EALTA www.ealta.eu.org | European Association for Language Testing and Assessment

Equality, Diversity and Inclusion (EDI) and Artificial Intelligence (AI) Joint SIG Meeting.

Salzburg, Austria.

Wednesday 28th May, 2025

2.00-5.00pm

(25 minutes for each presentation, including questions.)

2.00 *Introductions and welcome to the SIG.*

2.10 *Review of EDI policies and procedures for the AES of a French writing exam.*
Vincent Folny, France Education International.

2.35 *AI and item writing for fair testing: insights from text and multiple-choice question generation.*
Philip Horne, Trinity College London.

3.00 *Artificial Intelligence in Language Teaching, Learning, and Assessment: Insights from ALTE's CEFR and TECHNOLOGIES SIGs*
Letizia Cinganotto, University for Foreigners of Perugia.
Vincent Folny, France Education International.
Joaquin Manuel Cruz Trapero, University of Jaen.

3.25 **Break**

3.45 *Bias and Fairness in Automatic Spoken Language Assessment.*
Kate Knill, University of Cambridge.
Mengjie Qian, University of Cambridge.

4.10 Panel discussion
AI and the future of accessibility and inclusion in assessment: promised land or dystopian nightmare?
Vicky Allan, Berlin School of Economics and Law.
Judit Kormos, Lancaster University.
Okim Kang, Northern Arizona University.
Ari Huhta, University of Jyväskylä.

AOB - issues of interest to participants.

5.00 **Close**

Review of EDI policies and procedures for the AES of a French writing exam

Vincent Folny, France Education International.

The Test de Connaissance du Français (TCF) is a French language test used mainly for academic and immigration purposes. In 2019, France Education international (FEI) launched the development of an automated scoring solution for the digital writing using CamemBERT, a bidirectional deep learning model, combined with linguistic variables to ensure explainability.

The project / Solution, called "FIDELIA", prioritised equality, diversity, and inclusion (EDI) considerations from its inception, in line with the French legal concepts of equal treatment and non-discrimination. The development process followed rigorous statistical, psychometric, and scientific methodologies, guided by the Williamson et al.'s (2012) framework. FEI decided to develop this solution in collaboration with the CENTAL laboratory (UC Louvain). This laboratory specialises in NLP and French (L2).

The main steps of the project were:

1. Corpus development: Creation of training, validation, and evaluation datasets (27,000 calibrated tasks out of a total of 450,000).
2. Gold Corpus: 961 tasks rated by 55 raters and calibrated using a multifaceted Rasch model.
3. Selection of the “best” model out of 24 models.
4. Fairness analysis: assessment of potential biases (Rasch analysis and Loukina (2019) framework).
5. Explanation of the rules followed for the systematic independent double human/ machine rating.
6. Quality monitoring: ongoing assessment of the system's performance since implementation.

During the presentation we will review each step and explain how EDI has been considered in our procedures and what our concerns or questions were. The presentation will conclude by discussing the implications of human versus machine approaches to EDI and questioning the reduction of EDI to statistical thresholds or pure argumentative approaches.

AI and item writing for fair testing: insights from text and multiple-choice question generation

Philip Horne, Trinity College London.

We now live in an age of myriad opportunities to harness the power of artificial intelligence (AI), especially large language models (LLMs), in a range of educational contexts. From a testing perspective this has given item writers opportunities to develop secure and robust test content (including text passages and comprehension questions for reading and listening tasks) using AI content generators. In this session, we will present data from one Awarding Organisation based upon item acceptance rates. The statistics highlight the benefits of the model, as well bring important limitations and potential pitfalls to the fore. Rejected items were coded for their issues, which include superficiality and insufficient research. These points are especially important since they frequently result in the concomitant issue of poor representation and a lack of diversity in listening and reading texts. AI generated texts may, for instance, adopt more western-centric worldviews or rely upon insufficient language models to generate passages representative of global majority contexts or historically marginalised voices. We argue (a) that the role of the human item writer remains integral to ensure fair and ethical use of AI platforms and (b) that appropriate training is essential to ensure that test development adheres both to technical specifications as well as fairness and bias guidelines. There is an indubitable need for prompting and probing questions between the item writer and the mode. We conclude by examining specific AI generated samples to consider how to identify more insidious issues (such as the aforementioned superficiality) and how to frame feedback to best suit the 21st century needs of test developers for technological innovation and global English perspectives.

Artificial Intelligence in Language Teaching, Learning, and Assessment: Insights from ALTE's CEFR and TECHNOLOGIES SIGs

Letizia Cinganotto, University for Foreigners of Perugia.
Vincent Folny, France Education International.
Joaquin Manuel Cruz Trapero, University of Jaen.

The intersection of the ALTE CEFR Special Interest Group (SIG) and the TECHNOLOGIES SIG lies in their shared commitment to investigating and discussing real-world applications of the Common European Framework of Reference for Languages (CEFR). In particular, their focus is on the Companion Volume to the CEFR (CEFRCV, 2020) and its implications for assessment practices, which are increasingly being enhanced by digital technologies and, more recently, by Artificial Intelligence (AI). Given the central role of the CEFR CV in shaping teaching, learning, and assessment, as well as the growing need for alignment with the learning-oriented assessment framework, the integration of AI and other digital tools holds significant potential for improving language evaluation and pedagogical effectiveness.

Recognizing the transformative impact of AI on language education, the two ALTE SIGs have recently collaborated to explore its implications, particularly in relation to assessment, feedback, and personalized learning. Their joint efforts aim to examine how AI-driven solutions can enhance language assessment methodologies, optimize learning pathways, and refine feedback mechanisms, all while maintaining strict adherence to CEFR principles. This exploration is crucial in ensuring that emerging technologies are implemented in a way that supports fairness, validity, and ethical considerations in language assessment.

To further investigate these developments, a survey among all ALTE members has been delivered to gather insights on perceptions, experiences, and best practices concerning the use of AI in language assessment. The data collected provides a comprehensive overview of current trends, revealing both cautious optimism and enthusiasm regarding the advancements AI has introduced into the field. While respondents acknowledge the potential of AI to enhance assessment processes, they also highlight the need for careful implementation to ensure transparency, reliability, and alignment with established CEFR-based assessment frameworks.

The findings from this survey will be presented and analyzed during the session, offering valuable perspectives on the evolving role of AI in language testing and assessment. By reflecting on the opportunities and challenges posed by AI, the CEFR and TECHNOLOGIES SIGs seek to gather evidence and critically assess its impact on language assessment practices. Their ultimate goal is to ensure that CEFR-aligned approaches remain at the forefront of innovation, promoting inclusivity, adaptability, and responsiveness to the dynamic landscape of language education and assessment.

Bias and Fairness in Automatic Spoken Language Assessment

Kate Knill, University of Cambridge.
Mengjie Qian, University of Cambridge.

Increasing the automating of spoken language assessment, whether for full or hybrid auto-marking, has many benefits. Significant improvements in auto-marking performance have been made possible with the rise of AI-powered systems. Such systems, however, may be biased against or towards specific groups of users as they rely on data for training, which may be imbalanced and/or reflect human bias's in annotation. An automatic spoken language assessment (SLA) system should focus solely on a candidate's language proficiency, without being influenced by factors such as their first language (L1), gender, or age. This work introduces a novel and interpretable approach to bias detection in SLA AI-based models using Concept Activation Vectors (CAVs). CAVs provide an efficient and scalable way to assess whether machine learning (ML) models are sensitive to specific concepts such as a candidate’s L1. By analysing model activations, we can identify specific cases where non-linguistic characteristics may be inadvertently influencing scores. This allows a targeted collection of appropriate data to validate this hypothesis, limiting the data and time needed to check for bias. We apply this approach to various SLA models, including handcrafted feature-based models, BERT and wav2vec based models, and demonstrate how the model choice can affect the levels of bias. Our findings emphasise the need for fairness-aware AI development and highlight how CAV-based bias measurement can support the creation of more reliable and trustworthy AI-driven assessment models.