How We Measure Quality in Interpreting

Critical Comparison of Quality Evaluation Models and the Case for a New Framework

Centre for Translation Studies
UNIVERSITY OF SURREY

PHD researcher Hope Hao: qh00195@surrey.ac.uk
Supervisory Team: Dr Elena Davitti & Prof Constantin Orasan

Background

Quality has been studied since the 1950s, traditionally by rating a whole performance against fixed criteria. This is structured but subjective, and holds up poorly across large datasets and many language pairs (Davitti et al., 2025). Newer methods instead pinpoint individual errors and rate each by severity, giving more consistent, reproducible results. Such error-based models exist for respeaking (NTR; Romero-Fresco & Pöchhacker, 2017), subtitling (FAR; Pedersen, 2017) and translation (MQM; Mariana, 2014) but none for interpreting. This project adapts them to build one.

Research Question

How can existing quality evaluation models be adapted to create a systematic framework for measuring interpreting quality?
SQ1. Which components of existing error-based models (MQM, FAR, NTR) transfer to interpreting, and which do not?
SQ2. How can interpreted output be segmented into comparable, countable units for scoring?
SQ3. How can the framework account for both severe (critical-failure) errors and effective renditions, and apply consistently across interpreters, modes and settings?

Methodology

Review existing quality models and evaluation studies. Map how they connect, where they overlap, and where gaps remain for interpreting.

Apply existing metrics to pilot legal interpreting data; identify where they break down; build a revised taxonomy for interpreting.

Validate revised taxonomy on larger datasets across different domains

Data:
academic literature on quality assessment

Current Pilot data:
Court interpreting recording video
Duration: 30 min total
Participants: Two Professional interpreters, English → Chinese
Modes: Both simultaneous and consecutive

Future dataset (planned)
Legal and healthcare interpreting recording videos
Participants: both professional and student interpreters, English → Chinese
Modes: Both simultaneous and consecutive

Expected Outcomes

Segmentation method: a consistent way to divide an interpreted output transcript into countable units. In interpreting, no agreed basis for this segmentation yet exists, which is a precondition this project must address.

Critical-failure threshold:
Identifying errors severe enough to render an interpretation unacceptable, not just score it down.

Positive scoring:
Not only penalising errors but also crediting positive scoring for effective renditions.

Greater consistency:
Quantitative scores comparable across modes, and across settings, reducing the subjectivity and scalability limits.

References:

Davitti, E., Korybski, T. and Braun, S. eds., 2025. The Routledge handbook of interpreting, technology and AI. Taylor & Francis.

Mariana, V.R., 2014. The Multidimensional Quality Metric (MQM) framework: A new framework for translation quality assessment. Brigham Young University.

Pedersen, J., 2017. The FAR model: assessing quality in interlingual subtitling. The Journal of Specialised Translation, (28), pp.210-229. Romero-Fresco, P. and Pöchhacker, F., 2017. Quality assessment in interlingual live subtitling: The NTR Model. Linguistica Antverpiensia, New Series–Themes in Translation Studies, 16.