Presenter: Vinh Vo
Co-authors: Prof. Gang Chen
Dr. Maame Esi Woode

MONASH University

Test-retest reliability of case 1 best-worst scaling: evidence from relative attribute importance of preferences for AI use in healthcare

PCR162

1. Background

- Despite a number of existing DCE-related studies on test-retest reliability, the issue of repeatability and stability over time in best-worst scaling studies is important and under-researched.
- Best-worst scaling (BWS) has been applied across a wide range of disciplines, including health, agriculture, business, environmental studies, linguistics, and transportation (Schuster et al., 2024).
- Evidence on test-retest reliability using BWS is little known (Gandhi et al., 2025; Xiong et al., 2023).

2. Aims

1. Investigate the relative attribute of importance (RAI) of AI in two survey waves in a case 1 BWS study
2. Examine test-retest reliability of case 1 BWS study

3. Methodology

1 Study setting and participants

- Two-wave survey
- Participants living in Australia > 1 year, >18 yo, variables include age, gender, education level, household income, use of AI apps, health status

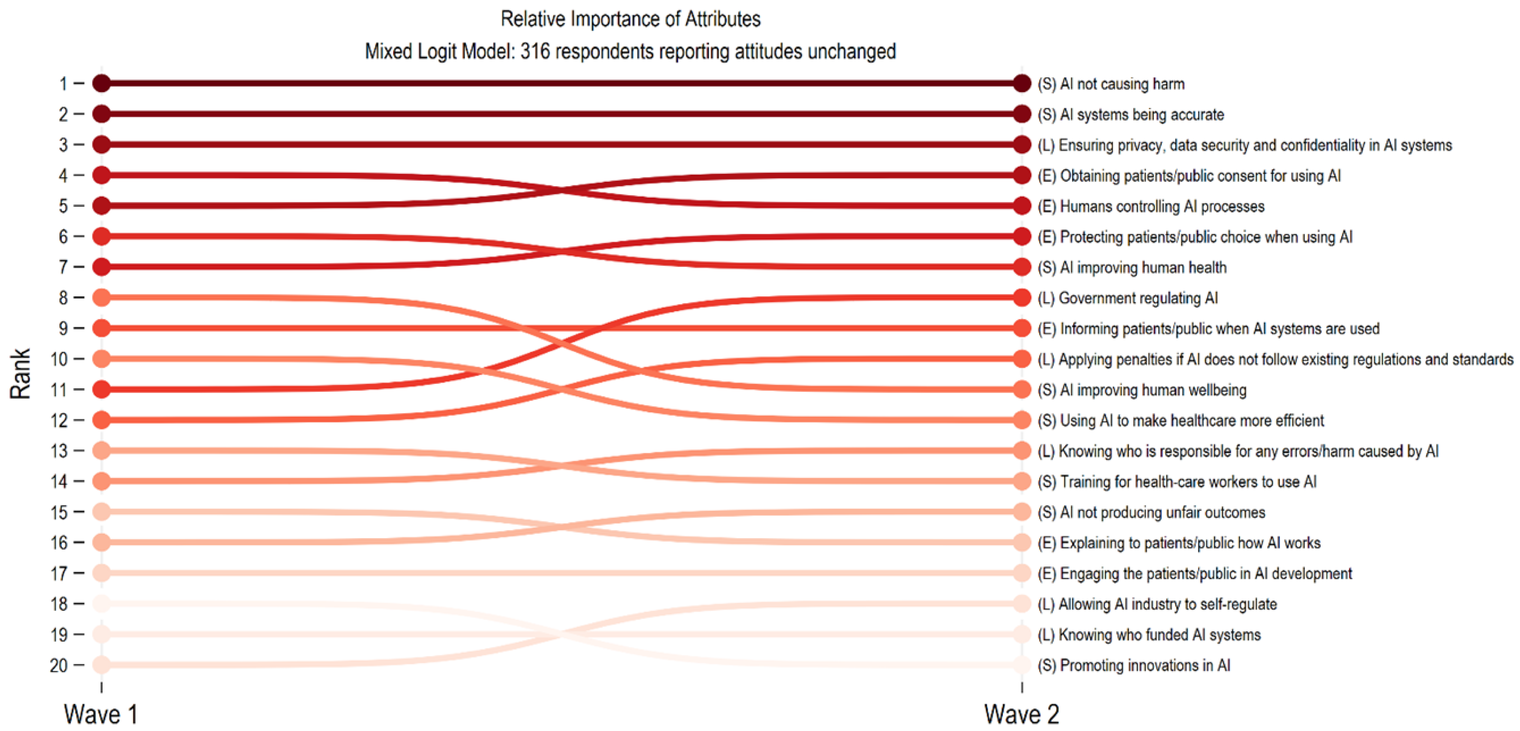
2 Case 1 BWS

- Attribute identification: 20 attributes into Social, Legal, and Ethical dimensions in using healthcare AI
- Experimental design: Balanced Incomplete Block Design
- 1st wave administered in Mar 2023 and 2nd wave in Sep 2023

3 Analysis

- Panel-data mixed logit for estimating RAI
- Test-retest reliability using two-way mixed effects intra-class correlation coefficient (ICC)
- Subsample analyses for gender and income
- Scenario ICC analyses: (i) same respondents in two waves; (ii) excluding those who reported using AI in two waves

4. Findings



Among the domains, attributes in the **Social (S) domain exhibited the greatest variability**, with 75% changing in rank between waves. This was **followed by the Ethical (E) domain**, where 67% of attributes shifted, and **the Legal (L) domain, with 50% showing change indicating relatively more stable compared with those in the Social and Ethical domains.**

Males vs females: Both males and females maintained stable priorities around core ethical principles - safety, accuracy, privacy, and consent - indicating shared concern for trustworthy and responsible AI in healthcare. Women's preferences reflected a more cautious and governance-oriented stance, whereas men's priorities leaned toward performance and system effectiveness.

High-income vs low-income groups: High-income respondents placed greater emphasis on individual autonomy and formal regulation, while low-income respondents focused more on accountability and protection.

The ICC for the full sample of 426 participants across both waves was **0.81** ($p < 0.01$), indicating strong test-retest reliability.
For those who reported having used AI in both waves, there are 70 participants from this group. After excluding these individuals, 356 participants remained across both waves. The ICC for this subsample was **0.85** ($p < 0.01$), suggesting improved reliability when respondents with potential exposure effects were removed.

1. **The preference elicitation method exhibits robust stability;**
2. **External informational or technological exposures may modestly affect measured consistency, highlighting the importance of accounting for such factors in this longitudinal BWS.**

References

Schuster, A. L. R., Crossnohere, N. L., Campoamor, N. B., Hollin, I. L., & Bridges, J. F. P. (2024). The rise of best-worst scaling for prioritization: A transdisciplinary literature review. Journal of Choice Modelling, 50, 100466. https://doi.org/10.1016/j.jocm.2023.100466

Gandhi, M., Ang, F. J. L., Neo, S. H. S., Gonzalez, J. M., Cheung, Y. B., Finkelstein, E. A., Tan, S. M., Ai, I. T. E., Wong, J., Kanesvaran, R., Yee, A., & Ozdemir, S. (2025). Quality of Care for Patients With Advanced Illness Scale: Development, Preference Elicitation, and Evaluation of Measurement Properties. Value in Health, 28(9), 1417–1425. https://doi.org/10.1016/j.jval.2025.05.006

Xiong, X., Dalziel, K., Huang, L., & Rivero-Arias, O. (2023). Test-Retest Reliability of EQ-5D-Y-3L Best-Worst Scaling Choices of Adolescents and Adults. Value in Health, 26(1), 50–54. https://doi.org/10.1016/j.jval.2022.07.007