T3 Benchmark Analysis Report - GroupJ (Social Science)

Fernando Torres

January 28, 2026

Executive Summary

This report analyzes the **GroupJ (Social Science)** dataset for the CS372 T3 Benchmark Assignment 2. The dataset contains **500 validated causal reasoning test cases** in the **Social Science** domain (D10), migrated to the Assignment 2 Appendix B schema format.

Key Metrics:

- Total Cases: 500
- Mean Quality Score: 8.53/10
- Schema Compliance: 100% (Appendix B compliant)
- Pearl Level Distribution: L1=50, L2=300, L3=150
- Trap Type Corrections: 74 cases corrected

1. Summary of Unvalidated vs. Validated Dataset

Metric	Before Migration	After Migration
Total Cases	500	500
Schema Version	V2.0 (Pre-remediation)	Appendix B (Assignment 2)
L1 Labels	W/S/A format	YES/NO/AMBIGUOUS format
L1 Trap Types	Mixed T/W types	W1-W10, S1-S8, A only
L3 Trap Types	Mixed T/F types	F1-F8, DomainExt only
variables.Z	Object format	Array of strings
trap field	Flat fields	Nested object
Required Fields	Partial	Complete (21 fields)

Key Improvements: - Standardized ID format: `T3-BucketLarge-J-{level}.{seq}` - Transformed L1 labels from W/S/A to YES/NO/AMBIGUOUS - Corrected 23 L1 trap types (T→W mapping) - Corrected 51 L3 trap types (T→F mapping) - Restructured variables.Z from object to array format - Created nested trap object with type, type_name, subtype, subtype_name - Added all missing required fields per Table 9

2. Pearl Level Distribution

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

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

3. Label Distribution

L1 Labels (YES/NO/AMBIGUOUS) - Per Table 10

Label	Count	Description
YES	32	Valid causal claim (SHEEP cases)
NO	16	Invalid causal claim (WOLF cases)
AMBIGUOUS	2	Unclear or conditional relationship
Total	**50**	

L2 Labels - Per Table 10

Label	Count	Description
NO	300	All L2 cases labeled NO (invalid causal claims)

L3 Labels (VALID/INVALID/CONDITIONAL) - Per Table 10

Label	Count	Percentage
VALID	43	28.7%
INVALID	44	29.3%
CONDITIONAL	63	42.0%
Total	**150**	**100%**

4. Trap Type Distribution

L1 Trap Types (W1-W10, S1-S8, A)

Category	Trap Types	Count
WOLF (W-series)	W1, W2, W3, W5, W7, W8, W9, W10	16
SHEEP (S-series)	S1, S2, S3, S4, S5	32
AMBIGUOUS	A	2

WOLF Trap Type Breakdown (After Correction)

Type	Name	Count
W1	Selection Bias	10
W2	Confounding	2
W3	Measurement Error	4
W5	Collider Bias	2
W7	Reverse Causation	12
W8	Ecological Fallacy	6
W9	Temporal Ambiguity	2
W10	Missing Data	1

L2 Trap Types (T1-T17)

Trap	Family	Count	Description
T1	F1: Selection	65	Selection Bias
T2	F1: Selection	8	Survivorship
T3	F1: Selection	35	Collider Bias
T4	F1: Selection	6	Immortal Time
T5	F2: Statistical	8	Regression to Mean
T6	F2: Statistical	35	Ecological Fallacy
T7	F3: Confounding	36	Confounder
T8	F3: Confounding	35	Simpson's Paradox
T9	F3: Confounding	10	Conf-Mediation
T10	F4: Direction	10	Reverse Causation
T11	F4: Direction	8	Feedback Loop
T12	F4: Direction	6	Temporal Precedence
T13	F5: Information	8	Measurement Error
T14	F5: Information	8	Recall Bias
T15	F6: Mechanism	8	Mechanism Confusion
T16	F6: Mechanism	8	Goodhart's Law
T17	F6: Mechanism	6	Backfire Effect

L3 Trap Types (F1-F8, DomainExt) - After Correction

Type	Name	Count
F1	Deterministic	14
F2	Probabilistic	12
F3	Overdetermination	13
F4	Structural	49
F5	Temporal	32
F6	Epistemic	8
F7	Attribution	8
F8	Moral/Legal	6
DomainExt	Domain Extension	8

5. Difficulty Level Distribution

Difficulty	Count	Percentage	Target Ratio
Easy	88	17.6%	~25%
Medium	212	42.4%	~50%
Hard	200	40.0%	~25%
Total	**500**	**100%**	1:2:1

Note: Distribution shows deviation from 1:2:1 target with fewer Easy cases and more Hard cases, reflecting the inherent complexity of Social Science causal reasoning scenarios.

6. Score Summary

Unvalidated Dataset Scores

Metric	Value
Mean Score	8.50
Min Score	8.00
Max Score	8.50

Validated Dataset Scores

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

Validation Impact

- Schema Compliance: 500/500 (100%)
- Duplicate Detection: 0 duplicates found
- All 21 required fields: Present in all cases
- L1 trap type corrections: 23 cases (T1→W1, T7→W7, T8→W8)
- L3 trap type corrections: 51 cases (T7/T9→F4, T10/T11/T12→F5)
- Total trap corrections: 74 cases

7. Prompt Setup

LLM Configuration

Parameter	Value
Model	Claude (Anthropic)
Temperature	0.7 (generation), 0.0 (validation)
Max Tokens	4096 per case

Multi-Agent Workflow

1. **Existing Case Transformation (240 cases)**:
   - Variable Parser Agent: Array → object format
   - Field Normalizer Agent: V1.0 → V4.0 field names
   - Metadata Enricher Agent: Add validation fields
2. **New Case Generation (260 cases)**:
   - Generator Agents by trap family
   - Schema and Content Validators
   - Quality Judges and Correction Agents

Validation Pipeline

- JSON schema validation (Appendix B format)
- Content scoring (threshold 8.0/10)
- Duplicate detection (similarity < 0.75)
- Trap type verification per level requirements
- Distribution balance checks

Quality Control Measures

- 95%+ pass rate threshold per batch
- Iterative correction loops until quality met
- Final validation sweep for missing fields
- Critical trap type correction phase

8. Example Case

L1 Example (Association Level)


```
{
  "id": "T3-BucketLarge-J-1.1",
  "bucket": "BucketLarge-J",
  "case_id": "0001",
  "pearl_level": "L1",
  "domain": "D10",
  "subdomain": "Digital Media",
  "difficulty": "Easy",
  "is_ambiguous": false,
  "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-dependent.",
  "claim": "An organization reports a very positive statistic for
    Average star rating based only on observations from a subset
    of people",
  "variables": {
    "X": {"name": "Who leaves reviews", "role": "Treatment/Factor"},
    "Y": {"name": "Average star rating", "role": "Outcome"},
    "Z": ["Underlying true outcome (positive/negative)"]
  },
  "trap": {
    "type": "W1",
    "type_name": "Selection Bias",
    "subtype": "Sampling-on-the-Outcome",
    "subtype_name": "Sampling Bias"
  },
  "label": "YES",
  "causal_structure": "Selection into review sample is correlated
    with outcome satisfaction",
  "key_insight": "Voluntary reviews over-represent satisfied
    customers",
  "hidden_timestamp": "What proportion of customers leave reviews?",
  "conditional_answers": {
    "answer_if_condition_1": "If all customers review, rating valid.",
    "answer_if_condition_2": "If only satisfied review, rating biased."
  },
  "wise_refusal": "The reported star rating may not represent the
    true customer experience because reviews are voluntary and
    those who had positive experiences are more likely to leave
    reviews.",
  "gold_rationale": "Selection bias occurs because review submission
    is correlated with satisfaction level.",
  "initial_author": "Fernando Torres",
```

```
  "validator": "Fernando Torres",
  "final_score": 8.5
}
```

Report generated for CS372 Assignment 2 - T3 Benchmark Expansion Migration Date: January 28, 2026 Author: Fernando Torres