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

EALTA Joint Special Interest Group Meeting for the *Assessment of Writing* and *Academic Purposes*

Wednesday, 28th of May, 14.00-17.00

Pädagogische Hochschule Salzburg Stephan Zweig

PROGRAMME

14.00-14.15 *Sonja Zimmermann and Olivia Rütti-Joy*
Welcome and introduction

14.15-14.45 *Aylin Ünaldı*
Unpacking Cognitive Strategies in Academic Reading: Sentence, Whole-Text, and Multiple-Text Level Assessments

14.45-15.15 *Lesya Lymar, Tetyana Kolomiiets*
Impact of Stressful Conditions on Testing Outcomes: Insights from English Language Examinations

15.15-15.45 COFFEE BREAK

15.45-16.15 *Atta Gebril, Khaled ElEbyary, Ramy Shabara*
AI as a Muse and Mentor: Enhancing L2 Writing Through Reformulation and Model Texts

16.15-16.45 *Alina Reid*
Experts at the core: integrating the comparative judgement technique in a mixed-method rating scale development process

16.45-17.00 *Sonja Zimmermann & Olivia Rütti-Joy*
Synthesising plenary discussion, future SIG activities and wrap-up

Attendance Registration

Please register for the this event via the EALTA conference registration page. Should you have any questions please do not hesitate to get in touch with the SIG coordinators Sonja Zimmermann zimmermann@gast.de and Olivia Rütti-Joy olivia.ruetti-joy@unifr.ch.

We look forward to welcoming you in Salzburg!

Abstracts

Aylin Ünaldı

Unpacking Cognitive Strategies in Academic Reading: Sentence, Whole-Text, and Multiple-Text Level Assessments

Academic reading skills are widely theorised and researched as they are core to the development of academic literacy and the assessment of it. Research in recent decades has shown interest in the assessment of higher-level reading skills, such as reading at the text level to construct the macro-structure of the text (text model) and reading across multiple texts to establish intertextual connections (documents model). Building on Khalifa and Weir's (2009) socio-cognitive framework for reading assessment, this study specifically examines three types of academic reading test item types in a comparative way —sentence-level, whole-text level, and multiple-text level items by synthesising three sources of data: test performance, simultaneous reporting of strategy use and retrospective think aloud. Three versions of the reading test consisting of sentence, whole-text and multiple-text item types were administered to three different groups of over 300 EAP students at A2-B1 proficiency levels in an English medium university in Turkey (i.e., Version 1: N=340, Version 2: N=347, Version 3: N=345). The students also responded to reading strategy use proforma. A group of approximately 10 students were interviewed giving us retrospective think aloud data for the verification of the processes identified for each item type. The results have shown that the level of comprehension and test difficulty can be largely affected specific features of the questions as well as item types. Text level macro-structure formation is a demanding, memory-laden task and contrary to expectations, certain multiple-text reading items can be completed through quite superficial reading strategies if they do not operationalise documents model formation effectively. This study has given us insight into how reading tasks can be designed to better reflect the multifaceted nature of academic literacy at different levels with implications for enhancing the validity of academic reading assessment.

Lesya Lymar, Tetyana Kolomiiets

Impact of Stressful Conditions on Testing Outcomes: Insights from English Language Examinations

Stressful conditions can significantly affect test performance and outcomes. Since February 24, 2022, Ukrainian citizens have been exposed to continuous psychological stress due to the war, which inevitably influences the learning process, including standardized assessments such as English language testing. Bogomolets National Medical University administers English proficiency tests, assessing the language competence of its faculty. This study aims to examine the impact of stress on testing performance.

In 2022, testing was suspended for six months due to the outbreak of war. When examinations resumed, 10 candidates completed the test and all achieved a "High Pass" grade. However, neither the exam itself nor the preceding night was disrupted by military attacks. In 2023, 18 candidates took the exam across five sessions. Each of these sessions, held on Saturdays, was

interrupted by air raid alerts, requiring examinees to relocate to shelters. Despite these interruptions, all candidates again achieved a "High Pass" grade. In 2024, 16 candidates participated in four exam sessions. Three of these sessions were interrupted by air alerts, and one took place following an exhausting night of military drone attacks. Of the 16 examinees, 12 obtained a "High Pass," three received a "Pass," and one candidate failed, attributing it to emotional exhaustion. Overall, we gathered qualitative insights through **informal surveys** with candidates. Participants were asked about their emotional state, the perceived effect of air raid alerts, and their ability to concentrate. Some responses indicated **emotional exhaustion, anxiety, and difficulty sleeping before the exam.**

The findings suggest that examinees demonstrate remarkable resilience and an ability to concentrate despite extreme external stressors. However, the authors express concerns regarding the potential for increased failure rates if the war persists, given the cumulative effects of mental exhaustion and depressive states among test-takers. Regarding the ongoing war-related stressors, we consider it necessary to have the testing protocols modified to include additional accommodations, such as flexible rescheduling or extended time for candidates under the war conditions.

Atta Gebril, Khaled ElEbyary, Ramy Shabara

AI as a Muse and Mentor: Enhancing L2 Writing Through Reformulation and Model Texts

Reformulation and exemplars have been identified as two effective feedback strategies in second language writing instruction. L2 writing research has confirmed the benefits of these strategies on uptake and noticing. However, little research has examined the effectiveness of AI-generated feedback on noticing and uptake, particularly their role in enhancing syntactic complexity—a persistent challenge for L2 writers. Given this gap, this presentation reports on a quasi-experimental study that primarily investigated the impact of two AI-mediated feedback strategies on error noticing and writing complexity among EFL university students (B1 Level on CEFR). Sixty undergraduate participants were randomly assigned to two experimental conditions: one group engaged with model texts produced by ChatGPT as writing exemplars while the second group employed ChatGPT to reformulate their initial drafts. Adopting a pre, post, and delayed test design, the participants completed three argumentative writing tasks over three different sessions to track potential development of writing complexity. Noticing sheets were also used in the middle stage to document and analyze the participants' awareness of errors. The results revealed distinctive patterns of efficacy between the two feedback modalities in facilitating error recognition and enhancing syntactic complexity. These findings could contribute to theoretical understandings of AI-mediated feedback in second language writing pedagogy and offer practical implications for educators who might be interested in using AI-generated feedback in writing classes.

Alina Reid

Experts at the core: integrating the comparative judgement technique in a mixed-method rating scale development process

This presentation reports on the creation of two analytic scales for measuring performance in writing from sources and written online communication tasks. Traditional rating scale development follows either a theory-driven or data-driven model. While theory-driven scales prioritise generalisability, they often lack descriptive precision and rely on debatable assumptions about language acquisition (Fulcher, Davidson, & Kemp, 2011). In contrast, data-driven scales capture real performance features but can be overly task-specific and require time-intensive analysis of small performance sets, sometimes based on as few as eight scripts (e.g., Upshur & Turner, 1995). Recent approaches integrate theoretical frameworks with empirical data to enhance both generalisability and operational usability (Banerjee, Yan, Chapman, & Elliott, 2015).

This study employed a mixed-methods approach, combining empirical data from 115 candidate performances and descriptive comments from 21 raters with expert knowledge from theoretical models. The 12-stage iterative process involved rank-ordering performances, capturing norm- and criterion-referenced descriptions, drafting scales, piloting, statistical analysis, and revision. A key innovation is the use of comparative judgement (CJ) for rank-ordering and descriptor collection. CJ involves holistic pairwise comparisons, aggregating decisions to rank performances. Unlike traditional data-driven methods, CJ enabled the evaluation of a much larger script set and generated 830 descriptive comments (38,422 words). While CJ is criticised as a standalone assessment (Kelly, Richardson, & Isaacs, 2022), this presentation highlights its value as a preliminary step in scale development, strengthening empirical grounding and construct coverage.

The presentation will explore key challenges and insights from each stage, offering perspectives on alternative rating scale development methods.