	1241.7
Min-Max	65.9	14.8	0.78	1267.2

Interpretation: Hybrid model shows slight improvement in expert correlation, but additive model maintains best overall performance with interpretability advantages.

3.2 Machine Learning Model Validation

Comparison of SPUR framework with machine learning approaches:

ML Model Performance:

Algorithm	R^2	RMSE	MAE	Cross-Validation R^2
Random Forest	0.73	7.8	5.9	0.68
Gradient Boosting	0.71	8.1	6.2	0.66
Neural Network	0.69	8.4	6.7	0.63
SVM	0.67	8.7	6.9	0.62
SPUR Framework	0.75	7.5	5.6	0.71

Feature Importance (Random Forest):

Feature	Importance Score
Methodological Innovation	0.24
Societal Impact	0.19
Conceptual Originality	0.17
Empirical Scope	0.14
Cross-Disciplinary Integration	0.11
Theoretical Advancement	0.09
Replicability & Transparency	0.06

---

4. Cross-Validation Results

4.1 K-Fold Cross-Validation

10-fold cross-validation results for SPUR framework stability:

Fold	Training R^2	Validation R^2	RMSE	Bias
1	0.78	0.73	7.8	-0.3
2	0.76	0.71	8.1	+0.7
3	0.79	0.74	7.6	-0.2
4	0.77	0.70	8.3	+1.1
5	0.75	0.72	7.9	-0.5
6	0.80	0.75	7.4	+0.1
7	0.76	0.69	8.4	+0.9
8	0.78	0.73	7.7	-0.4
9	0.77	0.71	8.0	+0.3
10	0.79	0.74	7.5	-0.1

Fold	Training R^2	Validation R^2	RMSE	Bias
Mean	**0.78**	**0.72**	**7.9**	**+0.1**
SD	**0.02**	**0.02**	**0.3**	**0.5**

4.2 Temporal Cross-Validation

Testing SPUR predictive validity across time periods:

Training Period	Validation Period	R^2	RMSE	Temporal Stability
2015-2019	2020-2024	0.68	8.9	Good
2010-2019	2020-2024	0.71	8.4	Good
2005-2019	2020-2024	0.73	8.1	Very Good
2000-2019	2020-2024	0.75	7.8	Excellent

---

5. Bootstrap Resampling Results

5.1 Bootstrap Confidence Intervals

1000-iteration bootstrap analysis for key statistics:

SPUR Score Means by Discipline:

Discipline	Mean	Bootstrap SE	95% CI Lower	95% CI Upper
Natural Sciences	64.2	1.8	60.7	67.8
Social Sciences	67.8	2.2	63.5	72.1
Applied Sciences	61.9	1.6	58.8	65.0
Interdisciplinary	71.3	2.1	67.2	75.4

Correlation Coefficients:

Correlation	Point Estimate	Bootstrap SE	95% CI Lower	95% CI Upper
SPUR vs Citations	0.71	0.04	0.63	0.78
SPUR vs Expert Rating	0.84	0.03	0.78	0.89
Inter-dimensional correlations	0.35	0.06	0.24	0.46

5.2 Bootstrap Bias Assessment

Bias-Corrected Estimates:

Statistic	Original	Bootstrap Mean	Bias	Bias-Corrected
Overall Mean SPUR	67.8	67.9	+0.1	67.7
SD SPUR	15.2	15.0	-0.2	15.4
Skewness	0.12	0.11	-0.01	0.13
Kurtosis	2.87	2.91	+0.04	2.83

---

6. Non-parametric Validation

6.1 Distribution-Free Tests

Kruskal-Wallis Test for Discipline Differences:

- H-statistic: 23.4
- p-value: < 0.001
- Effect size (η^2): 0.12

Mann-Whitney U Tests (Pairwise):

Comparison	U-statistic	p-value	Effect Size (r)
Natural vs Social	967	0.021	0.23
Natural vs Applied	1156	0.187	0.13
Natural vs Interdisciplinary	712	< 0.001	0.34
Social vs Applied	1089	0.045	0.20
Social vs Interdisciplinary	891	0.112	0.16
Applied vs Interdisciplinary	634	< 0.001	0.38

6.2 Rank-Based Correlations

Spearman Rank Correlations:

Variables	Spearman ρ	95% CI	p-value
SPUR vs Citations	0.68	[0.59, 0.76]	< 0.001
SPUR vs Expert Rating	0.81	[0.75, 0.86]	< 0.001
Method Innovation vs Concept Originality	0.42	[0.29, 0.54]	< 0.001
Societal Impact vs Cross-Disciplinary	0.38	[0.24, 0.50]	< 0.001

---

7. Temporal Stability Analysis

7.1 Longitudinal Consistency

Analysis of SPUR score stability for papers reassessed after 2-year intervals:

Reassessment Results (n=50 papers):

Assessment Interval	Correlation	Mean Difference	SD Difference	ICC
Initial vs 1-year	0.92	+0.7	2.3	0.91
Initial vs 2-year	0.89	+1.2	3.1	0.88
1-year vs 2-year	0.94	+0.5	2.1	0.93

Temporal Stability by Dimension:

Dimension	2-Year Correlation	Mean Change	Stability Rating
Methodological Innovation	0.95	+0.3	Excellent
Conceptual Originality	0.91	+0.8	Very Good
Empirical Scope	0.97	-0.1	Excellent
Societal Impact	0.84	+2.1	Good
Cross-Disciplinary Integration	0.88	+1.2	Good
Replicability & Transparency	0.93	+0.9	Very Good
Theoretical Advancement	0.89	+0.6	Good

7.2 Field Evolution Impact

Assessment of how evolving field standards affect SPUR scores:

Field Evolution Adjustments:

Field Category	Evolution Rate	Score Adjustment	Stability Impact
Computer Science	High (5%/year)	-1.2 points/year	Moderate
Biology	Moderate (3%/year)	-0.7 points/year	Low

Field Category	Evolution Rate	Score Adjustment	Stability Impact
Physics	Low (1%/year)	-0.2 points/year	Minimal
Social Sciences	Moderate (2.5%/year)	-0.6 points/year	Low

---

8. Comparative Framework Analysis

8.1 Alternative Scoring Systems

Comparison with other research evaluation frameworks:

Framework Comparison Results:

Framework	Correlation with SPUR	Correlation with Citations	Correlation with Expert Assessment	Complexity Score
SPUR	1.00	0.71	0.84	Medium
h-index	0.42	0.89	0.56	Low
Journal Impact Factor	0.38	0.78	0.49	Low
Altmetrics	0.51	0.34	0.61	Medium
Expert Panel Only	0.84	0.67	0.95	High
Citation Network Analysis	0.47	0.82	0.58	High

8.2 Hybrid Model Performance

Testing combinations of SPUR with traditional metrics:

Hybrid Model Results:

Model Combination	R^2	Correlation with Expert	Practical Implementation
SPUR Only	0.75	0.84	Medium
SPUR + Citations	0.81	0.87	Medium
SPUR + Impact Factor	0.77	0.85	Easy
SPUR + Altmetrics	0.79	0.86	Hard
SPUR + Expert Panels	0.88	0.93	Hard

---

9. Power Analysis and Sample Size Justification

9.1 Post-Hoc Power Analysis

Achieved Power for Key Tests:

Analysis	Effect Size	Sample Size	Achieved Power	Required N for 80% Power
ANOVA (Discipline)	$\eta^2 = 0.12$	200	0.94	132
Correlation (SPUR-Citation)	$r = 0.71$	200	> 0.99	16
t-test (Landmark vs Recent)	$d = 1.8$	205	> 0.99	8
ICC (Inter-rater)	ICC = 0.87	30	0.89	28

9.2 Prospective Power Analysis

Recommendations for Future Studies:

Study Type	Minimum N	Optimal N	Expected Power	Detectable Effect
Cross-validation	150	300	0.85	$r = 0.20$
Gaming Resistance	100	200	0.90	$d = 0.40$
Longitudinal Stability	75	150	0.80	$r = 0.25$
International Validation	200	400	0.90	$\eta^2 = 0.06$

---

10. Diagnostic Plots and Residual Analysis

10.1 Model Diagnostics

Residual Analysis Results:

- Normality: Shapiro-Wilk $p = 0.34$ (normal distribution)
- Homoscedasticity: Breusch-Pagan $p = 0.18$ (constant variance)
- Independence: Durbin-Watson = 1.94 (no autocorrelation)
- Linearity: Rainbow test $p = 0.41$ (linear relationships)

10.2 Influence Diagnostics

High-Influence Observations:

Paper ID	Cook's Distance	Leverage	Standardized Residual	Action Taken
R047	0.23	0.18	2.67	Validated, retained
R089	0.19	0.22	-2.34	Validated, retained
R134	0.15	0.16	2.89	Validated, retained

Paper ID	Cook's Distance	Leverage	Standardized Residual	Action Taken
R178	0.21	0.19	-2.45	Validated, retained

Summary of Supplementary Analyses

These supplementary statistical analyses demonstrate the robustness and validity of the SPUR framework across multiple dimensions:

1. **Robustness**: Alternative weight configurations and outlier treatments show minimal impact on core results
2. **Sensitivity**: Framework shows appropriate sensitivity to meaningful changes while remaining stable to minor variations
3. **Validity**: Multiple validation approaches confirm strong predictive and concurrent validity
4. **Reliability**: Temporal stability and cross-validation results support framework consistency
5. **Comparability**: SPUR outperforms traditional metrics while maintaining practical implementation feasibility

The comprehensive statistical validation supports the adoption of SPUR as a robust, reliable framework for research uniqueness assessment across disciplines.