What makes listening comprehension difficult?: A feature-based machine learning approach to understanding item difficulty

Xun Yan, Huiying Cai, Ping-Lin Chuang, Yulin Pan

University of Illinois at Urbana-Champaign

Abstract

Item difficulty in L2 listening assessment can be affected by textual and acoustic features of both listening inputs and items. Traditional statistical approaches use regression models to predict difficulty based on various features. In contrast, machine learning (ML) allows for more generalizable models with consistent predictive performances across datasets, offering a broader understanding of item difficulty. This study builds a feature-based ML model, incorporating textual, acoustic, and extra-linguistic features, to predict difficulty in 225 multiple-choice listening items from Taiwan's General English Proficiency Test. We extracted 950 textual (i.e., lexical/syntactic complexity and textual similarity indices) and acoustic features (i.e., pronunciation and fluency) at the option, stem, and stimulus levels. Because GEPT features different item types, we used two approaches to select features: item-type generic features and item-type specific features. For each feature type, we further reduced data dimension through either manual removal of redundant features or principal

component analysis. These two steps yielded four different feature sets. We subjected each feature set to mixed-effects ridge regression models along with extra-linguistics features (i.e., test focus and item type) and compared their performances. The best-performing model employed 27 item-type generic raw features after manual removal of redundant features (R^2 = 0.86). Results indicated meaningful relationships between item difficulty and lexical/syntactic complexity, similarities among options, stem and stimuli, pronunciation, test focus, and item type. The findings highlight how these features influence item keys and distractors, offering insights into item difficulty modeling and distractor writing. This study underscores the effectiveness of integrating computational linguistics and ML in L2 listening assessment research.

***Keywords*:** Listening assessment; Item difficulty; Machine learning; Construct validity