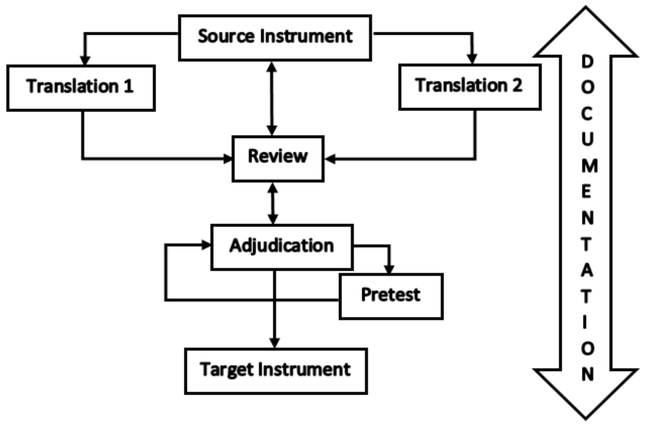
AI-assisted Survey Translation: Exploring the Role of LLMs as Expert Reviewers

Matt Dearstyne, Patricia Goerman, Betsarí Otero Class, Marcus Berger and Mikelyn Meyers
Center for Behavioral Science Methods

United States® Census Bureau

TRAP-D Method for Survey Translation



Review Method 1 (Traditional Team)

- 3 bilingual researchers
- Diverse backgrounds including survey methodology, social sciences, linguistics, translation, and pretesting
- Independent reviews → Consensus meetings → Documentation → Presentation of results
- Focus on Spanish translation review

Review Method 2 (Checklist)

- Structured approach to review using formal checklists (Willis & Lessler, 1999; Dean et al., 2007; Schaad et al., 2021)
- Designed to bring replicability, transparency, and systematic error detection to questionnaire review
- Focus on questionnaire design (English/Spanish)

Procedure

1. **Prompt engineering**: designed separate prompts to match traditional and checklist methods based on training materials that humans received
2. **Interact with LLM**: Sent prompts and data to model (gpt-5-mini), received structured output
3. **Analysis**: Compared results of AI reviews with human reviews

Research Questions and Results

1. Did comments made by AI reviewers match comments made by human reviewers?

Number of Questions Commented on Per Reviewer (N=101)

Review Method	Human only	AI only	Both	Level of Agreement (between human and AI)		
				Same	Partial	Different
Traditional	22	12	21	1	8	12
Checklist	35	0	39	0	7	32

- **Number**: Human reviewers consistently identified more issues with translated survey items than AI reviewers
- **Type**: Comments made only by AI reviewers were mainly focused on alternative word choice and grammatical or typographic errors

Conclusion: With these prompts, AI and human reviewers rarely made same recommendations.

2. Did different prompts change the type of comments made by AI reviewers?

Number of Questions Commented on Per Prompt (AI Only, N=101)

Traditional only	Checklist Only	Both	Level of Agreement (between prompts)		
			Same	Partial	Different
3	9	30	22	8	0

Conclusion: These two prompts made only minor differences in the comments generated by the AI reviewers.

Example Prompt

Evaluate the following Spanish translations of English survey items, paying particular attention to the following points:

Assess questions for inappropriate or ineffective cross-cultural references.

- **Example**: "Do you have health insurance?"
  "¿Tiene usted seguro médico?"
- **Issue**: In countries with nationalized healthcare, respondents may not understand the concept of private health insurance.

Example Output

English	Spanish	AI Comments	AI Suggested Translation
What is your current height?	¿Cuánto mide de altura actualmente?	The Spanish is understandable but slightly unnatural. The suggested translation is clearer and more idiomatic for interviewers.	¿Cuál es su estatura actual?

Future Research

- How do the translations generated by AI reviewers perform compared to human translations?
- What is the quality of the AI-generated comments and recommendations?
- Can Retrieval-Augmented Generation (RAG) improve the quality of AI recommendations and translations?
- Do different LLMs perform better or worse at survey translation tasks?

Contact: matthew.j.dearstyne@census.gov