Can Generative AI Enhance Spanish Questionnaire Translation Review?: An Experimental Comparison of Expert Review Methods

Patricia Goerman, Matt Dearstyne, Betsarí Otero Class, Mikelyn Meyers, and Marcus Berger, U.S. Census Bureau

Presentation for the 81st annual conference of the American Association for Public Opinion Research (AAPOR). Los Angeles, CA: May 13-15, 2026

Disclaimer: This presentation is intended to inform people about research and to encourage discussion. The views expressed are those of the authors and not those of the U.S. Census Bureau.

Introduction: Survey Translation Expert Review

- Questionnaire translation is a complex task
  - Lack of proper attention to different language versions of a survey can put comparability of data at risk
- Best practices: TRAPD: Translation, Review, Adjudication, Pretesting, Documentation. (Cross Cultural Survey Guidelines)
- This study compares AI generated recommendations to human recommendations for each of 2 expert review methods.
  1. Census Bureau team translation expert review method (Traditional Review)
  2. Questionnaire Appraisal Systems checklist method (Checklist review)
  3. Evaluation of usefulness of AI in review process

United States® Census Bureau

Expert Review Methods

Traditional

- 4 bilingual researchers:
  - lead/adjudicator and 3 additional reviewers
  - 2 experienced experts, 1 newer reviewer, 1 independent translator
- Language/cultural expertise and background in:
  - Survey methodology, social sciences, linguistics, translation, pretesting, subject matter expertise related to questionnaire topic/s
- Process:
  - Independent review of translation
  - Consensus meetings among reviewers
  - Write up/documentation
  - Results presentation/discussion with sponsors

Checklist

- Same types of reviewers and review process: individual reviews followed by consensus meetings/joint recommendations.
- Originally created by Willis and Lessler (1999), updated by Dean et al. (2007) and Schaad et al. (2021)
- Main difference: more structured approach to review with formal checklist
- Includes cross cultural and translatability issues such as politeness norms, naming conventions, and assumptions
- Minor modifications made to checklist for this study.

United States® Census Bureau

MD1

Survey Materials Reviewed

Separate Traditional and Checklist reviews of Spanish translations of segments of two survey instruments:

1. National Health Interview Survey (NHIS)
   - Survey used to monitor the health of the U.S. population on a broad range of health topics. Eg. demographic and socioeconomic characteristics of respondents, information on activity limitation, illnesses, chronic conditions, health insurance coverage, use of health care, and other health topics.
2. Census Household Panel (CHP)
   - U.S. Census Bureau survey panel. Purpose: to collect information on topics such as food and nutrition, transportation, employment, and education and to gather data that could be used to improve and inform future surveys. The panel consisted of individuals and households living across the U.S. that agreed to be contacted and invited to participate in surveys
   - *Now called HTOPS: In January 2025, the Census Household Panel transitioned to the Household Trends and Outlook Pulse Survey (HTOPS).*

United States® Census Bureau

Slide 4

MD1 What if we move this slide and the next slide to after the research design?
Matt Dearstyne (CENSUS/DSD FED), 2026-05-01T19:26:51.607

Types of Issues Flagged by Reviewers

- Missing terms or phrases
- Inconsistency in terms used throughout the instruments
- Whether other Census surveys use the same terms
- Adaptations needed across languages
- Mismatch between register or formality across languages
- Grammatical errors/typos
- New wording recommendations

United States® Census Bureau

Human Review Results: Traditional versus Checklist reviews

- Team of researchers coded the comments and recommendations from each method (results described in more detail elsewhere)
- Main impressions:
  - Timing: Checklist method involved more and lengthier consensus meetings
  - Checklist reviews prompted reviewers to look at more general survey methodology concepts, increased number of comments that spanned both English source and Spanish versions

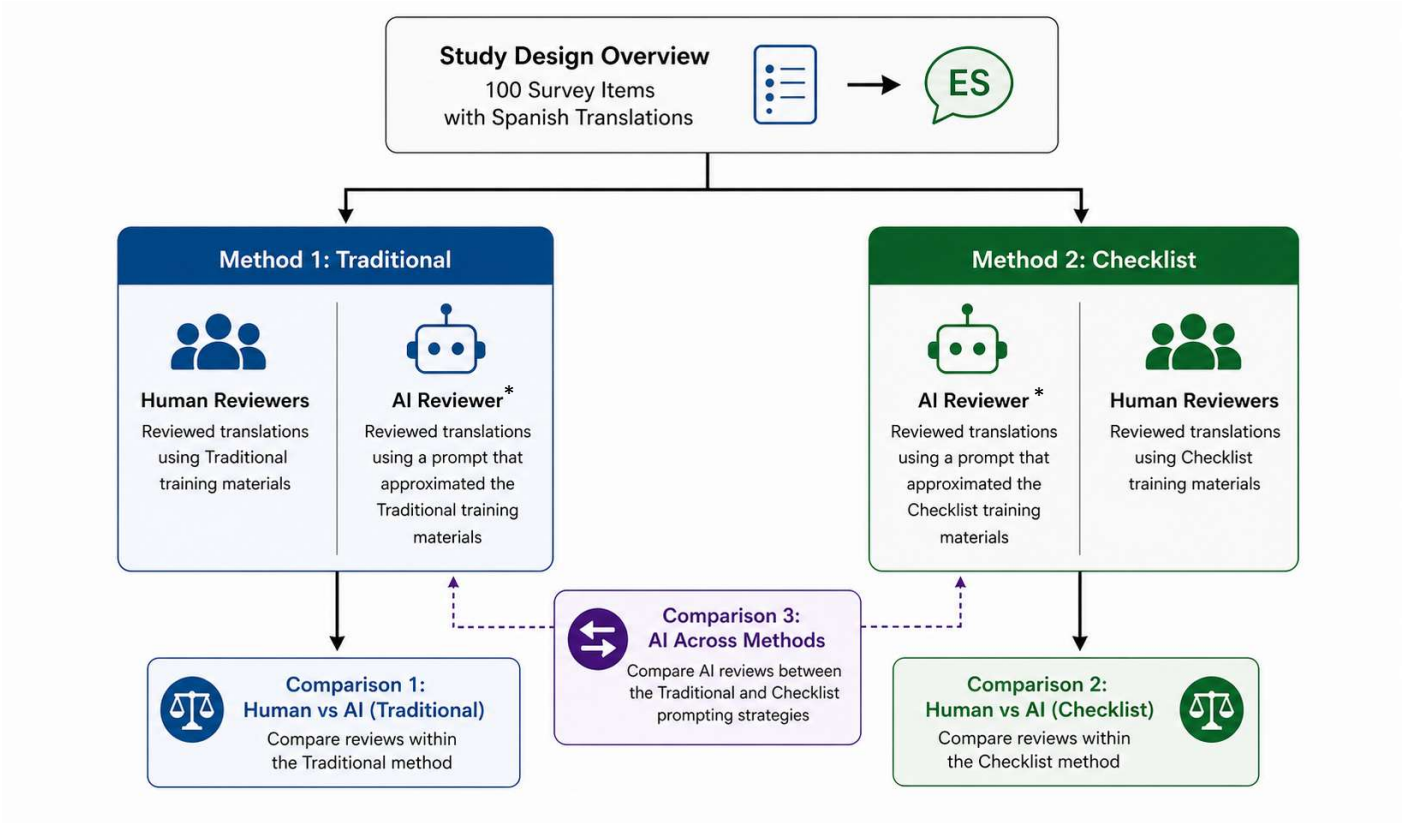
United States® Census Bureau

AI Review: Research Questions

1. Do comments made by AI reviewers match the comments made by human reviewers?
2. Do different prompts change the type of comments made by AI reviewers?
   - We tested prompts matching the Traditional review method compared to prompts matching the Checklist review method.

United States® Census Bureau

AI Research Design



**gpt-5-mini* used for all results in this presentation

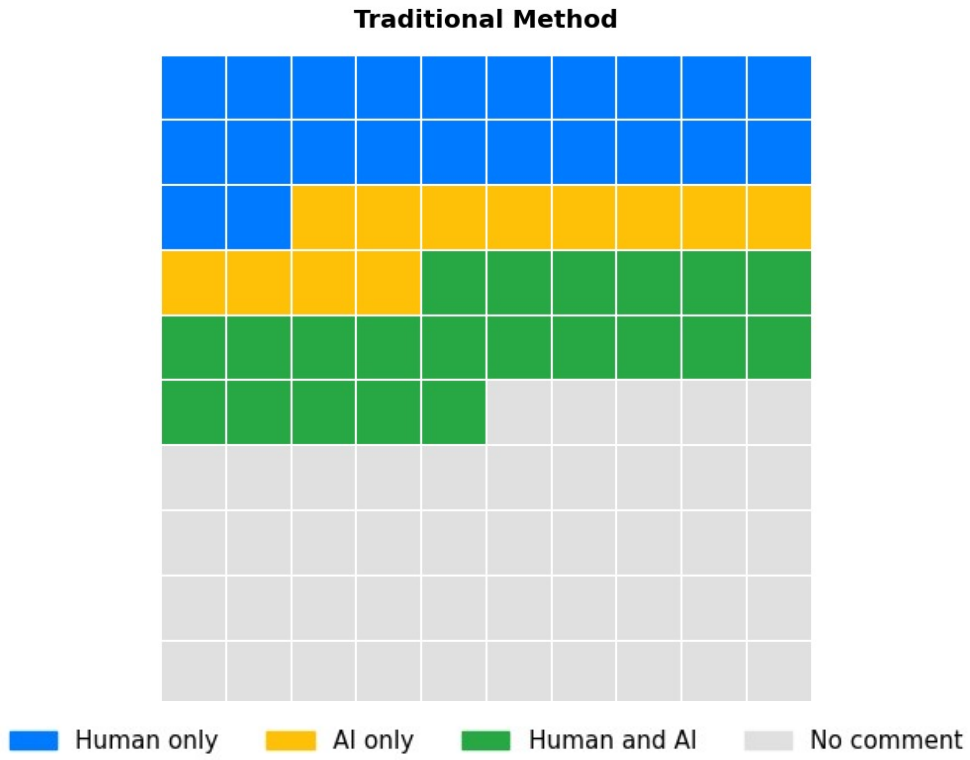
United States® Census Bureau

Prompting strategy

- Prompts designed to approximate the training materials used to train human reviewers
- Each prompt contained:
  - **Persona**: “You are a professional bilingual survey translation reviewer…”
  - **Examples**: Possible translation issues and relevant examples where appropriate, based on training materials
  - **Output constraints**: Structured CSV output containing only specific columns:
    - **item_id**: unique identifier
    - **comments**: feedback based on specified guidelines
    - **suggested_translation**: a revised translation based on any issues identified

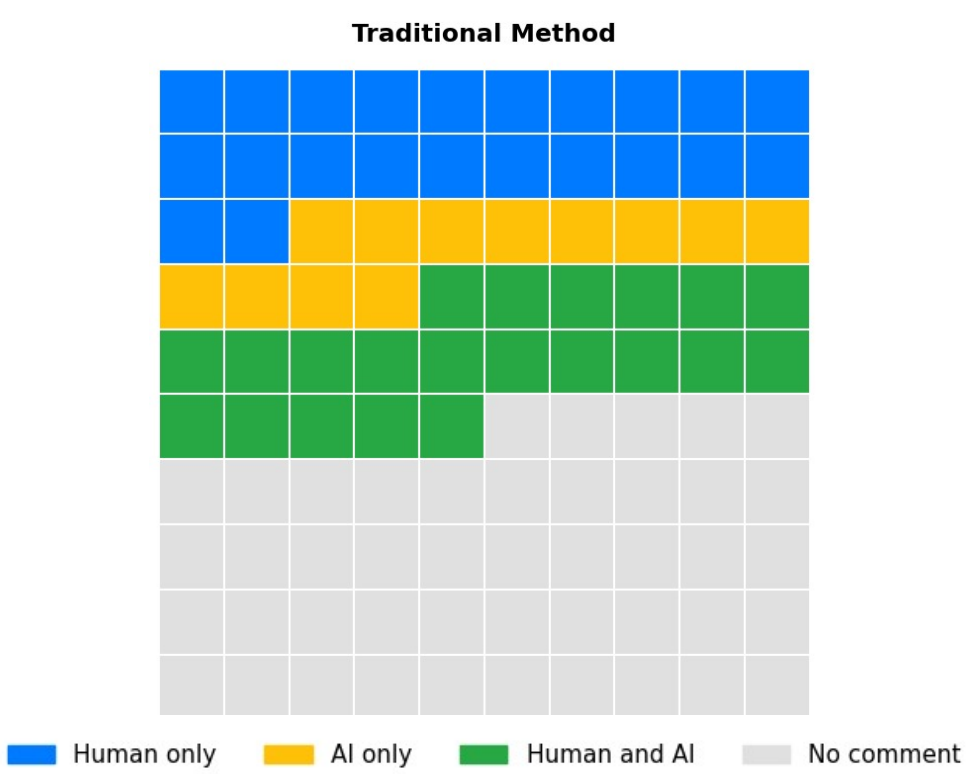
United States® Census Bureau

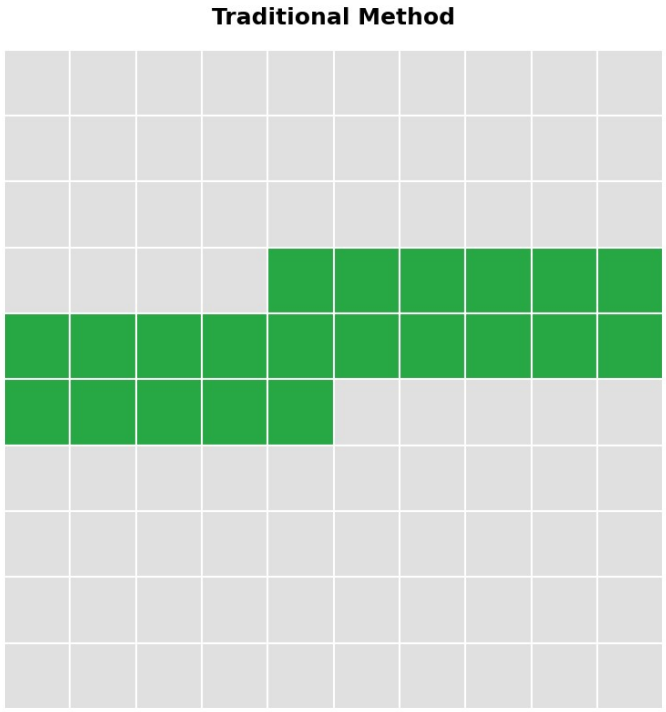
Results: Human vs. AI (Traditional)



United States® Census Bureau

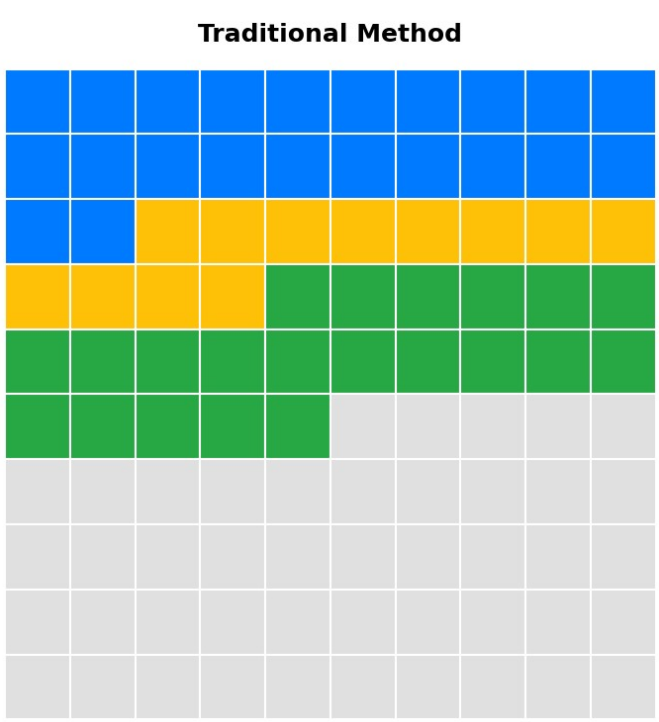
Results: Human vs. AI (Traditional)

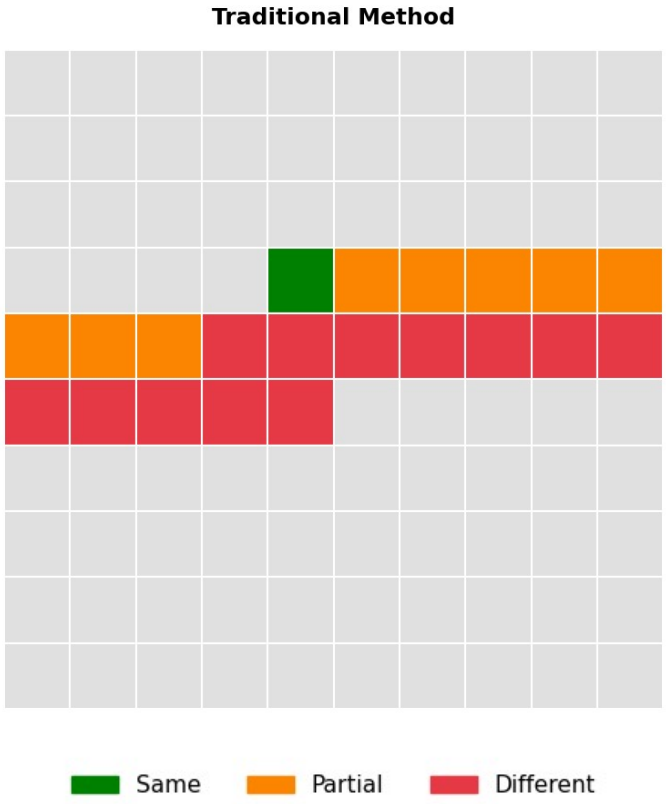




United States® Census Bureau

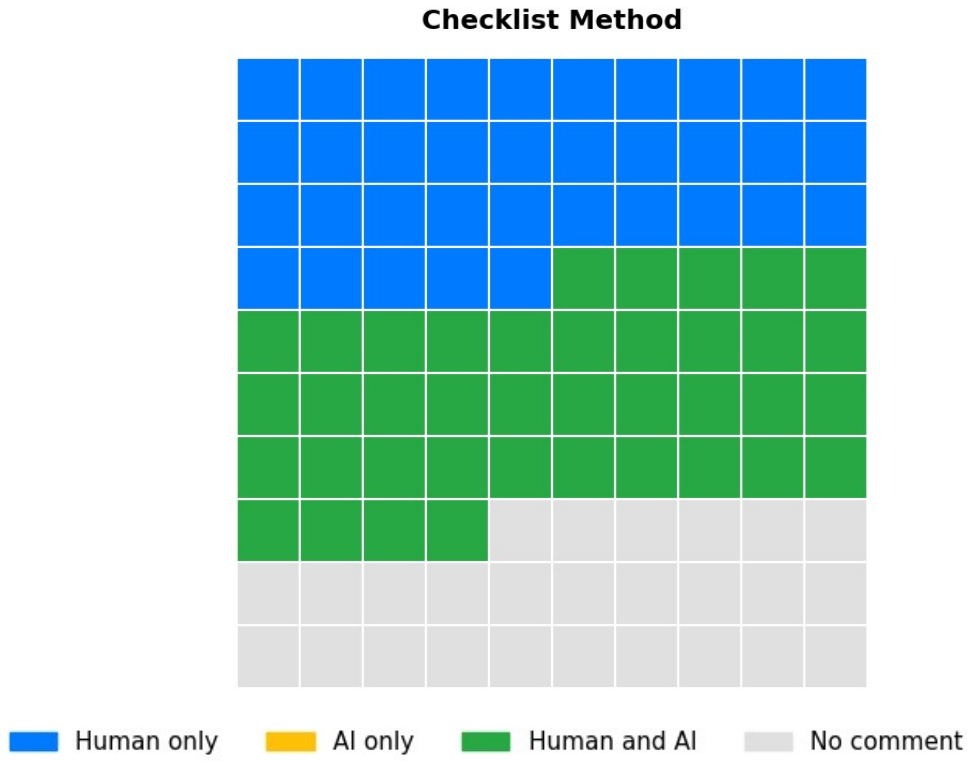
Results: Human vs. AI (Traditional)





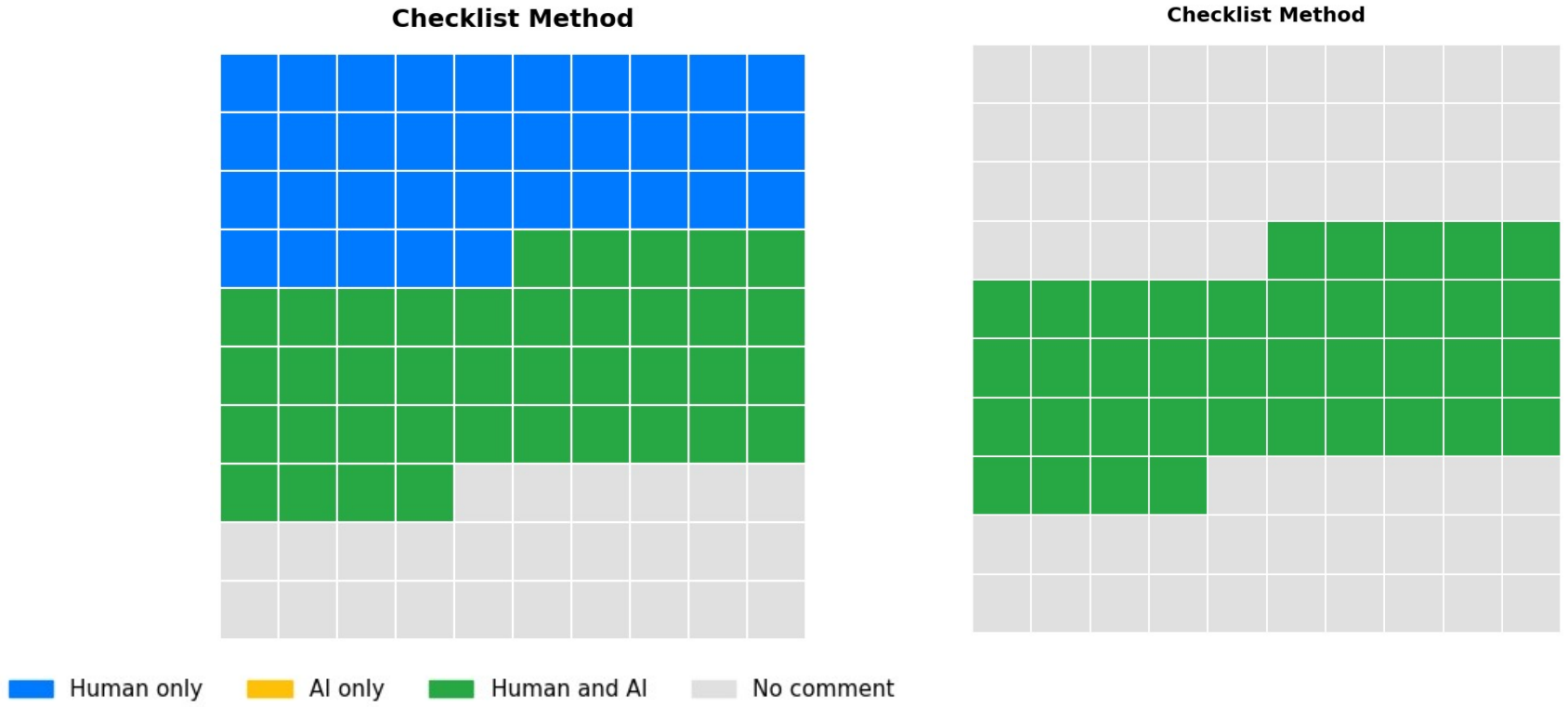
United States® Census Bureau

Results: Human vs. AI (Checklist)



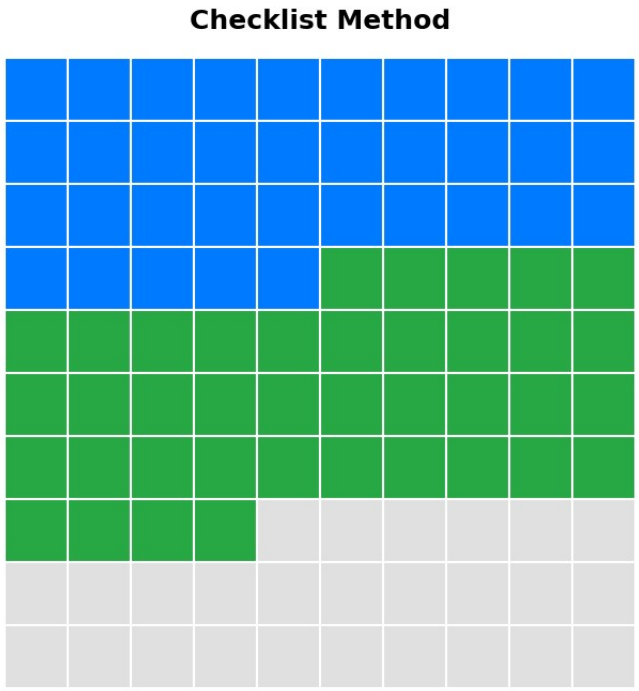
United States® Census Bureau

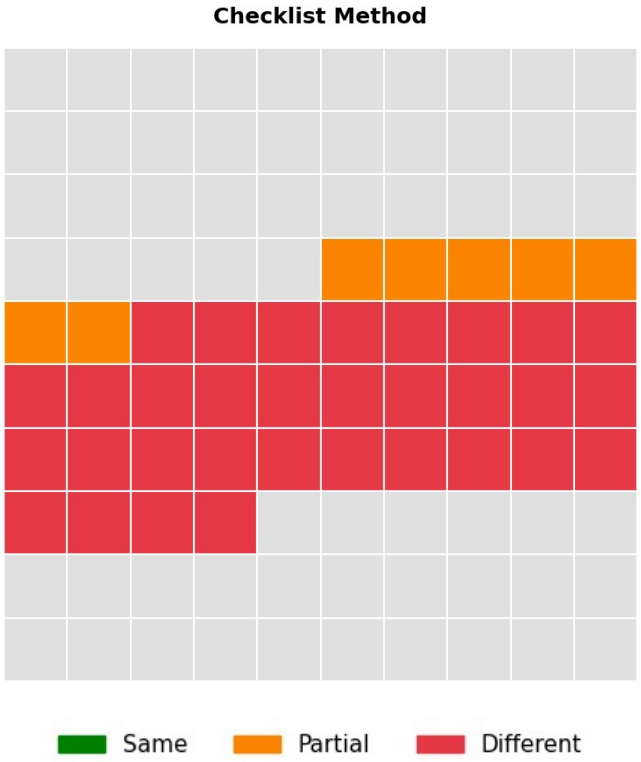
Results: Human vs. AI (Checklist)



United States® Census Bureau

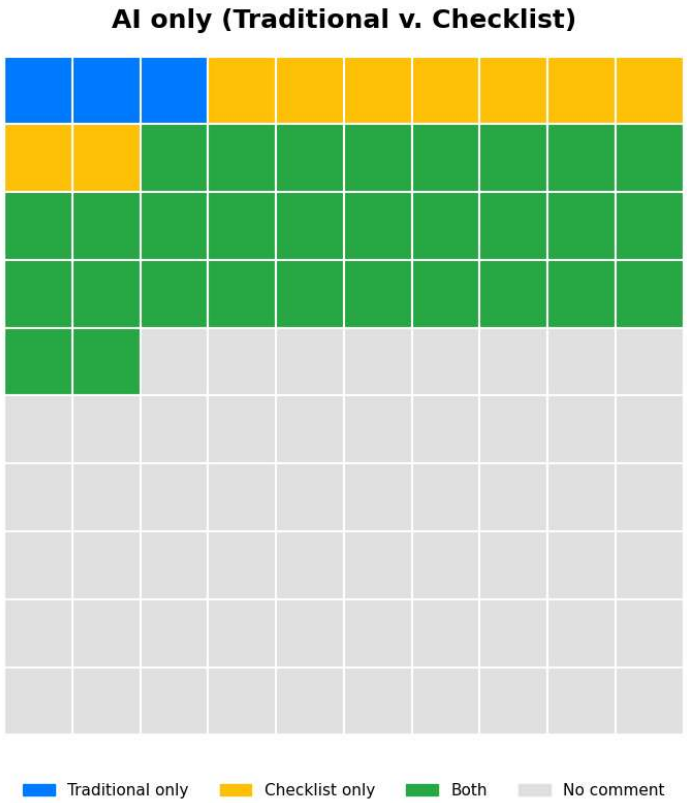
Results: Human vs. AI (Checklist)



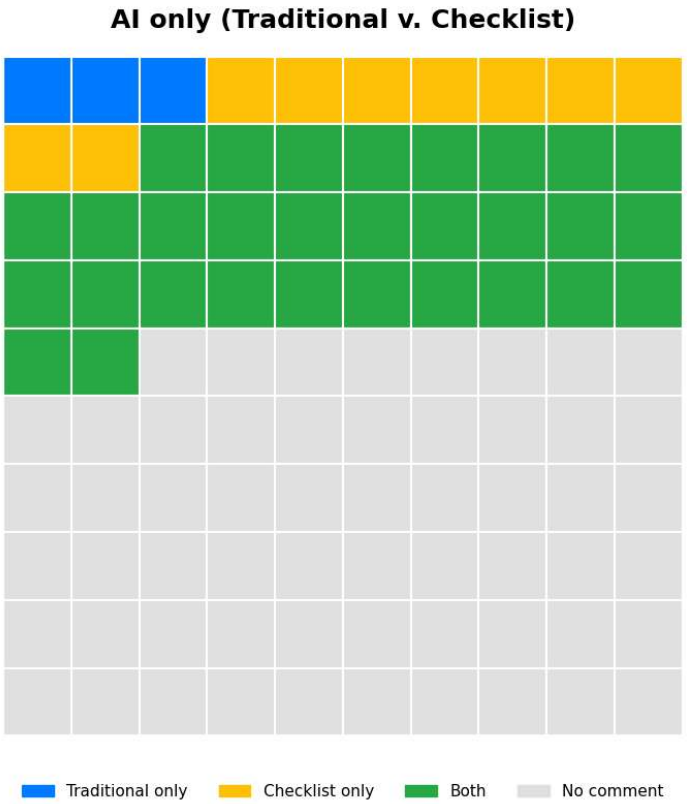


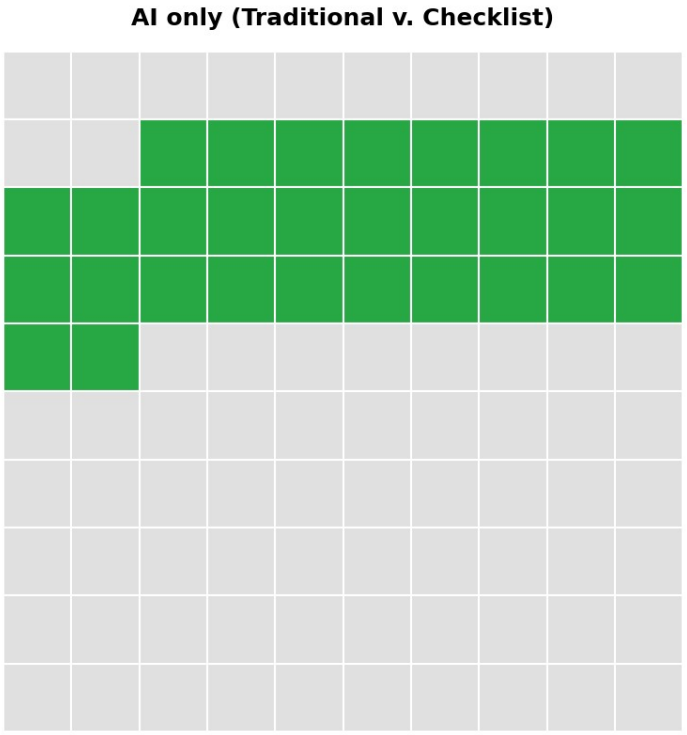
United States® Census Bureau

Results: AI Only (Traditional vs. Checklist)



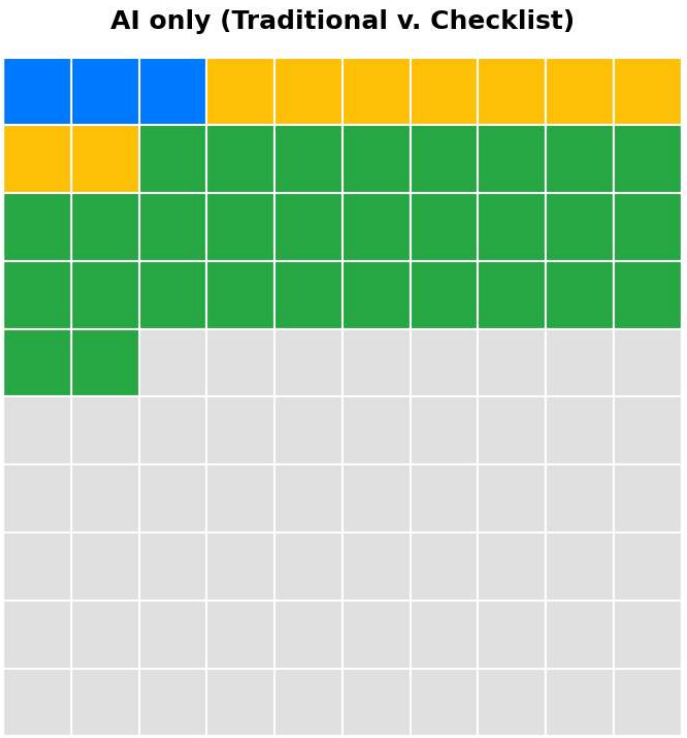
Results: AI Only (Traditional vs. Checklist)

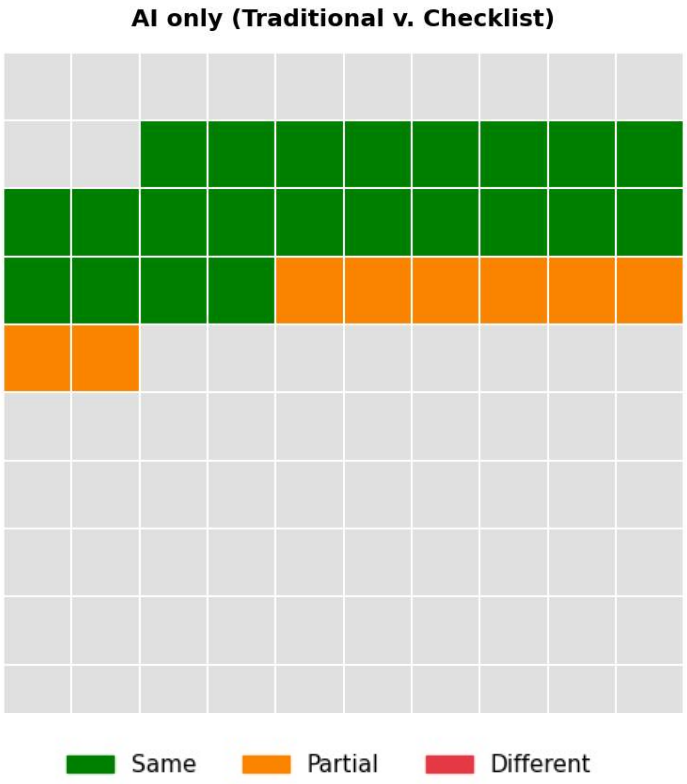




United States® Census Bureau

Results: Types of AI comments by Prompting Strategy





United States® Census Bureau

Overall Comparison of Human and AI Recommendations

- Human and AI made the same recommendation (n=1)
  - Both the human and AI reviewers noted a translation error (an incorrect translation for “LED lightbulbs”)
- Human and AI recommendations partially agreed (n=15)
  - Both the human and AI reviewers identified an issue with phrasing of the translation but provided different suggested translations
    - For example, both pointing out a missing phrase or error but recommending differing new wording
- Human and AI recommendations were different (n=44)
  - AI recommended different word choice, humans recommended modifying response options
    - For example, no overlap in noting missing phrases or recommending new translations
- Recommendation from only one reviewer (humans or AI) (n=69)
- Quirky AI recommendations: Translate interviewer instructions, use “common survey language”

United States® Census Bureau

Discussion

- Overall, human reviewers made more recommendations than AI reviewers regardless of review method
- Human reviewers and AI reviewers rarely made the same recommendations
- Recommendations by AI reviewers were largely the same regardless of prompting strategy

United States® Census Bureau

Next Steps

- Incorporate additional LLM models into the analysis
- Evaluate the quality of the LLM model recommendations
- Questions for future research:
  - How do final translations generated by AI reviewers perform compared to human translations in a fielded survey?
  - Would including an AI reviewer version prior to human consensus meetings be more useful?
  - Can Retrieval-Augmented Generation (RAG) improve the quality of AI recommendations and translations?

United States® Census Bureau

Can Generative AI Enhance Spanish Questionnaire Translation Review?: An Experimental Comparison of Expert Review Methods

Thank you!

For more information:

E-mail: Patricia.L.Goerman@census.gov

United States® Census Bureau