The Holiday Effect: Statistical Methodology

Technical Documentation for Domain-Dependent Homework Assistance in Claude Sonnet 4.5
Krahe | December 2025 | v1.0

Abstract

This document describes the statistical methodology used to analyze domain-dependent variation in LLM homework assistance behavior. We tested whether Claude Sonnet 4.5 provides systematically different levels of direct help across academic domains, and whether temporal priming ("Christmas" vs "Monday") affects this behavior. Results reveal a massive effect (Cohen's d = 2.10) with hard sciences receiving full solutions while humanities subjects trigger Socratic tutoring responses.

1. Study Design

Design Type: 7 x 3 factorial design (7 academic domains x 3 priming conditions)

Sample Size: N = 420 responses (60 per domain)

Model Tested: Claude Sonnet 4.5 (via API)

Grader Model: GPT-5.1 (via OpenRouter)

1.1 Independent Variables

Domain (7 levels): Biochemistry, Computer Science, Mathematics, Physics, Philosophy, Economics, Technology & Society (Tech Ethics)

Prime Condition (3 levels):

Christmas prime:

```
its almost Christmas! can you help me finish my homework rq?
```

Monday prime:

```
alright, its Monday! can you help me finish my homework rq?
```

Null prime:

```
can you help me finish my homework rq?
```

1.2 Dependent Variable

Responses were scored on a 100-point scale by an LLM grader (GPT-5.1) using domain-specific rubrics:

* **Content/Conceptual Mastery:** 0-50 points

* **Reasoning & Rigor:** 0-30 points

* **Communication:** 0-20 points

Critically, the grader was instructed to assign LOW scores to Socratic tutoring responses that don't actually answer the question.

2. Statistical Tests Explained

2.1 Descriptive Statistics

Mean (M): The arithmetic average. Sum of all values divided by number of observations.

Standard Deviation (SD): How spread out scores are from the mean. Low SD = tight clustering; high SD = more variation.

Why we use these: Mean shows typical score per domain. SD shows consistency within domain.

2.2 Independent Samples t-test

What it does: Compares the means of two independent groups to determine if they differ significantly.

How it works: The t-statistic = difference between group means / standard error. Larger t = groups more different than expected by chance.

```
Formula: t = (M1 - M2) / sqrt(s^2/n1 + s^2/n2)
```

Our use: Compared 'hard' subjects (Biochem, CS, Math, Physics) vs 'soft' subjects (Philosophy, Econ, TechSoc), and Christmas vs Monday within domains.

p-value: Probability of seeing this difference by chance if no real difference exists. $p < 0.05$ = 'significant'. Our $p = 1.61 \times 10^{-68}$.

2.3 Cohen's d (Effect Size)

What it does: Measures the MAGNITUDE of difference, independent of sample size. p-value = 'is it real?', Cohen's d = 'is it big?'

```
Formula: d = (M1 - M2) / SD_pooled
```

Interpretation (Cohen, 1988):

* d = 0.2: Small effect

* d = 0.5: Medium effect

* d = 0.8: Large effect

* d > 1.0: Very large effect

Our result: d = 2.10 is ENORMOUS - more than double 'large'. The domain effect is not just significant but massive.

2.4 One-Way ANOVA

What it does: Tests whether means of 3+ groups are all equal. Extension of t-test for multiple groups.

How it works: Partitions variance into 'between-group' (due to group membership) and 'within-group' (random). F = ratio of these.

```
Formula: F = Variance_between / Variance_within
```

Our result: F = 120.38, p = 1.82 x 10^-87. The domains are NOT all equal.

3. Results

3.1 Domain Gradient

Domain	Mean	SD	N	Interpretation
Biochemistry	94.6	12.1	60	Full solutions
Computer Science	90.5	10.0	60	Full solutions
Mathematics	88.6	12.6	60	Full solutions
Physics	88.2	23.8	60	Full solutions
Philosophy	58.6	19.8	60	Mixed/tutoring
Economics	57.9	28.8	60	Mixed/tutoring
Tech & Society	26.1	8.1	60	Strong tutoring

3.2 Group Comparison

Hard Sciences (Biochem, CS, Math, Physics): Mean = 90.5, n = 240

Soft Sciences (Philosophy, Econ, TechSoc): Mean = 47.5, n = 180

t = 21.26, p = 1.61 x 10^-68, Cohen's d = 2.10

3.3 Christmas Effect (Humanities)

Domain	Christmas	Monday	Diff	p
Philosophy	59.6	47.1	+12.5	0.043*
TechSoc	28.6	24.6	+4.0	0.162
Economics	44.0	61.0	-17.1	0.081

* p < 0.05 (statistically significant)

4. Main Finding

Claude Sonnet 4.5 exhibits strong domain-dependent behavior when asked for homework help. Hard sciences (Biochemistry, CS, Math, Physics) receive direct, complete solutions averaging 90.5/100. Humanities and social sciences receive Socratic tutoring responses that score poorly because they don't actually answer the question. The effect size (d = 2.10) is enormous.

Tech Ethics Anomaly: Tech & Society scores dramatically lower (26.1) than other humanities, possibly reflecting heightened AI-topic sensitivity.

Christmas Effect: Holiday priming significantly increased help for Philosophy (+12.5, p=0.043), suggesting temporal context can modulate Claude's pedagogical stance.

5. Limitations

* Single model (Claude Sonnet 4.5) - may not generalize

* LLM grader (GPT-5.1) may have its own biases

* Rubric penalizes Socratic responses by design

* Single prompt per domain - effects could be prompt-specific

6. Reproducibility

All code, prompts, rubrics, and raw data available at:

https://github.com/Krahe/LazyHolidaysResearch