Keynote Lecture

Applications of Linguistic and Psychometric Methods in AI-driven Language Assessments

Authors: Geoffrey T. LaFlair
Affiliation: Duolingo English Test

Outline

Recent advances in AI have transformed digital-first language assessment development and administration (LADA). However, the development process of modern, large-scale, high-stakes language assessments requires more than "just AI". The development process should be rooted in an interdisciplinary, human-centered approach that leverages AI as a tool, the outputs of which are evaluated. In this presentation, I will discuss applications of empirical linguistic and psychometric methods to three key challenges in large-scale language assessment: content generation, item parameter estimation, and automated scoring.

Content generation (stimuli and items) is often a bottleneck for large-scale LADA systems, which require very large item banks. Large item banks are especially important for assessments that adapt the difficulty of questions to test taker ability in order to ensure no two tests are the same and that the test can measure a wide range of abilities. Artificial intelligence can generate content en masse. The challenge for parameter estimation (e.g., item difficulty) is that traditional methods require extensive pilot testing, which is time-consuming, costly, and poses security risks. Artificial intelligence coupled with psychometric methods, or computational psychometrics (von Davier, 2022), offers solutions for incorporating features of the stimuli in addition to response data to create high-quality and stable "synthetic" parameter estimates. Automated scoring systems have a long history of operational use in language assessment (e.g., e-rater, Attali & Burstein, 2006). However, some subconstructs of writing and speaking have been difficult to capture using traditional approaches to feature engineering. Artificial intelligence now offers opportunities to better measure these subconstructs, such as cohesion (Naismith et al., 2023) and intelligibility (Cai et al., 2025).

However, with these opportunities comes the responsibility to ensure the systems are working as intended. One of the primary assumptions of AI-generated content is the ability to automatically generate text that is linguistically appropriate for the purposes of measurement. I will discuss how register-based corpus linguistics (Biber & Conrad, 2019) offers approaches for shaping prompts, or test specifications, and for evaluating the linguistic features of the generated content. An assumption of synthetic parameter estimates is that they are stable, accurate, and capture difficulty in a way that is construct-relevant. I will show how synthetic parameter estimates can be evaluated empirically through natural language processing, corpus-linguistic and psychometric methods, such as examining feature importance as well as their effects on the reliability of test scores. Additionally, I will demonstrate how salient quantifiable aspects of spoken language, such as intelligibility, can be included as features in automated scoring systems in order to improve construct coverage. Through the lens of responsible AI (Burstein, 2026), I will present systems and methods (built on linguistics and psychometrics) for ensuring that test scores from digital-first assessments remain valid, reliable, and fair.