# A Brief Translation Quality Measure for Patient-reported Outcome Measures Using Machine Translations

Chris Gibbons[1], Sheng-Chieh Lu[2]

[1]Department of Symptom Research, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA
[2]Department of Cardiology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA

THE UNIVERSITY OF TEXAS MD Anderson Cancer Center
Making Cancer History®

## Results

### Psychometric properties

- 735 participants enrolled, with a mean age 40.5±15.9; 62.9% female; 46.4% white.
- The Mokken AISP unidimensionality analysis suggests a scalability of 0.86 after removing items 4 and 6.
- Confirmatory factor analysis of the four items loading onto one dimension yielded Root Mean Square Error of Approximation (RMSEA) = 0.01 with $p < 0.05$.

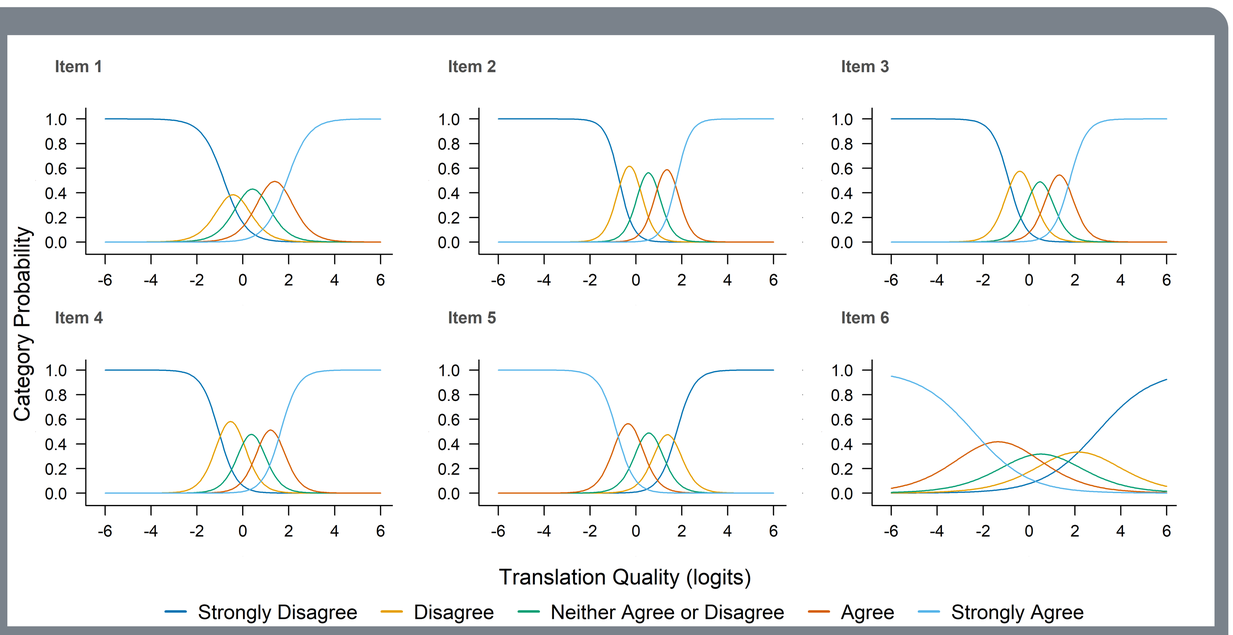


Figure 1. The response category probability curve.

Table 1. Psychometric parameters of translation quality measure (TQM) for graded response model (GRM) scoring

| # | Items | Discrimination (a) | Threshold 1 (b1) | Threshold 2 (b2) | Threshold 3 (b3) | Threshold 4 (b4) |
|---|---|---|---|---|---|---|
| 1 | The translation is difficult to read and understand | 2.10 | -0.83 | -0.05 | 0.86 | 1.91 |
| 2 | The translation contains unnecessary elements | 3.64 | -0.72 | 0.14 | 0.92 | 1.75 |
| 3 | The translation does not feel authentic. | 2.75 | -0.89 | 0.07 | 0.87 | 1.79 |
| 4 | Parts of the original text are missing or changed in the translated version* | - | - | - | - | - |
| 5 | The meaning of the translation and the original is different | 2.67 | -1.10 | -0.05 | 0.81 | 1.76 |
| 6 | The translation of all questions is perfect* | - | - | - | - | - |

Note: Items 4 and 6 were not selected by the Mokken item selection procedure, thus not included in the graded response model analysis.

## Key Findings

**❖ Tools evaluating LLM translation quality from multiple aspects, such as readability, semantic equivalence, and cultural adaptation, are needed.**

**❖ We developed a machine translation assessment tool (TQM) consisting of 4 5-point Likert-type scale items. The TQM demonstrated good psychometric properties.**

**❖ There is a potential to use the TQM in combination with metrics for LLM translation quality assessments for PROMS and other healthcare-related materials.**

## Background

- The rise in large language models (LLMs) enables opportunities to enhance patient-centered care. One might be the ability to rapidly and inexpensively create accurate translations of English language patient-reported outcome measures (PROMs) to facilitate global uptake [1].
- Automatic metrics, such as METEOR, have been developed and widely utilized for machine translation quality assessments [2].
- Tools are lacking that evaluate machine translations from various aspects, in addition to objective scores of word matching.

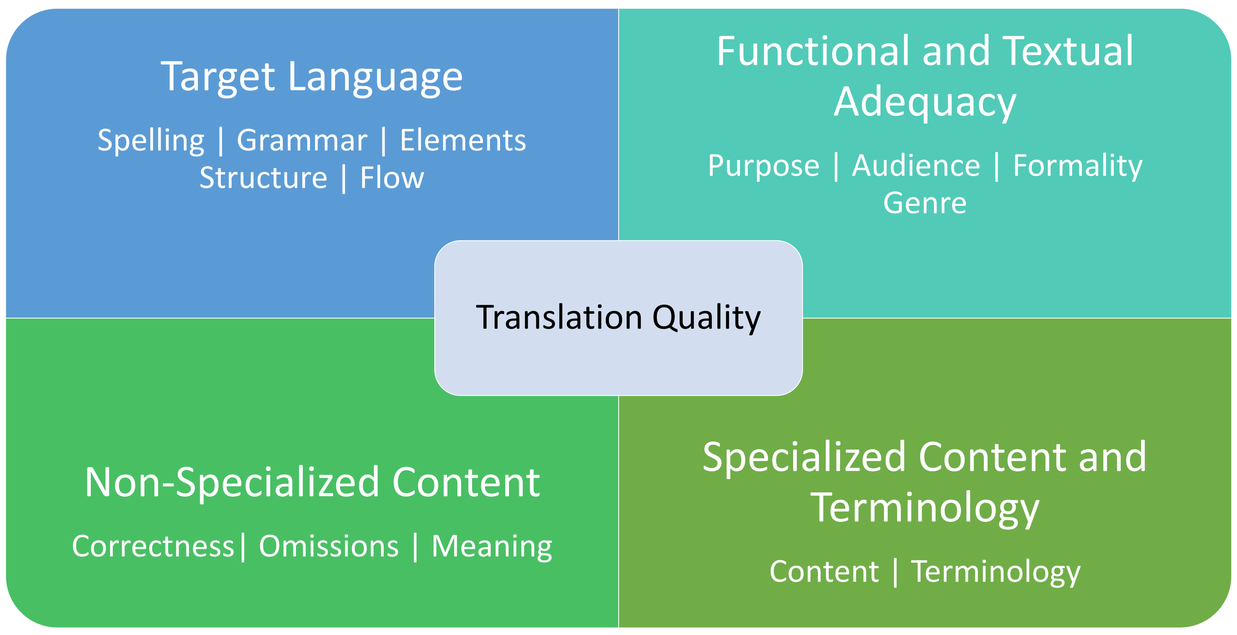


Figure 2. Conceptual components for translation quality evaluation

## Objective

To develop and evaluate the Translation Quality Measure (TQM) for assessing machine translation quality for PROMs from the aspects of readability, completeness, cultural adaptation, and semantic equivalence.

## Methods

- We developed the nascent questionnaire based on a tool originally developed for expert linguist use [3]. Figure 2 demonstrates the conceptual framework of the tool.
- We modified the tools by adding a 5-point Likert scale, ensuring that each question only contained a single clause to reduce variation in interpretation. We focused TQM on the target language and non-specialized content domains.
- We enrolled participants on the Prolific platform to collect data for the psychometric evaluation. Each participant evaluated the translation quality of Breast Q or Face Q by different translators (GPT-4, Google Translate, and Human experts) using TQM.
- We examined the psychometric properties of TQM using a multi-step psychometric approach involving the Mokken automated item selection procedure (AISP; Hi < .30 criteria for item removal), Item Response Theory (IRT), and confirmatory factor analysis (CFA).
- The institutional review board of the MD Anderson Cancer Center approved this study

## Conclusions

- The TQM can be applied to the evalution of machine translation for PROMs from patient perspectives, in addition to semantic equivalence evaluation.
- There is potential to apply the instrument to evaluate the quality of machine translation for other healthcare-related instruments.

## References

1. Chan, V., Tang, W.KW. GPT for Translation: A Systematic Literature Review. SN COMPUT. SCI. 5, 986 (2024).
2. Lu, SC., Xu, C., Kaur, M. et al. Can machine translation match human expertise? Quantifying the performance of large language models in the translation of patient-reported outcome measures (PROMs). J Patient Rep Outcomes 9, 94 (2025).
3. Colina S. Translation quality evaluation: Empirical evidence for a functionalist approach. Translator. 2008;14(1):97-134.