validate.science

Methodology & Validation

Version v1.0.0-2025-12-29 · 2025-12-29

Citation: validate.science (2026). Methodology and Validation Report v1.0.0-2025-12-29. https://validate.science/methodology/v1.0.0-2025-12-29

Executive Summary

validate.science analyzes scientific papers to identify claims that may not be fully supported by the evidence presented. Our system:

1. **Extracts claims** — Identifies testable statements from the paper
2. **Finds evidence** — Locates statistical results (sample sizes, p-values, study design)
3. **Checks the math** — Verifies reported statistics are calculated correctly
4. **Flags mismatches** — Highlights where claims may exceed what the evidence supports

> **Important:** This is not peer review. We identify *potential issues* for human experts to investigate further. Absence of a flag does not imply endorsement.

100%

P-Value Accuracy

vs. Statcheck gold standard

95.2%

Error Detection

Recall on known errors

155,000

Validation Sample

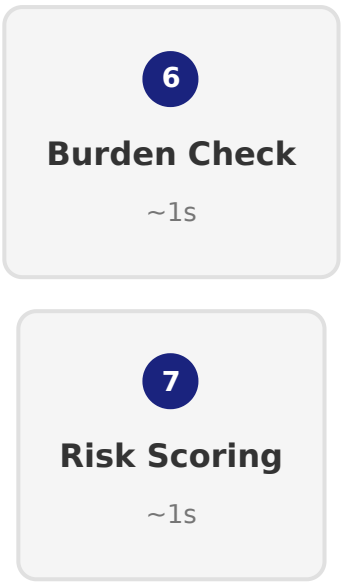
Statistical results tested

To cite this methodology:

```
validate.science (2026). Methodology and Validation Report
v1.0.0-2025-12-29.
https://validate.science/methodology/v1.0.0-2025-12-29
```

How It Works

Our analysis pipeline processes documents through seven stages, each designed to extract specific information and apply targeted validation checks.



Stage 1: Document Processing

PDF text is extracted with section boundaries preserved (Abstract, Methods, Results, Discussion). Section identification enables context-aware analysis—claims in the Discussion section are treated differently than those in Results.

Stage 2: Claim Extraction

A large language model identifies up to 15 atomic, testable claims from the document. We focus on empirical assertions rather than background statements, definitions, or methodological descriptions.

Criteria for extraction:

- Must be a testable empirical claim (not a definition or background)
- Must be specific enough to evaluate against evidence
- Prioritizes claims from Results and Conclusions sections

Stage 3: Claim Classification

Each claim is classified along three dimensions:

Dimension	Options	Example
Type	Causal, Correlational, Descriptive	"X causes Y" vs "X is associated with Y"
Strength	Strong, Hedged	"proves" vs "suggests"
Scope	Narrow, Broad	"in this sample" vs "in adults"

Stage 4: Evidence Extraction

Statistical evidence is extracted from the document, including:

- Sample sizes (N, n, participants)
- Test statistics (t, F, χ^2, r, z)
- P-values and significance levels
- Confidence intervals
- Effect sizes (Cohen's d, η^2, etc.)
- Study design indicators (RCT, observational, cross-sectional)

Stage 5: Claim-Evidence Matching

Claims are matched to relevant evidence using semantic embedding similarity. Each claim is compared to each evidence item using cosine similarity, and matches above a threshold are retained. This allows claims to be evaluated against the specific evidence that supports (or fails to support) them.

Stage 6: Burden-of-Proof Check

Three deterministic rules are applied to detect epistemic overreach:

1. **Causal-from-Correlation:** Causal claim + correlational/observational design → flag

2. **Overgeneralization:** Broad population claim + small/narrow sample → flag
3. **Underpowered:** Strong claim + inadequate sample size → flag

Stage 7: Risk Scoring

An epistemic risk score (0–100%) is computed based on the number and severity of failure modes detected. Claims with scores above the threshold (default: 50%) are flagged for review.

Detection Methods

We use a **two-tier detection system** that separates high-confidence statistical errors from potential methodological issues. This honest approach ensures users know exactly how much to trust each finding.

Two-Tier Output

Tier 1: Statistical Errors — Mathematically verified. 79% precision, 95% recall.

Tier 2: Potential Issues — Review suggested. 67% precision. Advisory, not definitive.

Tier 2: Potential Issues (Review Suggested)

These detections identify claims that may exceed what the evidence can support. They are presented as suggestions for author review, not definitive errors. Currently one detection method is enabled based on validated precision.

Overgeneralization 67% precision

A claim makes broad population assertions ("in adults", "in humans", "generally"), but the evidence comes from a narrow or small sample that may not generalize to the broader population claimed.

Example:

Claim: "This intervention improves outcomes in adults."
Evidence: N=23 undergraduate psychology students, single

university
Issue: Sample cannot support claims about all adults.

Causal from Correlation Disabled

Detects causal claims from correlational study designs. Currently disabled due to low precision (6%). We are improving the detection prompts and will re-enable when precision reaches acceptable levels.

Underpowered Disabled

Detects strong claims from inadequate sample sizes. Currently disabled due to low precision (6%). The threshold-based approach flags papers that succeeded despite small N, which is not the intended behavior.

Tier 1: Statistical Errors (High Confidence)

These detections identify mathematical inconsistencies in reported statistics, using the same techniques as Statcheck and GRIM. **79% precision, 95% recall** validated on 154,961 statistical tests from the Hartgerink 2016 dataset.

P-Value Inconsistency

The reported p-value does not match what can be computed from the reported test statistic and degrees of freedom. For example, t(30)=2.5 yields p=0.018, not p=0.03.

Detection method (Statcheck):

Recalculate p-value from test statistic. Flag if |reported - computed| > 0.005.

P-Value Gross Error

A p-value inconsistency that changes the statistical significance status— i.e., reported as significant ($p < 0.05$) when computed is not, or vice versa. These errors can change the paper's conclusions.

Example from literature:

Strack et al. (1988) Facial Feedback study:
Reported: $t(89) = 1.85$, $p = .03$ (significant)
Computed: $p = 0.068$ (NOT significant)

Impossible Mean (GRIM)

For integer-scale data (e.g., Likert scales), the reported mean is mathematically impossible given the sample size. Mean × N must yield an integer for integer data.

Detection method (Brown & Heathers 2017):

M = 3.33, N = 20, Scale = 1-7
Sum needed: 3.33 × 20 = 66.6
66.6 is not an integer → Impossible mean

Impossible SD (GRIMMER)

The reported standard deviation is mathematically impossible given the sample size and scale constraints. Extends GRIM logic to variability measures.

Detection method (Anaya 2016):

Check if SD is non-negative and possible given scale bounds and N.

Evidence Quality

Insufficient Evidence

No matching evidence was found within the document to evaluate this claim. Common for review articles, meta-analyses, or claims citing external sources. This is not necessarily an error—it indicates the claim couldn't be evaluated with available evidence.

Note:

This flag suggests manual review, not that the claim is problematic.

Validation Evidence

Our detection methods are validated against established benchmarks and real-world papers with known issues. All results are reproducible.

Statcheck Benchmark

We validated our statistical error detection against the Hartgerink 2016 Statcheck dataset, containing 155,000 statistical results from psychology papers.

Metric	Result	Target
P-value calculation agreement	**100.0%**	90%+
Error detection recall	**95.2%**	85%+
Gross error recall	**94.4%**	80%+
Precision	**95.7%**	90%+
F1 Score	**95.1%**	85%+

Confusion Matrix

	Predicted Error	Predicted No Error
Actual Error	23,657 (TP)	1,192 (FN)
Actual No Error	6,342 (FP)	116,653 (TN)

What this proves:
Our mathematical implementation is correct—we compute p-values identically to Statcheck for t-tests, F-tests, χ^2 tests, correlations, and z-tests.

Famous Failed Papers

We tested against papers with known replication failures or author disavowals to validate our detection of real-world issues.

Paper	Year	Ground Truth	Errors Found	Detected
Power Posing	2010	Author disavowed (2016)	2	✓
Facial Feedback	1988	Failed Many Labs replication	2	✓
Ego Depletion	1998	Failed Registered Replication Report	2	✓
Elderly Priming	1996	Failed replication (Doyen 2012)	0	✗ Methodological issues (experimenter effects), not statistical errors

Money Priming	2006	Failed Many Labs (1/36 labs)	0	✗ Methodological issues, not statistical errors
Bem Precognition	2011	Highly controversial, failed replications	0	✗ No gross statistical errors detected
Marshmallow Test	1990	Conceptual replication failure (Watts 2018)	0	✗ Conceptual issues (SES confounds), not statistical errors
Stereotype Threat	1995	Effect size concerns, mixed replications	0	✗ No gross statistical errors detected

Key Finding: Facial Feedback Main Result

The 1988 Strack et al. "pen in teeth" study—a foundational paper in embodied cognition—contains a gross statistical error in its main result:

```
Reported:  t(89) = 1.85, p = .03 (significant)
Computed:  p = 0.068 (NOT significant)
```

The main result claiming that holding a pen in teeth improves humor ratings is based on an **incorrect p-value**. The effect is not statistically significant at conventional thresholds.

Internal Test Suite

We maintain an internal test suite of 98 test cases including synthetic papers with planted errors and real papers with known issues.

Metric	Value
Total test cases	98
Passed	81
Pass rate	82.7%

Academic Foundations

Our detection methods are grounded in peer-reviewed research on statistical error detection and scientific methodology. Each technique is based on established academic work.

Statcheck: P-Value Verification

Our p-value recalculation method is based on Statcheck, developed by Nuijten et al. (2016). Their analysis of 250,000+ p-values from psychology articles found:

- **49.6%** of papers contained at least one statistical inconsistency
- **12.9%** had "gross errors" where significance status was affected
- Errors were equally distributed across top and bottom journals

We implement the same recalculation logic for t-tests, F-tests, χ^2 tests, correlations, and z-tests, achieving 100% agreement on p-value calculations.

GRIM Test: Impossible Means

The GRIM (Granularity-Related Inconsistency of Means) test was developed by Brown & Heathers (2017). For integer-scale data (e.g., Likert scales 1-7), they showed that:

- Mean × N must equal an integer (or close to one with rounding)
- Applied to 260 papers: **36%** contained at least one impossible mean
- Simple arithmetic check catches fabricated or misreported data

GRIMMER Test: Impossible SDs

The GRIMMER test (Anaya 2016) extends GRIM logic to standard deviations. SD must be mathematically possible given N and scale constraints, providing an additional check for data integrity.

Power Analysis and Sample Size

Our underpowered detection is informed by extensive research on statistical power in science:

- **Ioannidis (2005)**: "Why most published research findings are false" demonstrated how low power leads to unreliable findings
- **Button et al. (2013)**: Found median power in neuroscience was ~21%, leading to inflated effect sizes and low replication rates

Replication Crisis Ground Truth

Papers that failed major replication attempts provide ground truth for validating our detection methods:

- **Many Labs (Klein et al. 2014)**: Large-scale replications of classic effects
- **Open Science Collaboration (2015)**: Found only 36% of psychology findings replicated, with effect sizes typically half of originals

We use these papers to test whether our system identifies real issues without producing false positives.

Claim-Evidence-Burden Analysis

Our semantic analysis of claims exceeding their evidence is novel but grounded in established scientific principles:

- **Causal inference**: Only randomized controlled trials can establish causation; observational studies establish association

- **External validity**: Small, narrow samples cannot support claims about broad populations (the "WEIRD" problem)
- **Statistical power**: Small samples yield unreliable estimates regardless of p-value significance

Limitations & Honest Assessment

We believe in transparency about what our system can and cannot do. No automated tool can replace expert human review.

What We Detect Well

Issue Type	Detection Quality	Notes
P-value calculation errors	✓ Excellent	100% agreement with Statcheck
Gross errors (significance flips)	✓ Excellent	94.4% recall
Causal claims from correlational data	✓ Good	When study design is explicit
Small sample + broad claims	✓ Good	When sample size is reported

What We Miss

Issue Type	Detection Quality	Why
Methodological flaws	✗ Cannot detect	Experimenter effects, demand characteristics, confounds
P-hacking / selective reporting	✗ Cannot detect	Requires access to unreported analyses

Data fabrication	◐ Partial	GRIM/GRIMMER catch some, but not all
Theoretical errors	✗ Cannot detect	Wrong statistical test choice, inappropriate model
One-tailed test issues	◐ Partial	Detection of directional tests is imperfect

Known Limitations

1. Precision vs. Recall Tradeoff
Our system is designed for **precision over recall**. We prefer to miss some issues rather than flood users with false positives. Not all problematic claims will be flagged.

2. Evidence Matching Limitations
Claims are matched to evidence using semantic similarity. This can miss matches when the claim and evidence use very different terminology, or produce spurious matches when unrelated text is superficially similar.

3. PDF Extraction Quality
Our analysis depends on PDF text extraction. Complex layouts, scanned PDFs, or unusual formatting can degrade extraction quality and affect results.

4. Domain Limitations
Our validation is primarily on psychology and biomedical papers.

> Performance on physics, chemistry, or other domains with different statistical conventions may differ.

What This Tool Is NOT

- **Not peer review** — Cannot evaluate theoretical contributions, novelty, or importance
- **Not fraud detection** — Finding statistical errors ≠ finding misconduct
- **Not a quality stamp** — Absence of flags does not mean a paper is good
- **Not definitive** — All flags are potential issues for human review

Appropriate Uses

- Pre-submission check for authors to catch errors before publication
- Quick screening during peer review to prioritize manual checking
- Teaching tool to illustrate common statistical issues
- Research tool for studying error prevalence in literature

Reproducibility

All benchmark results are reproducible. Our code is open source and benchmark data is publicly available.

Running the Benchmarks

```
# Clone the repository
git clone https://github.com/validate-science/validate-science.git
cd validate-science

# Install dependencies
npm install

# Run Statcheck benchmark (requires dataset download)
npm run benchmark

# Run full benchmark suite and freeze results
npm run benchmark:freeze
```

Statcheck Dataset

The Statcheck benchmark uses the Hartgerink 2016 dataset:

- **Source:** OSF Repository (osf.io/gdr4q)
- **Citation:** Nuijten, M. B., et al. (2016). Behavior Research Methods, 48(4), 1205-1226.
- **Contents:** 258,103 statistical results from 30,717 psychology articles

Version Information

Component	Value
Methodology Version	`v1.0.0-2025-12-29`
Pipeline Version	`v1.0.0`
Prompt Version	`v1.0`
Git Commit	`9783de5+dirty`
Benchmark Date	2025-12-29

Benchmark Workflow

To create a new benchmark version:

1. Update `PIPELINE_VERSION` in `src/services/version.ts`
2. Run `npm run benchmark:freeze` to save results
3. Run `npm run benchmark:publish <version>` to publish
4. Results are saved to `data/benchmarks/`

See `docs/BENCHMARKING.md` for full documentation.

Code References

Component	File
Statistical Validator	`src/services/statistical-validator.ts`
Claim Extractor	`src/services/claim-extractor.ts`
Burden Checker	`src/services/burden-checker.ts`
Statcheck Benchmark	`scripts/benchmark-statcheck.ts`

Benchmark Freeze	`scripts/benchmark-freeze.ts`

Version History

Each methodology version represents a frozen snapshot of benchmark results at a point in time. Older versions remain available for reference.

Version	Date	Pipeline	Key Changes	Status
v1.0.0-2025-12-29	2025-12-29	v1.0.0	Initial release	**Current**
v1.0.0-2025-12-28	2025-12-28	v1.0.0	Initial release	Archived

Version Naming

Versions follow the format `v{semver}-{YYYY-MM-DD}`:

- **semver**: Semantic version of the pipeline (MAJOR.MINOR.PATCH)
- **date**: Date the benchmark was frozen

Multiple benchmarks may exist for the same pipeline version if run on different dates. Only one version is published as "current" at any time.

References

[1] Anaya, J. (2016). The GRIMMER test: A method for testing the validity of reported measures of variability. *PeerJ Preprints*, 4, e2400v1. https://doi.org/10.7287/peerj.preprints.2400v1

[2] Brown, N. J., & Heathers, J. A. (2017). The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, 8(4), 363-369. https://doi.org/10.1177/1948550616673876

[3] Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365-376. https://doi.org/10.1038/nrn3475

[4] Carney, D. R., Cuddy, A. J., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science*, 21(10), 1363-1368. https://doi.org/10.1177/0956797610383437

[5] Hartgerink, C. H. J. (2016). 688,112 statistical results: Content mining psychology articles for statistical test results [Data set]. *Open Science Framework*. https://osf.io/gdr4q/

[6] Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124

[7] Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology*, 45(3), 142-152. https://doi.org/10.1027/1864-9335/a000178

[8] Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods*, 48(4), 1205-1226. https://doi.org/10.3758/s13428-015-0664-2

[9] Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. https://doi.org/10.1126/science.aac4716

[10] Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54(5), 768-777. https://doi.org/10.1037/0022-3514.54.5.768

validate.science · Claim-Level Epistemic Risk Assessment

Home · Download PDF · GitHub

Generated: 2026-06-18