“What Were They Thinking?”

Using Open-Text Responses to Validate Constructs in Survey Experiments

Steven Denney

March 2026

Leiden University

Motivation

The Construct Validity Problem

Survey experiments tell us **what** respondents chose.

But did they think about **what we assumed** they were thinking about?

- Conjoint profiles contain many attributes
- Closed-ended items constrain expression
- Treatment effects on Likert scales are not evidence of intended reasoning

The Solution: Open-Text Questions

Key idea: Embed an open-text question after the conjoint task

"Why did you choose this person?"

Three things we can learn:

1. Did respondents **notice** the experimental manipulation?
2. Did they **interpret it as intended**?
3. What **other considerations** shaped their choice?

Challenge: How do we systematically analyze thousands of responses?

Two Complementary Approaches

1. Topic Modeling (STM)

Discovery

- What did respondents discuss?
- Data-driven, no prior categories
- Treatment as covariate

2. LLM Classification

Confirmation

- Do responses match theory?
- Theory-driven codebook
- Scalable "close reading"

Triangulation: If both approaches converge, we have strong validity evidence.

Empirical Application

The "Cues of Commitment" Conjoint

Data: Conjoint surveys on immigrant naturalization

- South Korea ($N = 1{,}999$) and Taiwan ($N = 2{,}050$), fielded 2024
- Paired conjoint with 12 randomized immigrant attributes

Key attribute: **Social integration**

- "Volunteers weekly at community center"
- "Co-hosts multicultural events"
- "Only interacts with foreign residents"
- *Not shown* (randomly withheld)

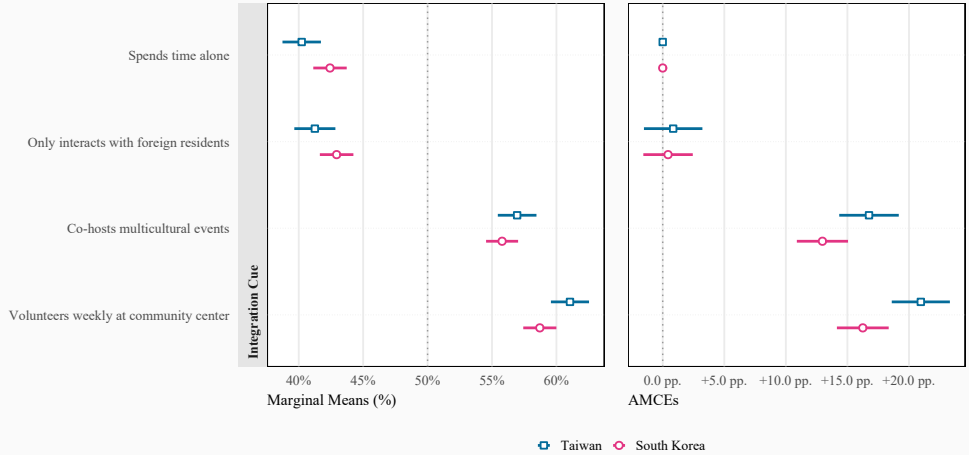
Validation question: Does showing the social integration cue shift what respondents *write about*?

See the Conjoint Design

Preview the survey instrument:

https://leidenuniv.eu.qualtrics.com/jfe8/preview/previewId/cb34058b-3eda-44a1-8abc-ca173b62e5cb/SV_eXS0hZgBUOCBaZM?Q_CHL=preview&Q_SurveyVersionID=current

The Integration Cue: Main Effects



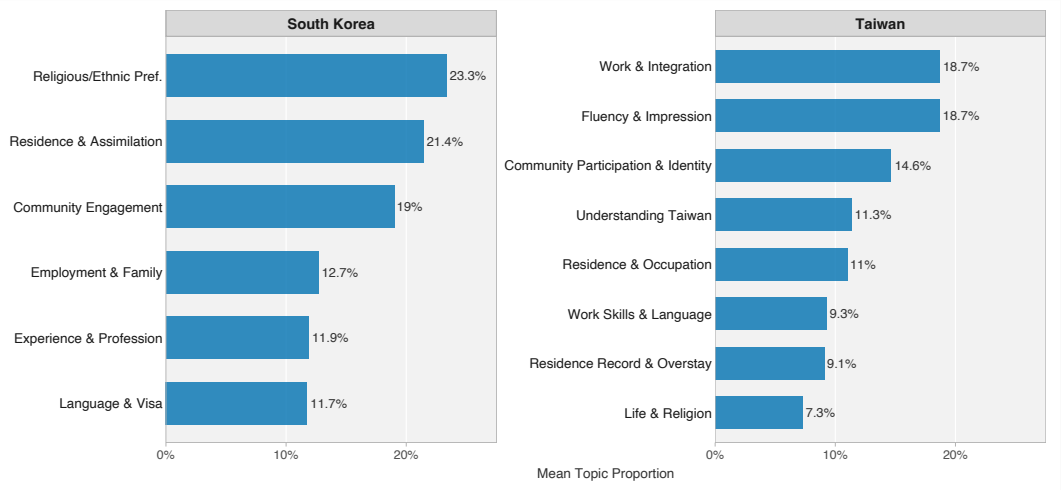
Approach 1: Structural Topic Models

STM: Incorporating Treatment into Estimation

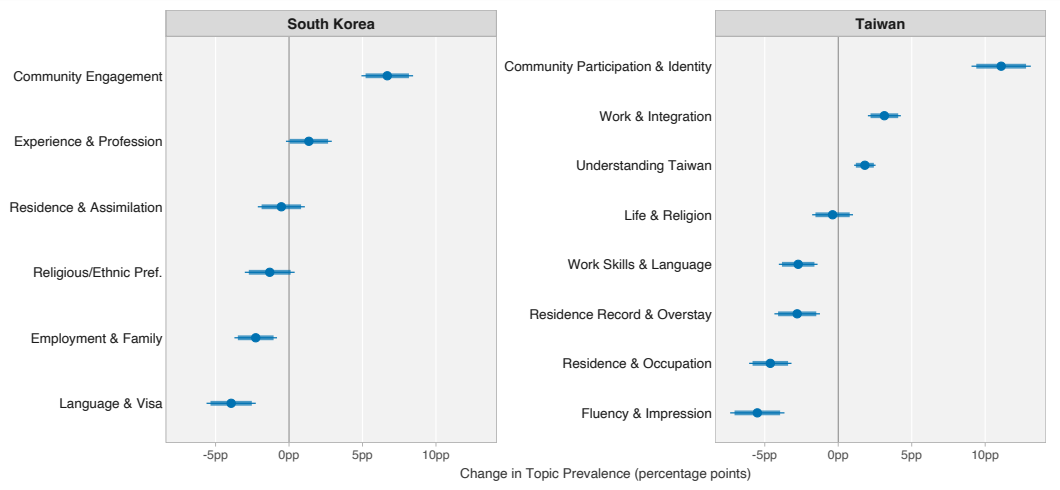
Structural Topic Models (Roberts et al. 2014):

- Treatment affects topic prevalence **during estimation**, not post-hoc
- Separate models per country (Korean and Chinese require different preprocessing)
- South Korea: $K = 6$ topics; Taiwan: $K = 8$ topics

What Did Respondents Write About?



STM: Effect of the Social Integration Cue



STM Takeaway

The civic cue shifts reasoning **toward community engagement**.

But language and residence topics *decrease* with the cue.

These are **substitutes** for civic engagement, not signals of it.

This finding informed the codebook design for LLM classification.

Approach 2: LLM Classification

Codebook v1: The Initial Scheme

Following Halterman & Keith (2025, *Political Analysis*):

Code	Definition
`civic_commitment`	Civic engagement, volunteering, commitment signals, *including language and residence*
`identity_concern`	Ethnic, cultural, or national identity concerns
`economic_focus`	Jobs, skills, welfare, economic contribution
`other`	Too vague, too short, or off-topic

Each code includes definition, inclusion/exclusion criteria, and few-shot examples in Korean and Chinese.

The Problem with v1

civic_commitment was over-determined

- **561 responses** (13.9%) coded as civic but only about language or residence
- STM showed these topics *decrease* with the civic cue

The STM (discovery) informed the codebook (confirmation)

Codebook v2: The Fix

Split `functional_integration` from `civic_commitment`:

- `civic_commitment`: participation in **community life** (volunteering, organizing, prosocial engagement)
- `functional_integration`: individual-level **adaptation** (language, residence, law-abiding behavior)

Now we can test whether the cue specifically increases active civic reasoning, or just practical credentials.

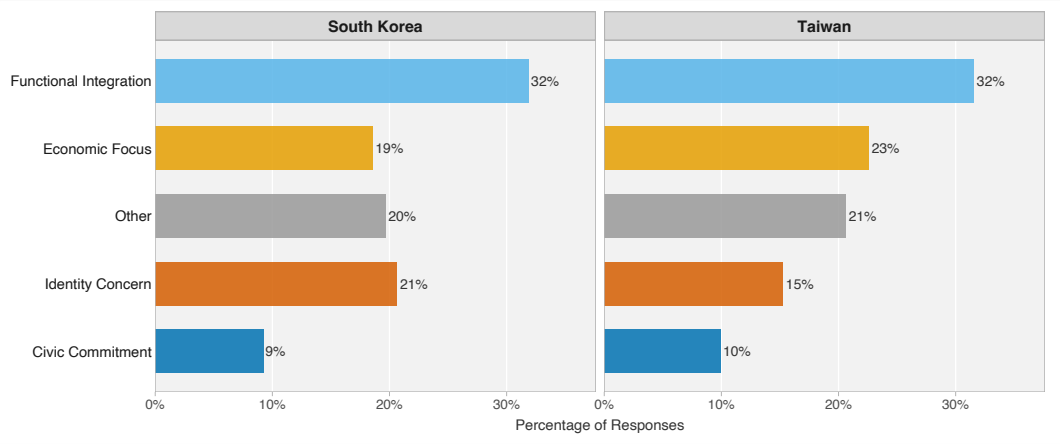
LLM Classification Pipeline

6 models: 1 proprietary (GPT-4o-mini) + 5 open-weight (3B to 72B)

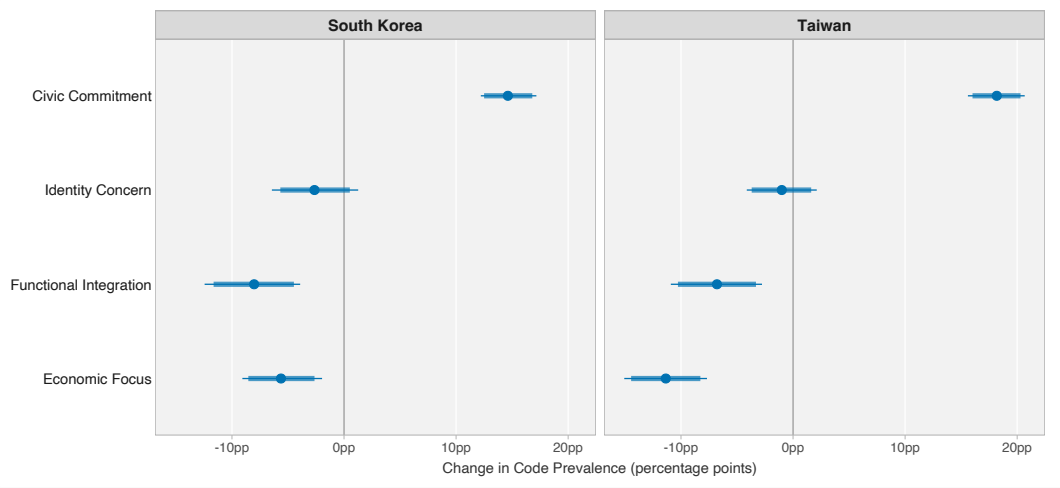
- Open-weight models ensure reproducibility (Barrie et al. 2025)
- Majority vote: 4+ of 6 models must agree
- Cross-model agreement: Fleiss' $\kappa = \mathbf{0.745}$

{Even a 3B model on a consumer GPU produces directionally consistent results. Agreement with GPT-4o-mini improves with model size (Qwen family: κ 0.74 $\rightarrow$ 0.84 $\rightarrow$ 0.86).}

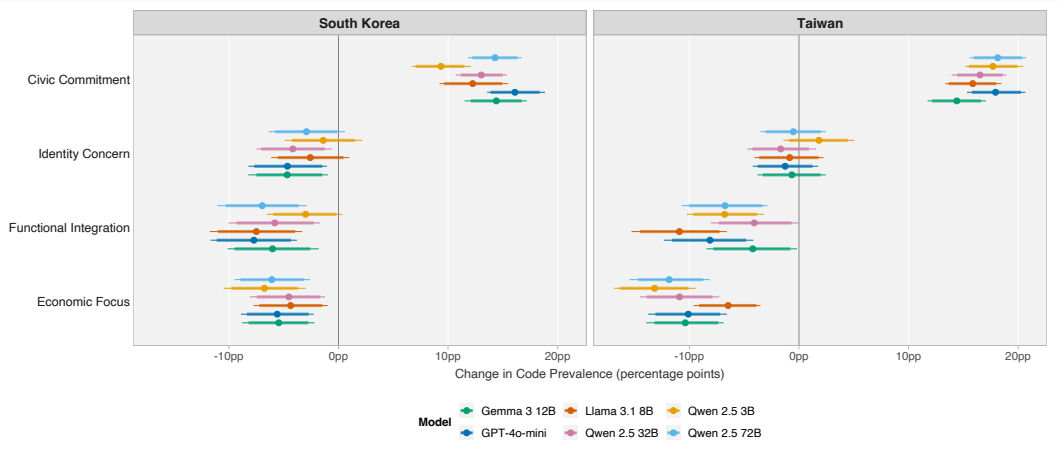
Code Prevalence by Country



LLM: Effect of the Social Integration Cue



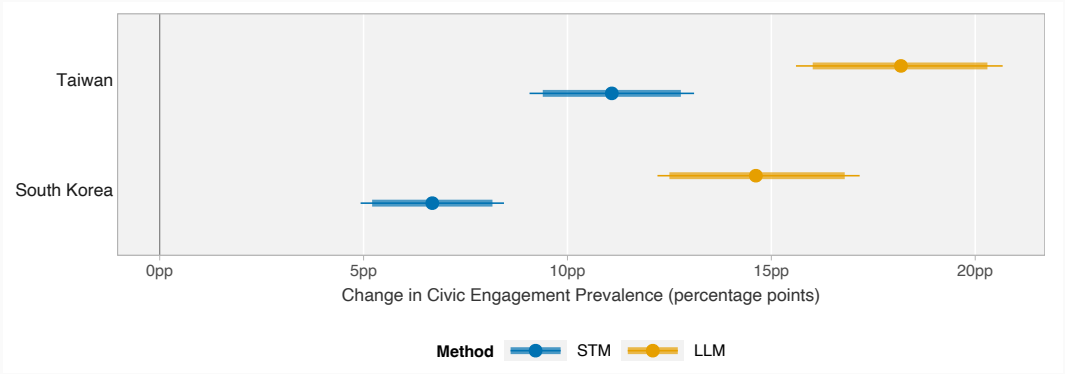
Cross-Model Robustness



{Each model's own classifications (no majority vote). All 6 agree on direction for every code.}

Triangulation

STM + LLM Convergence



Two Methods, One Story

Finding	STM	LLM
Civic engagement ↑	✓	✓
Language/residence ↓	✓	✓
Economic criteria ↓	✓	✓
Identity stable	✓	✓
Larger effects in Taiwan	✓	✓

Convergent validity: data-driven discovery and theory-driven classification agree.

Extensions

Portability: Immigration Admission Conjoint

Applied the same framework to a standard conjoint on immigrant *admission* (Denney & Green 2021), with no within-design treatment randomization.

- LDA + LLM classification converge on the same constructs
- Identifies functional integration as a previously underspecified category
- Works for both intervention-based and conventional conjoint designs

Ground Truth: LLM vs. Human Coders

Can LLMs match human coders? Tested on a migrant entrepreneurship conjoint (Denney, Ward & Green 2023) with **2,009 human-coded responses**.

Comparison	Cohen's κ	Raw agreement
Human vs. Human (PW vs CG)	0.881	92.3%
GPT-4o-mini vs. Human	0.673	78.8%
Llama 3.1 8B vs. Human	0.611	74.5%
GPT-4o-mini vs. Llama 8B	0.717	81.5%

Both LLMs reach **substantial** agreement ($\kappa > 0.6$) with human coders.

The Codebook Iteration Matters

	v1	v2	v3
GPT-4o-mini vs. Human	0.40	0.45	**0.67**
Llama 3.1 8B vs. Human	0.52	0.60	**0.61**

Two rounds of codebook refinement moved GPT-4o-mini from "fair" to "substantial" agreement.

LLM classification is not one-shot-and-go. It requires:

- Classifying a sample, reviewing disagreements, revising the codebook
- Being transparent about the iteration process
- Theoretically motivated categories with precise definitions

Takeaways

Practical Recommendations

1. **Design**: Embed open-text questions after key experimental tasks
   - Short responses (2–20 words) are sufficient
2. **Discovery first**: Fit STM with treatment as covariate
   - Let the data show what respondents discussed
3. **Then confirm**: LLM classification with a rigorous codebook
   - **Iterate**: classify, review, revise, repeat
   - Cross-validate with multiple models (Barrie et al. 2025)
4. **Accessibility**: Open-weight models on consumer hardware work

Summary

1. Open-text questions provide **construct validation** for survey experiments
2. **STM** discovers themes; **LLM classification** tests theory; each can **inform the other**
3. After codebook iteration, LLMs achieve **substantial agreement** with human coders ($\kappa = 0.61$–0.67)
4. The framework is **portable** across conjoint designs and policy domains

References i

Barrie, C., Palmer, A., & Spirling, A. (2025). Replication for language models. Working paper.

Denney, S. & Green, C. (2021). Who should be admitted? *Ethnicities*, 21(1), 120–145.

Denney, S., Ward, P., & Green, C. (2023). Public support for migrant entrepreneurship. *International Migration Review*.

Halterman, A. & Keith, K. (2025). Codebook LLMs. *Political Analysis*.

Roberts, M. E. et al. (2014). Structural topic models for open-ended survey responses. *AJPS*, 58(4), 1064–1082.

Tornberg, P. (2025). LLMs outperform expert coders. *Social Science Computer Review*, 43(6), 1181–1195.