# T3 Benchmark Analysis Report - GroupJ (Social Science)

Fernando Torres

January 22, 2026

## Executive Summary

This report analyzes the **GroupJ (Social Science)** dataset for the CS372 T3 Benchmark assignment. The dataset contains **500 validated causal reasoning test cases** in the **Social Science** domain (D10).

**Key Metrics:**

- Total Cases: 500
- Mean Quality Score: 8.53/10
- Schema Compliance: 100%
- Pearl Level Distribution: L1=50, L2=300, L3=150

## Pearl Level Distribution

| Level | Count | Percentage | Target |
|---|---|---|---|
| L1 (Association) | 50 | 10.0% | 50 (10%) |
| L2 (Intervention) | 300 | 60.0% | 300 (60%) |
| L3 (Counterfactual) | 150 | 30.0% | 150 (30%) |
| **Total** | **500** | **100%** | **500** |

### Level Descriptions

- **L1 (Association)**: Tests whether LLMs can distinguish justified from unjustified causal claims.
- **L2 (Intervention)**: Tests causal disambiguation and wise refusal generation.
- **L3 (Counterfactual)**: Tests reasoning about alternative worlds.

## Label Distribution

### L1 Labels (WOLF/SHEEP/AMBIGUOUS)

| Label | Count | Description |
|---|---|---|
| W | 16 | WOLF - Unjustified claim |
| S | 32 | SHEEP - Valid inference |
| A | 2 | AMBIGUOUS |

### L2 Labels

All 300 L2 cases are labeled **NO**.

### L3 Ground Truth

| Ground Truth | Count |
|---|---|
| VALID | 43 |
| INVALID | 44 |
| CONDITIONAL | 63 |

## Trap Type Distribution (L2)

| Trap | Family | Count | Description |
|---|---|---|---|
| T1 | F1 | 65 | Selection Bias |
| T2 | F1 | 8 | Survivorship |
| T3 | F1 | 35 | Self-Selection |
| T4 | F1 | 6 | Attrition |
| T5 | F2 | 8 | Regression Mean |
| T6 | F2 | 35 | Base Rate |
| T7 | F3 | 36 | Confounding |
| T8 | F3 | 35 | Mediated |
| T9 | F3 | 10 | Collider |
| T10 | F4 | 10 | Reverse Cause |
| T11 | F4 | 8 | Bidirectional |
| T12 | F4 | 6 | Feedback |
| T13 | F5 | 8 | Ecological |
| T14 | F5 | 8 | Simpsons |
| T15 | F6 | 8 | Proxy |
| T16 | F6 | 8 | Oversimplify |
| T17 | F6 | 6 | Black Box |

## Difficulty Distribution

| Difficulty | Count | Percentage |
|---|---|---|
| Easy | 88 | 17.6% |
| Medium | 212 | 42.4% |
| Hard | 200 | 40.0% |

## Quality Metrics

| Metric | Value |
|---|---|
| Mean Score | 8.53 |
| Min Score | 8.00 |
| Max Score | 9.50 |
| Std Dev | 0.36 |

### Validation Results

- Schema Compliance: 500/500 (100%)
- Duplicate Detection: 0 duplicates
- All required fields: Present

# Methodology

## Multi-Agent Workflow

1. Generator Agents: Created cases
2. Schema Validator: JSON compliance
3. Content Validators: Quality scoring
4. Cross Validator: Duplicate detection
5. Quality Judges: Trap verification
6. Correction Agents: Issue resolution

## Validation Pipeline

- JSON schema validation (V4.0)
- Content scoring (threshold 8.0/10)
- Duplicate detection (similarity less than 0.75)
- Trap type verification
- Distribution balance checks

## Example Cases

### L1 Example

Case ID: T3-J-L1-0001

Scenario: An organization reports a very positive statistic for Average star rating based only on observations from a subset of people. The subset is formed by Who leaves reviews that is voluntary or outcome-de...

Label: S

### L2 Example

Case ID: T3-J-L2-0051

Trap Type: T8

Scenario: A manager must choose between two interventions (Program A vs Program B) to improve test pass rate. A pilot dataset reports that, overall, intervention A has a lower test pass rate than intervention B...

### L3 Example

Case ID: T3-J-L3-0351

Scenario: A city implemented rent control in 2017. Five years later, the rental housing stock had decreased by 15% as landlords converted units to condos or let buildings deteriorate. New construction also slow...

---

*Generated by Claude Code for CS372 T3 Benchmark*