The Holiday Effect: Statistical Methodology

Technical Documentation for Domain-Dependent Homework Assistance in Claude Sonnet 4.5
Krahe | December 2025 | v2.0

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 domain effect (Cohen's d = 2.10) with hard sciences receiving full solutions while humanities subjects trigger Socratic tutoring responses. The Christmas priming effect was significant only for Philosophy (+12.5 points, p=0.043); no overall Christmas effect was found when collapsing across domains (p=0.950).

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, temperature=0)

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 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 deflect rather than 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): Measure of score dispersion. Low SD = tight clustering; high SD = more variation.

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: (1) Compared 'hard' vs 'soft' subjects for main domain effect. (2) Compared Christmas vs Monday within each domain. (3) Compared overall Christmas vs Monday collapsed across domains.

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, d=0.5 medium, d=0.8 large, d>1.0 very large

2.4 One-Way ANOVA

What it does: Tests whether means of 3+ groups are all equal.

How it works: F = between-group variance / within-group variance. Significant F means at least one group differs.

2.5 p-value Interpretation

The p-value is the probability of observing results this extreme if the null hypothesis (no effect) were true.

* $p < 0.05$: Conventionally 'statistically significant'
* $p < 0.01$: Highly significant
* $p < 0.001$: Very highly significant

Note: Statistical significance does not imply practical importance. Always consider effect size.

3. Results

3.1 Main Finding: Domain Effect

Claude Sonnet 4.5 exhibits massive domain-dependent behavior. Hard sciences receive direct solutions (mean=90.5). Humanities receive Socratic tutoring that scores poorly because it doesn't answer the question (mean=47.5). Effect size d=2.10 is enormous.

Domain	Mean	SD	N	Tutoring %	Interpretation
Biochemistry	95	12.1	60	3%	Full solutions
Computer Science	90	10.0	60	3%	Full solutions
Mathematics	89	12.6	60	3%	Full solutions
Physics	88	23.8	60	10%	Full solutions
Philosophy	59	19.8	60	47%	Mixed/tutoring
Economics	58	28.8	60	45%	Mixed/tutoring
Tech & Society	26	8.1	60	100%	Strong refusal

Group Comparison (Hard vs Soft):

* 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

One-Way ANOVA (all 7 domains): F = 120.38, p = 1.82 x 10^-87

3.2 Christmas Effect Analysis (Original Research Question)

We originally set out to test whether temporal priming ('it's almost Christmas!') would make Claude lazier or more helpful. We tested this across ALL 7 domains.

Domain	Monday	Christmas	Diff	p-value	Significant?
Philosophy	47.1	59.6	+12.5	0.043	YES *
Biochemistry	90.5	96.3	+5.8	0.210	No
Tech Ethics	24.6	28.6	+4.0	0.162	No
CS	86.9	90.9	+4.0	0.293	No
Math	86.0	87.2	+1.2	0.792	No
Physics	92.5	80.5	-12.0	0.128	No
Economics	61.0	44.0	-17.1	0.081	No

* $p < 0.05$ (statistically significant)

Overall Christmas Effect (collapsed across all domains):

Christmas mean: 69.6 (n=140) vs Monday mean: 69.8 (n=140)

t = -0.06, p = 0.950 - NO OVERALL EFFECT

Interpretation:

The original 'holiday effect' hypothesis (LLMs are lazier during holidays) was NOT supported. There is no overall difference between Christmas and Monday priming (p=0.950). However, Philosophy showed a significant positive Christmas effect (+12.5 points, p=0.043), suggesting the holiday context made Claude MORE willing to help with philosophy homework specifically. Economics showed a trending negative effect (-17.1 points, p=0.081) but did not reach significance. The domain effect massively overwhelms any priming effect.

4. Key Takeaways

1. DOMAIN EFFECT IS MASSIVE: Claude's willingness to help with homework depends dramatically on subject. Cohen's d = 2.10 is more than double what is typically considered 'large'.

2. TECH ETHICS IS SPECIAL: 100% tutoring rate - Claude NEVER gave a direct answer to tech ethics homework. This may reflect heightened sensitivity around AI-related topics.

3. CHRISTMAS EFFECT IS DOMAIN-SPECIFIC: Only Philosophy showed significant Christmas effect. No overall 'lazy holiday Claude' phenomenon exists.

5. Limitations

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

* LLM grader (GPT-5.1) may have systematic biases

* Rubric explicitly penalizes Socratic responses by design

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

* Christmas effect tested with multiple comparisons (7 domains) - Philosophy result should be interpreted cautiously

6. Reproducibility

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

https://github.com/Krahe/LazyHolidaysResearch