LTRC/ALTAANZ celebratory event - Programme

(with Conference abstracts)

Thursday November 19th (all times are in AEDT, GMT+11)

	Session A	Session B
8.45	Opening remarks – Dean, LTRC (Ute), ALTAANZ (Rosemary)	
9.00	**Paper**: Ana Maria Ducasse & Kathryn Hill (Chair: Ute) Translating a feedback intervention across four university language programs	
9.30	**Work-in-progress session 1** (Chair: Ute) 9.30 Yuyun Lei Comparing fluency and disfluency features of L2 speaking performances on spontaneous and controlled speaking tasks 9.45 Olivia Beggins Measuring student perception on the effectiveness of authentic audiovisual materials on Spanish L2 vocabulary acquisition 10.00 Vashti Lee, Margaret Malone & Charlene Polio A comparison of PT analysis of mandarin Chinese OPIs at two sublevels of the ACTFL Intermediate Level 10.15 Agustinus Hardi Prasetyo Designing a scenario-based language assessment literacy test for teachers of English as a Foreign Language in Indonesia	
10.30	**Paper**: Mehdi Riazi (Chair: Ivy) The realisation of Kane's argument-based approach to test score interpretation and use in articles published in Assessing Writing	
11.00	**Paper**: Diep Tran (Chair: Ivy) Argument-based validation of a high-stakes listening test in Vietnam	
11.30	**Paper**: Hang Nguyen (Chair: Ivy) Development and initial validation of the Academic Collocation Recall Test	
12.00	**ALTAANZ AGM**	
12.30		
1.00		

1.30	**Paper**: Jon Clenton & Dion Clingwall (Chair: Susy) Investigating whether speaking tasks influence second language vocabulary and fluency	
2.00	**Paper**: Jianda Liu & Xuelian Li (Chair: Susy) The Validation of the Critical Thinking Ability Questionnaire for EFL learners	
2.30	**Paper**: Harsha Dulari Wijesekera (Chair: Ksenia) Assessments for Learning and promoting 21st-century skills through alternative assessments in English Language classrooms	
3.00	**Paper**: Neslihan Onder-Ozdemir (Chair: Ksenia) Assessment practices in medical English during medical education	
3.30	**LTRC 30-year anniversary symposium: Mapping tensions between language testing, policies, and practices** (Chair: Carsten) Cathie Elder: Framing the LTRC's policy contribution in the languages area Ute Knoch: The challenges of providing expert advice in policy contexts Jason Fan & Jin Yan: Navigating tensions between language testing intentions and policy imperatives: The case of the College English Test (CET) in China Kellie Frost: Negotiating the boundaries of English: the role of tests and language testers in Australia's skilled migration policy space Discussants: Tim McNamara & Joe Lo Bianco Closing remarks: Lesley Stirling (Head of School, School of Languages and Linguistics) Followed by LTRC anniversary video	
5.00		
5.30	**ALTAANZ social event**	
6.30	**Paper**: Ruolin Hu & Danijela Trenkic (Chair: Jason) Understanding the effects of IELTS test preparation and repeated test-taking on candidates' IELTS scores, proficiency, and IELTS's predictive validity	**Paper**: Eleni Meletiadou & Dina Tsagari (Chair: Ute) Peer Assessment: A dynamic learning-oriented tool for the development of writing skills

7.00	**Work-in-progress session 2** (Chair: Jason) 7.00 Ester Dominguez Lucio & Vahid Aryadoust Understanding the impact of textual features during listening: An on-going neuroimaging and eye-tracking study 7.15 Bassey Antia & Karin Vogt Dynamic assessment of academic reading: Peeping into a corpus of annotated texts 7.30 Eleni Meletiadou The use of Peer Assisted Learning/Mentoring (PAL/M) as an inclusive peer assessment strategy within foundation year practice	7.00pm **Paper**: Albert Weideman (Chair: Ute) Complementary evidence in the early stage validation of language tests: Classical Test Theory and Rasch analyses 7.30pm **Paper**: Rose Clesham & Sarah Hughes (Chair: Ute) Language Testing Concordance Studies: Valid Alignment by Design

Friday, 20th November (all times are in AEDT, GMT+11)

	Session A	**Session B**
9.30	**Paper**: Okim Kang, Hyunkee Ahn, Kate Yaw & Soh Yon Chung (Chair: Kellie) Impact of test-taker's background on score gain on IELTS	
10.00	**Paper**: Noriko Iwashita (Chair: Kellie) Stakeholders' perceptions of IELTS speaking and writing tests and their impact on communication and achievement	
10.30	**Work-in-progress session 3** (Chair: Morena) 10.30 Aynur Ismayilli Karakoc Developing a marking rubric for the integrated reading-writing test 10.45 Viola Lan Wei The measurement of implicit linguistic knowledge: a meta-analysis 11.00 Liz Kose Assessing speaking in a Post-Entry Language Assessment 11.15 Leila Zohali Investigating the pedagogical usefulness of Automated Writing Evaluation (AWE) System in academic writing instruction 11.30 Tracy Ware & Andrew Kelly Integrating Post-Entry Language Assessments into the Curriculum. What works well?	

	11.45 Xuan Minh Ngo A sociocultural analysis of teacher assessment literacy development: The promises and pitfalls of taking a Vygotskian perspective on assessment	
12.00		
12.30	**Paper**: David Wei Dai (Chair: Andrew) Expanding the interactional competence construct: Criteria from domain experts on interactional success in everyday life	
1.00	**Paper**: Tingting Liu & Vahid Aryadoust (Chair: Andrew) The effect of public speaking anxiety on the public speaking performance of tertiary-level students	1pm **ALTAANZ student event** (Chairs: Maria/Phuc)
1.30	**Paper**: Huawei Shi & Vahid Aryadoust (Chair: Andrew) A comprehensive review of research on automated written feedback	
2.00 2.30	**Editor session: Publishing in language assessment journals** (Chair: Maria) Q&A session with journal editors for graduate students and early career researchers Lyn May & Sally O'Hagan: Papers in Language Testing and Assessment Martin East: Assessing Writing Paula Winke/Dylan Burton: Language Testing Aek Phakiti: Language Assessment Quarterly	
3.00	**Paper**: Denise Angelo & Catherine Hudson (Chair: Johanna) Moderation as a mechanism for collecting nationally consistent EAL/D data inclusive of Indigenous EAL/D learners	
3.30	**Paper**: Rie Koizumi, Akiyo Watanabe, Makoto Fukazawa, Chihiro Inoue (Chair: Johanna) Examining learner perception toward test feedback in classroom-based speaking assessment using a validity framework	
4.00	**Teacher sharing session** (Chair: Sally) 4.00 Cathy Bow & Susy Macqueen Assessing learners of Indigenous languages - balancing perspectives 4.15 Fan Chen From structure to substance: scaffolding the development of argumentation in EFL academic writing 4.30 Maha Hassan Upgrading critical evaluation and skill development through Group Assessment 4.45 Gaby Lawson Turbocharge your peer feedback with summative criteria	
5.00	**Closing comments (LTRC, ALTAANZ)**	

Conference Abstracts

Thursday November 19th (all times are in AEDT, GMT+11)

Papers

9.00 AM

Translating a feedback intervention across four university language programs

Ana Maria Ducasse, RMIT University, Australia

Kathryn Hill, Deakin University, Australia

While the importance of feedback for promoting learning is well-documented, learners can only benefit to the extent they actually do something with it (Boud & Molloy, 2013). Hence a lot of recent research focusing on how to increase learner engagement with feedback. However, there is increasing recognition of the situated nature of assessment and feedback practices (Larenas & Brunfaut, 2018; Scarino, 2013; Xu, 2015) with the result that interventions which are successful in one setting will not necessarily succeed in another (Ajjawi, Molloy, Bearman & Rees 2017).

This paper describes the extension of an intervention designed to improve learners' engagement with written feedback in a university-level Spanish language program (Ducasse & Hill, 2020) to three additional language programs (French, Chinese and Japanese) at the same institution.

Research questions:

1. Were there any differences in outcomes for the four languages?

2. What factors accounted for these differences?

Method

Participants were 272 university students enrolled in pre-intermediate or upper-intermediate language classes (Spanish, French, Chinese and Japanese) and the class teachers for each language (n=4).

Based on an adaptation of the 'reflective feedback conversation' (Cantillon & Sargeant, 2008) the intervention focused on written texts submitted for feedback and/or assessment.

The procedure

1. Students produce a written text, self-assesses and request specific feedback;

2. The teacher provides the requested feedback and returns an electronic copy of the document to the students for filing (e.g., in an e-portfolio);

3. Students interact with the teacher (online or in person) about their feedback;

4. On each subsequent occasion, students need to review previous feedback and specify action taken before self-assessing and requesting specific feedback.

Data collection & analysis

Data included:

1. Online student questionnaires;

2. Data on the number of students who:

- submitted work for feedback

- uploaded feedback to their e-portfolio

- interacted with teachers about feedback, and

- acted on their feedback, as well as the number of times they accessed their e-portfolio;

3. Recordings of teacher discussions;

4. Teacher questionnaires, and

5. Copies of corrected work.

Quantitative data were analysed using descriptive and inferential statistics. Qualitative data were analysed using thematic content analysis. Copies of student work were analysed for evidence of learning from feedback. Contextual influences (RQ2) will be analysed using Bronfenbrenner's ecological framework of human development (1977, 1979).

Results

While the results of the intervention were positive overall, there were some marked differences between the respective languages and class levels. Discussion will focus on the extent to which these were due to intrinsic differences between the languages themselves (e.g., character vs. Latin script) or to contextual differences. While further refinements are indicated, this approach has demonstrated benefits for student and teacher assessment literacy, for sustainability of assessment and for improving student learning.

10.30 AM

The realisation of Kane's argument-based approach to test score interpretation and use in articles published in Assessing Writing

Mehdi Riazi, The Macquarie University, Australia

Two concepts have played a pivotal role in the testing and assessment field. These two concepts are validity and validation. As Kane (2001) explicated, the current conceptions of validity are based on Messick (1989) and the 1999 Standards for Educational and Psychological Testing. The two concepts have a history of almost 100 years and have developed over time. The conceptualisations and operationalisations of these two concepts have developed over time to respond to the developments in the theoretical frameworks in the field of language testing and assessment. Argument-based framework has been one of the crucial operationalisations of Messick's unitary theory of validity and its

corresponding validation procedure. Kane has pioneered the development and dissemination of the argument-based framework to test validation.

In this paper, I am going to report on and discuss 8 studies that have purely used Kane's argument-based framework in investigating issues related to writing assessment. These 8 articles are extracted from the Journal of Assessing Writing. There are many articles in this journal that have used a sort of argument-based framework to address issues related to the writing assessment. However, it seems that these 8 articles are the only articles in the journal's life cycle that explicitly use Kane's argument-based framework. The focus of the presentation will thus be on how the writing assessment researchers have used Kane's argument-based framework to address issues related to the writing assessment. In particular, the presentation will focus on how the "interpretive" and "validity" arguments are designed and supported with relevant evidence in these articles. Details of the kind of inferences, assumptions, and backups installed in the interpretive and validity arguments will be explicated in the presentation. A critical review of the interpretive and validity arguments in the 8 articles will help to make suggestions for future studies that are going to use argument-based framework to address issues related to writing assessment.

11.00 AM

Argument-based validation of a high-stakes listening test in Vietnam

Diep Tran, Victoria University of Wellington, New Zealand

This paper presents the highlights of my recently completed PhD thesis which focuses on validating a high-stakes standardized listening test in Vietnam.

More than a decade ago, the Vietnamese Government announced an educational reform to enhance the quality of English language education in the country. An important aspect of this reform is the introduction of the localized test of English proficiency which covers four language skills, namely listening, speaking, reading, and writing. This high-stakes English test is developed and administered by only a limited number of institutions in Vietnam. Although the validity of the test is a considerable concern for test-takers and test score users, it has remained an under-researched area. This study aims to partly address the issue by validating a listening test developed by one of the authorized institutions in Vietnam. This test consists of 35 multiple-choice items, as directed by the Ministry of Education and Training. In this presentation, it is referred to as the Locally Created Listening Test (LCLT).

Using the argument-based approach to validation, this research aims to develop a validity argument for the evaluation, generalization and explanation inferences of the LCLT. Three studies were carried out to elicit evidence to support these inferences. The first study investigated the statistical characteristics of the LCLT test scores, using the Rasch model as the main analytical tool. The second study, by means of think-loud protocols, shed light on the extent to which test items engaged the target construct. This study will be the focus of my presentation. The third study examined whether test-takers' scores on the LCLT correlated well with their scores on an international English test that measured a similar construct. Expert judgements, Rash analysis and correlational analysis were the research methods used in this study.

The LCLT was found to have major flaws that affected the validity of score interpretations. They were: very high level of difficulty, construct under-representation, construct-irrelevant factors, low correlation

between test-takers' scores on the LCLT and their scores on an IELTS Listening test. The multiple-choice format was a source of invalidity since it unreasonably increased the cognitive load of many test items. It also promoted the use of test-taking strategies and dependence on the written text, especially when the recording was played only once. Test-takers' performances were also affected by other factors such as the number of speakers in the spoken text, speech rate and order of items.

The research findings raised an alarm about the technical quality of the LCLT, suggesting that the localization of language tests is indeed easier said than done. Better construct conceptualization, clear guidance and effective quality control measures are needed to create good localized English tests. For the development of listening tests, a variety of response methods and the double-played format should be taken into consideration. To effectively control item difficulty, it is important to consider the listening subskill that an item targets in relation with text- and task-related factors.

11.30 AM

Development and initial validation of the Academic Collocation Recall Test

Hang Nguyen, Victoria University of Wellington, New Zealand

Academic collocations (i.e., two-word combinations such as 'empirical evidence' and 'key element') are important as they occur in a wide range of registers and disciplines. Without knowledge of academic collocations, learners may use incorrect or deviant collocations, which may affect intelligibility and formality. For example, instead of using collocations such as 'central issue' or 'external factor', learners may use 'centre issue' or 'outside factor', which are not appropriate in academic writing. Although several lists of academic collocations have been compiled (e.g., Ackermann & Chen, 2013; Lei & Liu, 2018) to assist learners in developing their academic collocational knowledge, there are still gaps in taking steps beyond creating a word list, such as including it in vocabulary testing (Nation, 2016). The present study is intended to fill that gap by creating an Academic Collocation Recall Test (ACRT) based on the Academic Collocation List (Ackermann & Chen, 2013) to test learners' ability to produce academic collocations.

As the ACRT is a newly created test, an empirical study was conducted to investigate whether the test is of good quality and achieves its intended purpose. A total of 343 students in Vietnam and New Zealand participated in the study. The participants took the ACRT and the Vocabulary Size Test (VST) (Nation & Beglar, 2007) via Qualtrics as a testing platform. The ACRT has 60 items written in a gap-filling format. Test-takers were asked to provide a suitable academic collocation for a given context. The VST has 140 multiple-choice questions to measure learners' vocabulary size (i.e., how many English words they know). Following the tests, 44 test-takers participated in post-test interviews to give their reflections about the ACRT and to retake the test verbally. The results indicated that the ACRT exhibited high reliability with items sharing the same construct and ranging in difficulty levels, which means it is suitable for learners at a wide range of English proficiency. In addition, there was a positive moderate correlation between the scores on the ACRT and scores on the VST.

The ACRT is beneficial for language teachers as a new and reliable measure of learners' knowledge of academic collocations. The test can help to distinguish between learners who have enough collocational knowledge for academic study from those who need additional support. The ACRT can be used as one single administration at the beginning of an English course to help teachers plan an appropriate course of instruction. It can also be used as pre- and post-tests to measure learning gains. Furthermore, the

ACRT is a useful research tool to investigate how knowledge of academic collocations correlates with general vocabulary knowledge.

The focus of the presentation is to briefly introduce the ACRT and provide preliminary validity evidence for the ACRT based on the Rasch model analysis. Interpretations of the tests scores and how to use the test will also be discussed. The presentation will end with recommendations for future research.

1.30 PM

Investigating whether speaking tasks influence second language vocabulary and fluency

Jon Clenton, Hiroshima University, Japan

Dion Clingwall, Hiroshima University, Japan

When examining oral fluency, researchers (e.g. De Jong, 2016; Tavakoli, 2016) suggest that fluency varies according to task. Also, research (e.g. Clenton, De Jong, Clingwall, & Fraser,

2020; Uchihara & Clenton, 2018; Uchihara & Saito, 2016) indicates that vocabulary knowledge varies according to speaker fluency. When taken together, to date, no single study has concurrently investigated whether both fluency and vocabulary knowledge are influenced by speaking task. The current paper, therefore, investigates multiple tasks' influences on both oral fluency and lexical knowledge.

Participants were 44 L1 Japanese learners of L2 English, (CEFR: B1-B2). Participants responded to the three IELTS speaking section (monologic, quasi-dialogic, and dialogic) tasks, and vocabulary knowledge was elicited with both a productive vocabulary knowledge task (Lex30; Meara & Fitzpatrick, 2000), and a receptive vocabulary knowledge task (X_lex; Meara & Milton, 2003). Our fluency measures (De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012; Tavakoli, 2016) included pause frequency, pause length, and articulation rate. Following Clenton et al. (2020), we explore the vocabulary produced in response to speaking tasks with the Academic Spoken Word List (ASWL; Dang, Coxhead & Webb, 2017).

Findings suggest that relationships between the two main variables, fluency and vocabulary knowledge, vary according to speaking task. We report three main findings: i) overall fluency indices were higher for the dialogic than for the monologic or quasi-dialogic tasks; ii) significant differences between the three speaking tasks can be seen in the pausing data; and, iii) corpora analysis shows significantly different vocabulary production in each of the three speaking tasks. We discuss the implications of these findings in terms of assessment research and practical classroom applications.

2.00 PM

The Validation of the Critical Thinking Ability Questionnaire for EFL learners

Jianda Liu, Guangdong University of Foreign Studies, China

Xuelian Li, Guangdong University of Foreign Studies, China

In China, the teaching of critical thinking ability (CTA) is required as a part of the English curriculum, as shown in English Curriculum Standards for Chinese Senior Middle Schools (ECSCSMS) (MOE 2018). The construct of CTA is integrated in English learning, and it is different from the general concepts of CTA (e.g., Facione 1990; Paul & Elder 2006; Watson & Glaser 1994; Ennis 1991; Black 2008). Due to the different construct, relevant instruments were urgently needed to measure students' CTA. This study

intended to validate the Critical Thinking Ability Questionnaire, which was developed in previous research (Li 2020).

Guided by Messick's validity framework (1989, 1996), this study investigated structure validity and content validity of the questionnaire. A big-scale questionnaire survey was conducted to confirm its theoretical structure. 1052 valid questionnaires were collected by senior middle school students, from eight schools in six provinces in China. Then, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were conducted via SPSS 24.0 and CFA lavaan .6-5 in R package Version 3.6.2 respectively.

To further validate the structure validity of the questionnaire and its content validity, semi-structured interviews were conducted with two experts and four senior middle school students, following the questions in the interview outline (expert version and student version). WinMAX 2000 was applied to analyse the interview data.

Research results identified and confirmed a seven-factor model: analysis, inference, argument evaluation, language construction, self-reflection/correction, cultural comparison and credibility evaluation of language sources. The questionnaire was confirmed to be of high quality. The structure validity was quite good: a clear seven-factor structure model was identified via EFA; The model fitted very well with the data (Ï‡2=477.682, p=.812>.05, RMSEA=.000<.06, SRMR=.043<.08, CFI=1.000>.95, TLI=1.001>.95); the interviewed teachers and students confirmed the consistency between the construct and the theory. The content validity was supported by the interviewed experts, who confirmed the coverage, relevance, clarity and specificity of the questionnaire. In addition, the internal reliability coefficients of the questionnaire (.928) and subscales (.673-.863) were higher than those of other CTA instruments published so far.

The theoretical contribution of this research lies in the construct framework of CTA put forward in this study. Unlike general CTA frameworks, this framework integrates CTA with senior middle school students' English learning. The first five factors are important dimensions in other influential CTA frameworks and they were supported by empirical data in this research. Additionally, this research confirmed that cultural comparison and credibility evaluation of language sources are especially crucial for second language learners.

This study may also contribute pedagogically to Chinese education departments, SMS teachers and SMS students. The construct framework can be added into ECSCSMS and used by education departments to instruct and supervise teachers' cultivation of CTA, and to assess the effective implementation of the English curriculum. With the questionnaire, SMS teachers may precisely judge students' strengths of CTA, and teach CTA more effectively. Students may learn CTA more effectively due to a correct evaluation by using the questionnaire.

2.30 PM

Assessments for Learning and promoting 21st-century skills through alternative assessments in English Language classrooms

Harsha Dulari Wijsekera, Open University of Sri Lanka, Sri Lanka

A country's education system is envisioned to inculcate 21st-century skills in students: Critical thinking, Creativity, Communication, Collaboration. However, the educational authorities tend to focus only on curriculum and material development to achieve this end, whereas assessment processes are neglected.

Sri Lanka too perpetuates examination-oriented education, which debars deep learning and deprives the acquisition of social and soft skills while producing highly competitive, self-centered individuals. It is also questionable if only high-stake examinations (or even term/semester tests) provide valid measurements of English as a second language (ESL) skills. Mainly, in the absence of listening and speaking test items the content validity of ESL high-stake examinations is questionable, which also has a negative washback impact on classroom learning/teaching. Though there is tumult for learner-centered teaching, learner-centered assessments seem to be taken-for-granted. Contrastingly, Alternative Assessments (AA) may provide solutions for these issues, which are also considered as Assessments for learning or Learning through Assessments. This paper investigates how classroom-based AA can compensate these lapses. It discusses the experience of secondary and university teachers who conducted a four-week AA project to assess their students, which included planning AA activities, making students aware of the procedures such as peer feedback/self-assessment, utilization of rubrics, self-reflection, etc. Based on the analysis of the project reports, an online questionnaire comprising Likert-scale items and guided-reflections was administered to respective teachers (n- 16). Thematic content analysis was utilized to analyze qualitative data. The analyses illustrate that, irrespective of the proficiency levels of the students, AA makes testing more meaningful and personalized; promotes motivation, engagement, autonomy, self-confidence, deep and life-long learning; increases cooperation and interdependence among students; and reduces anxiety. AA also facilitates noticing the gaps, consciousness-raising, hypothesis testing, students' efforts evenly on all language skills, teacher-student rapport, and reduces teacher domination on assessment while facilitating students' better understanding of what good performances in terms of learning goals and expected standards. Some respondent teachers expressed that the quality of self and peer assessment increasingly became similar to teachers' feedback. These findings provide evidence that AA is an approach to foster 21st-century skills and achieve SDGs. The challenges reported were time-constraints given the compact syllabuses to cover, discomfort towards AA among the students who are so accustomed to conventional exams, and lack of resources. Also, most teachers doubt institutional support for AAs given the lack of awareness among the authorities and academic staff. Overall, utilized well, AA can facilitate bridging the gaps in high-stake tests and provide better assessments on learning which cannot be fully assessed only through high-stake conventional examinations, while stimulating professional growth and best practices in ESL teachers. The paper recommends legitimizing AA in education as a policy given the intersectionality between assessment and learning since a country will not reap benefits despite high investment in other areas of education unless assessment procedures are adjusted for deep learning.

3.00 PM

Assessment practices in medical English during medical education

Neslihan Onder-Ozdemir, Bursa UludaÄŸ University, Turkey

Testing language for specific purposes (LSP) is discussed with a focus on two key aspects in the literature (i.e., the authenticity of the task and the interaction between language knowledge and specific content knowledge), which may distinguish LSP testing from general purpose language testing (Douglas, 2000, p. 2). There are very few studies on testing in LSP, which remained under-researched. This longitudinal study sets out to describe medical English examinations accompanied by undergraduate medical students' voices. Medical English was an elective ESP course in the Faculty of Medicine. Questions for Medical English examinations were built from current medical research articles from international journals with a high impact factor and also from medical students' self-reflection reports about their

undergraduate research experience using medical English knowledge. Medical students' written feedback before clinical years in the lecture halls in the Faculty of Medicine and during clinical years in the university hospital were collected to assess the testing to investigate whether the medical English course and the testing conducted have had any effects on medical students' life (n=79). In this study, the data were collected using interviews from medical students before clinics and also email interviews in the clinical years given that clinical years are hectic as medical students have responsibilities at the university hospital. In light of data, the findings will elucidate how Medical English assessment brings the added value through connecting teaching, peer-assisted learning, curriculum, assessment and also real-life use of medical English. Sample questions from the medical English examinations and also interview data extracts will be shared.

6.30 PM (Session A)

Understanding the effects of IELTS test preparation and repeated test-taking on candidates' IELTS scores, proficiency, and IELTS's predictive validity

Ruolin Hu, UCL Insitution of Education, UK

Danijela Trenkic, University of York, UK

Washback on high-stakes language proficiency tests, such as IELTS, has been intensely debated. Dedicated test preparation and repeated test-taking are its two main manifestations. Although both are common among IELTS candidates (Hawkey, 2006), little is known about the concurrent effect of test preparation and repeated test-taking on candidates' scores and proficiency, and their consequential effect on IELTS's predictive validity. We conducted two studies to address this gap.

Study 1 investigated the concurrent effects of IELTS preparation using a 4-week pretest-posttest design with two groups (IELTS-prep n=45, Control n=44) and two measures (IELTS and Oxford Online Placement Test, OOPT). OOPT scores were similar in both groups at both times; by contrast, there was a large gain in IELTS scores, but only in the IELTS-prep group. This suggests that dedicated test-training could lead to IELTS gains that are not reflected on other measures of English proficiency.

Study 2 explored the predictive validity of IELTS and how test preparation and repeated test-taking influence the relationship between IELTS and alternative measures of proficiency in international students. We found for international Chinese students (n=153), those who met the entry requirements without attending IELTS-preparation programmes and/or achieved the required score with fewer attempts had significantly higher English proficiency scores on alternative measures (C-test; Duolingo English Test) than students who underwent such preparations and/or attempted the test more times. Overall, the findings confirm that IELTS scores can be boosted beyond what can be detected by alternative tests by attending dedicated test preparation and to a lesser extent by taking the test repeatedly. Despite that, IELTS was a good predictor for academic achievement, after factoring in the linguistic demand of the students' academic programme. We offer insights on the validity of IELTS as a measure of language proficiency and as a predictor and consider theoretical and practical implications.

6.30 PM (Session B)

Peer Assessment: A dynamic learning-oriented tool for the development of writing skills

Eleni Meletiadou, London South Bank University, UK

Dina Tsagari, Olso Met University, Norway

In the last few decades, researchers and educational authorities express their concern for EFL students' poor writing performance and lack of motivation (Shang, 2019; Challob, Bakar & Latif, 2016). Researchers indicate that peer assessment (PA) can support a better integration of teaching with assessment of progress in learning (Hovardas, Tsivitanidou & Zacharia, 2014; Tsagari & Meletiadou, 2015). Bearing this in mind, the current study employed a pre-test post-test quasi-experimental design and aimed to explore: the effect of PA and teacher assessment (TA) on EFL students' writing performance, the impact of PA and TA on EFL students' writing quality as opposed to TA only, and EFL students and teachers' attitudes towards PA. Participants of the study were: (a) twenty groups of ten Cypriot intermediate adolescent EFL students (200 students in total); (b) 20 qualified EFL teachers, and (c) an external assistant. All participants received adequate training in PA methods. Several instruments of data collection were used. 1,300 student writing samples were generated and 400 students' pre- and post-test essays were analysed both quantitatively (by comparing students' marks) and qualitatively (by analysing students' texts and comparing some of their features among the experimental and control groups). Lengthy questionnaires with closed and open items were administered to both 200 experimental group students and 20 teachers to explore their attitudes towards PA of writing. Interviews were conducted only with teachers, while students took part in whole-class discussions. Data collected via the questionnaires, interviews and whole-class discussions were analysed both quantitatively and qualitatively. The study outcomes indicated that PA and TA can have a moderately positive impact on students' writing performance and a similarly significant impact on EFL students' writing quality. EFL teachers and students' attitudes towards PA of writing were positive and they both expressed their wish for multiple forms of assessment. The present study strongly suggests that PA is a valuable tool to use in developing adolescent EFL students' writing skills. Students were very supportive of the activity and showed a significant grade improvement following revision subsequent to PA and TA (also in Ying, Schunn & Yu, 2019) with lower graded papers showing the greatest improvement. Learners increased their writing performance in all five aspects (mechanics, organization, content, focus, vocabulary and language use), and improved the writing quality of their essays pertaining to lexical complexity, accuracy and some aspects of grammatical complexity and fluency. Teachers were also satisfied with the overall outcome of the PA implementation. With the support of students and teachers' favorable reactions, moderately positive improvement in students' writing performance in almost all aspects of writing and benefits brought about by PA, this study concludes that PA is a viable alternative to involve students in the assessment process and promote independence in secondary education (also in Shen, Bai & Xu, 2020). In response to the need for more experimentation, the present study provides a PA implementation model for secondary school EFL writing classes enabling teachers to enhance students' performance and motivation which in so far has been absent (Lee, 2019).

7.00 PM

Complementary evidence in the early stage validation of language tests: Classical Test Theory and Rasch analyses

Albert Weideman, University of the Western Cape and University of the Free State, South Africa

While a good test gains in reputation as it is administered over time, the early stages of its validation are perhaps the most critical. The paper takes the early stage validation of an Assessment of Language for Economics and Finance (ALEF) to demonstrate the employment of an extended framework of principles

for the initial validation of a test. ALEF is an assessment of the ability of prospective or entry-level employees in the banking sector, to determine whether the language ability of those employees is at the required level for entry into post-school training. Six claims are investigated, linked to the same number of principles in the framework that is employed. The claims are that ALEF

[1] exhibits a sufficient degree of homogeneity, for which there are several warrants;

[2] shows that it is a reliable measure of language ability on several counts: at test level, as well as at item level;

[3] is organised as a differentiated whole, with each subtest functioning both uniquely and together with others in contributing to the viability of the measurement;

[4] exhibits an adequate degree of fit by distributing candidates normally as regards language ability, while it simultaneously has an acceptable degree of difficulty; moreover, it can be demonstrated that the test fits the ability of candidates, in that there is a likelihood of minimal misfit either of items or persons in its measurements;

[5] yields scores that are clear, meaningful and intelligible;

[6] measures so consistently that the number of potential misclassifications it produces is smaller than 5% of the total test population, and the test developers have a way of identifying such misclassifications in order to give those potentially misclassified a fair chance of taking a similar test.

For each claim there is at least one, but often multiple warrants. By relating the claims and their warrants to principles of test design, a fair measure of integration is obtained for the argument. The investigation of the claims uses two of the methodological tools most frequently employed to muster empirical evidence for validating test design, namely Classical Test Theory (CTT) and Rasch analyses. While most language tests designed in South Africa have used CTT, the employment of Rasch analyses has been more limited. The paper provides an example of how the latter kind of analysis can complement the former.

There is now general agreement that the validation process should be reported in the form of an argument. The format of such integration is, however, still contestable ground. Since the claims investigated in this instance derive from a framework of principles for responsible test design, the analyses also show a possible format for bringing together multiple sets of evidence to justify the design and implementation of a language test, a process that is conventionally termed 'validation'. The conclusion is that test validation may more aptly be conceived of as the process of designing language tests responsibly. For that, we need to relate the process of responsible design to a theory of applied linguistics.

7.30 PM

Language Testing Concordance Studies: Valid Alignment by Design

Rose Clesham, Pearson, UK

Sarah Hughes, Pearson, UK

At the core of the language teaching, learning and assessment, there are three key alignments:

1. The alignment of learning objectives to curriculum or content standards

2. The alignment of performance standards (test outcomes) to content standards

3. The alignment of test measurement scales- Performance Standards alignment.

Alignments of language learning objectives to a recognised schema or framework (eg. the CFER) are essential in order to support teaching and learning programmes and to indicate general levels of progression or attainment. There are guides and manuals publicly available to show how to develop such content alignments. Having said this, it should be remembered that different methodologies can lead to different alignment outcomes and human judgemental exercises can be influenced by unconscious bias and heuristics. Most language content standards were never designed to be treated empirically, and as such there are no definitive empirical alignments, just best fits. It is also the case that such alignments can change if either learning objectives or content standards change.

This also applies to the alignment of performance standards to content standards. If the construct, design or scoring of a test changes over time, or the content standards change, it follows that the alignment of a test to content standards also needs to change. This has happened a number of times in the area of language testing, for example where test alignments to the CFER have changed. Irrespective of any construct or content standards changes, alignment studies over time are necessary to present concurrent test validity.

The final piece of the jigsaw is the alignment of the performance standards of different language tests, which of course carry significant currency in terms of academic or professional entry requirements. In many ways this should be the easiest alignment exercise because it is simply comparing performance standard outcomes. The simple question is if the same person took different tests at approximately the same time, how would their scores compare?

Any performance standards comparative analyses work is complicated as the tests themselves are invariably different in terms of the test constructs, item types, rubrics, marking methods and standard setting, however for the sake of argument, we must assume that all high stakes language tests are assessing language proficiency as modelled by frameworks such as the CFER.

Although the concept of implementing an alignment study sounds simple, the detail is important. What is the concordance study design, how is the sample selected, how representative is the sample across the measurement range of the tests, how far apart do test scores need to be in order to be 'approximately the same time', how is test data collected and what is the rationale for the selected concordance equating methodology?

This presentation will describe an alignment study Pearson has recently completed to revalidate the concordance of performance standards between the PTE and IELTS Academic tests. This study is part of a project spanning several years to continuously monitor the relationship between performance outcomes of different tests. We will describe the research design, the alignment methodology used and show the research outcomes.

Work-in-Progress session 1

9.30 AM

Comparing fluency and disfluency features of L2 speaking performances on spontaneous and controlled speaking tasks

Yuyun Lei, University of Illinois at Urbana-Champaign, USA

This ongoing project compares fluency and disfluency features of second language (L2) speaking performances on elicited imitation (EI) tasks and the ACTFL Oral Proficiency Interview (OPI), as an effort to elucidate the constructs measured by these two types of speaking tasks. Although EI has been shown to be an effective proficiency measure and can predict scores on spontaneous speaking tasks, its construct validity remains questioned mostly due to the imitative nature of the task, with some scholars arguing that EI does not elicit the same kind of language processing in real-life speech or conversation. To address this concern, this study aims to examine characteristics of speaking performances that would inform the underlying processes of speech production on both EI and OPI tasks. Fluency and disfluency features will be focused on in this study as they can offer a window into the investigation of cognitive processes involved in speech production (Segalowitz, 2010). A total of 80 L2 Chinese learners at different curricular levels in U.S. universities were recruited to complete an EI test and a simulated OPI test of Chinese. The simulated OPI test comprises six tasks targeting intermediate to superior levels in the ACTFL Proficiency Guidelines (2012). The EI test consists of 72 sentences featuring varying lexico-grammatical complexity levels, targeting the same range of proficiency levels. Participants' performances will be holistically scored and analyzed for a wide array of fluency and disfluency features, including amount and rate of speech, pausing and repairs. The features will be compared across proficiency levels and between EI and OPI tasks. The purpose of the presentation is to seek feedback on appropriate comparison of the analyzed features between tasks or additional features that would help identify similarities and differences in the speaking performances on EI and OPI tasks. It is hoped that the findings of the study can add validity evidence for EI as well as bridging the gap in constructs between controlled and spontaneous speaking tasks.

9.45 AM

Measuring student perception on the effectiveness of authentic audiovisual materials on Spanish L2 vocabulary acquisition

Olivia Beggins, Temple University, USA

This study seeks to investigate the effectiveness of YouTube grammar tutorials on L2 vocabulary acquisition. More specifically, this study tests learners' ability to distinguish between the Spanish verbs â€œsaberâ€ and â€œconocerâ€, both of which translate to â€œto knowâ€ in English. An objective pre and posttest were administered to students during two separate lessons that incorporated YouTube videos that were selected to present the target lexical pair. Two, university-level basic Spanish classes participated in this study. Each of the two classes received a different treatment during the lesson on the target forms: while Class A was shown a grammar tutorial that used explicit grammar teaching to explain the target forms, Class B viewed an authentic video in the target language that provided implicit instruction regarding the target forms. In addition to completing the pre- and post-tests, the students completed a survey to ascertain their perception of each of the videos, and which of the two best helped their acquisition of the target forms. This study aims to answer the following questions:

a. Do authentic materials that include implicit grammar instruction facilitate learner acquisition of vocabulary in an L2?

b. What are learners' perceptions of authentic materials vs. explicit grammar instruction?

Key words: Authentic materials, vocabulary acquisition, materials use, Spanish language teaching.

10.00 AM

A comparison of PT analysis of mandarin Chinese OPIs at two sublevels of the ACTFL Intermediate Level

Vashti Lee, Georgetown University, USA

Margaret Malone, Georgetown University, USA

Charlene Polio, Michigan State University, USA

The popularity of Chinese as a foreign language has increased rapidly in North America over the past decade (Looney and Lusin, 2018). As enrollment has increased, it has been accompanied by a steady rise in demand for studies of Chinese second language acquisition (Jiang, 2014). Over the years, there has been much attention on L2 Chinese interlanguage development. While there are only a small number of studies that has applied Processability Theory (PT) to Chinese learning contexts since the theory was put forth by Pienemann just over two decades ago (Wang, 2011), the few studies that have attempted to investigate the developmental sequence of learner's grammar agree to a large degree on their findings. This suggests that the developmental stages identified by these studies may be used as a means to conduct linguistic profiling on Chinese L2 learners.

The present study aims to contribute to this growing body of literature by investigating the extent to which the stages of Chinese L2 learners' syntactic development informed by previous research on Chinese grammatical structure acquisition pattern with the ACTFL Oral Proficiency Interview (OPI) guidelines. Two different sublevels of the ACTFL Intermediate level, Intermediate Mid (IM) and Intermediate High (IH), are targeted. This narrow range of proficiency is of particular interest, because, in the transition from Intermediate mid to Intermediate High, the descriptors of ACTFL Proficiency Guidelines (2012) shift from sympathetic interlocutors to less sympathetic interlocutors. This shift is significant as it marks a significant change in expectations regarding the test taker's language performance during the interview.

Data informing this study include eight OPI recordings rated by ACTFL certified raters, with four rated at the IM level and another four rated at the IH level. All recordings are from Chinese L2 learners who studied in the same language program in Northern America. The eight OPIs have been transcribed, and using the conclusions drawn by Brolin (2017) on the hierarchies of Chinese interlanguage grammatical structure, the transcriptions were coded through the application of an emergence criterion and categorized across PT stages. This process allows for in-depth comparison between the L2 Mandarin IM OPIs and IH OPIs.

Discussion in the presentation of this work-in-progress study will center on possible explanations for the relationship between the emergence of certain Chinese syntactic structures and the communicative ease between learners and sympathetic interlocutors versus native speakers not familiar with interlanguage of learners of Chinese. Finally, the pedagogical implications of possible study findings will also be explored.

10.15 AM

Designing a scenario-based language assessment literacy test for teachers of English as a Foreign Language in Indonesia

Agustinus Hardi Prasetyo, Iowa State University, USA

Assessment has become an integral part of teaching; therefore, it is necessary for teachers to be assessment literate. Assessment literate teachers will be able to make informed decisions in their teaching to help students learn and improve their teaching. However, studies in language assessment literacy (LAL) show that teachers lack core knowledge of LAL and need language assessment training. Since few studies have been conducted to investigate English as a foreign language (EFL) teachers' LAL especially in Indonesian EFL context, this mixed-method approach study aims to 1) elucidate what LAL core knowledge and skills are by surveying Indonesian EFL teachers on what constitutes LAL knowledge and skills based on their EFL context, 2) determine whether they have those knowledge and skills, and 3) identify assessment trainings they need. This study is also intended to design a scenario-based LAL test based on the survey to facilitate EFL teachers' professional development. The survey will use the questionnaire adapted from Kremmel and Harding's (2019) Language Assessment Literacy survey and will be distributed to Indonesian EFL teachers through social media and mailing lists. Follow-up interviews will be conducted for those teachers who are willing to be interviewed. They will also be asked to submit any assessment-related documents. Based on the analysis of the questionnaire, interview transcripts, assessment-related documents, and Standards documents issued by the Indonesian government, a scenario-based LAL test will be constructed. This test will then be piloted to some pre-service Indonesian EFL teachers to measure their LAL level and identify LAL training needs. The results of the piloting stage will also be used to determine the validity of the test and to improve the test performance. The results of this study can help advance the field of language assessment and specifically language assessment literacy by proposing the core knowledge and skills of language assessment literacy which can be useful in the discussion of the construct of language assessment literacy.

Work-in-Progress session 2

7.00 PM

Understanding the impact of textual features during listening: An on-going neuroimaging and eye-tracking study

Ester Dominguez Lucio, Nanyang Technological University, Singapore

Vahid Aryadoust, Nanyang Technological University, Singapore

Listening comprehension can be defined as the process of extracting meaning from an oral or/and visual stimuli. In this process there are bottom-up and top-down processes. During these processes, listeners decode the message into smaller segments and then build up the message by recoding the segments together. Moreover, the listener incorporates his/her own knowledge from previous experiences and cognitive processes that help to recreate a mental representation of the message. Oftentimes, this representation is created with little difficulty for English as L1 speakers.

Traditionally, the listening comprehension ability is assessed using English listening tests. The main assumption based on which these tests are developed is that they tap into an array of listening processes. However, there is little evidence to show what constructs listening assessments measure and how the features of listening texts affect listeners' processes under assessment conditions. To address this research gap, we employed advanced technologies in a study with an experimental design in which we controlled extraneous variables. Under the controlled conditions of the study, we measured the effect of textual features on listeners' neurocognitive processes (by using neuroimaging technology) and biological processes (by using eye tracking technology).

To our knowledge, there is little experimental research in which the effect of textual factors on listener' cognitive processes are examined. The aim of this on-going study is to incorporate and triangulate evidence from functional near-infrared spectroscopy (fNIRS), test scores, and eye-tracking technology to address this gap. We examined the test scores, brain activity patters, and gaze behaviour of a group of university students across two lectures that had similar test items but different textual features. Both of these tests require test takers to read and answer the questions while the audio text is played. The evidence collection phase of the study is completed and we are in data analysis phase.

We believe understanding the cognitive processes of listeners through the triangulation approach invented in this study will yield a profound understanding of listening under assessment conditions. The results would also generate knowledge about the role of textual features in listening, which can be translated to test development guidelines for teachers and curriculum developers and ultimately help students to face challenges of listening.

7.15 PM

Dynamic assessment of academic reading: Peeping into a corpus of annotated texts

Bassey Antia, University of the Western Cape, South Africa; University of Education Heidelberg, Germany

Karin Vogt, University of Education Heidelberg, Germany

Text annotations are literacy practices that are common in the reading experience of university students. Annotations may be multilingual, monolingual, or multimodal. Despite their enormous diagnostic potentials, they have not been widely investigated as a modality for both describing and assessing the cognitive processes that attend reading. In other words, there has been limited exploration of the heuristic value which signs (verbal and non-verbal) inscribed by students on texts have for both descriptive and prescriptive diagnostics of their reading practices. We report on work in progress that explores the possible role of annotations in reading diagnostics and assessment. On the basis of a corpus of text annotations obtained from teacher trainee students (n=7) enrolled in a German university, we seek to understand and evaluate what different students attend to while reading, their problem-solving strategies, the role of language and other semiotic systems, and their level of engagement with text. The study is undergirded by two notions, viz. multisemioticity and text movability. The former underscores the interaction as well as the transformation of a range of verbal and non-verbal resources in communication (Antia & Mafofo, in press; LeppÃ¤nen & KytÃ¶lÃ¤, 2017). The latter accounts for the ways individuals read or relate to texts and 'express their understanding' (Liberg et al 2011: 79; Halleson & VisÃ©n, 2018). We conclude by framing the findings within dynamic

assessment (Davin, 2016; Poehner, 2008) and by exploring their methodological implications for university-level teaching and assessment.

7.30 PM

The use of Peer Assisted Learning/Mentoring (PAL/M) as an inclusive peer assessment strategy within foundation year practice

Eleni Meletiadou, London South Bank University, UK

Higher education (HE) has gradually moved away from an elitist and exclusive mindset (based on power and privilege claims) and towards a more democratic and inclusive mindset (based on justice and human rights claims). In considering the need for HEIs to put into place support mechanisms to assist students adapt to university, it is important to take account of arguments that most students' failure or withdrawal tends to reflect difficulties in adjusting to the environment rather than intellectual problems (Wang, 2016; Yan & Berliner, 2013).

Peer Assisted Learning/Mentoring (PAL/M) has gained much importance in educational learning and educational research (Hansman, 2012; Hilsdon, 2014; Topping, 1998). PAL/M provides a setting for students to collaborate in discussing and solving problems, working through examples, reviewing the content of lectures and providing feedback (Arendale, 2020; Capstick & Fleming, 2002; Tonna, Bjerkholt & Holland, 2017). Whilst PAL/M enhances student learning by promoting a deep approach to learning in which individual students are able to develop high level cognitive skills, it also provides the medium through which students are able to develop as independent learners (Garcia-Melgar, 2018; Wallace, 1997) offering and receiving emotional support.

The current study is conducted with the aim of exploring the effectiveness of a well-developed university-based PAL/M implementation design to promote a peer assessment strategy which will foster writing improvement and student well-being amidst the Covid-19 pandemic. Bearing this in mind, 100 Foundation Year (FY) students at London Met University and 100 Foundation Year students at London South Bank University have been working in groups of four for two semesters providing support to each other. They have been encouraged to complete a variety of assignments providing online peer feedback to their fellow students in their groups. A questionnaire has been used to explore 200 FY students' attitudes towards PAL/M before and after an online teaching/learning intervention employing PAL/M to foster inclusivity and help diverse student cohorts. Students' writing performance has also been assessed before and after the intervention to explore its impact on students' writing performance. The study outcomes are anticipated to make a significant contribution to the context field of education by providing a rigorous evaluation of the impact of the use of PAL/M as an inclusive strategy in FY practice. Recommendations for training lecturers and students in PAL/M and for implementing PAL/M successfully in FY courses as well as suggestions for future research will be provided. In response to the need for more information, this study will contribute a students' voice regarding the use of PAL/M as an inclusive strategy within FY practice which in so far has been absent (McCarthy & Armstrong, 2019).

Symposium

3.30 PM

Cathie Elder: Framing the LTRC's policy contribution in the languages area

Ute Knoch: The challenges of providing expert advice in policy contexts

Jason Fan & Jin Yan: Navigating tensions between language testing intentions and policy imperatives: The case of the College English Test (CET) in China

Kellie Frost: Negotiating the boundaries of English: the role of tests and language testers in Australia's skilled migration policy space

Discussants: Tim McNamara & Joe Lo Bianco

Closing remarks: Lesley Stirling (Head of School, School of Languages and Linguistics)

Followed by LTRC anniversary video

LTRC 30-year anniversary symposium: Mapping tensions between language testing, policies, and practices

Language tests are widely used in Australia and elsewhere to implement institutional policy, to serve educational goals and to regulate transitions into study and employment. The different roles testing activities can serve in these contexts and the fairness and justice implications of using or misusing language tests for such consequential purposes have been the focus of growing attention in recent years. Accordingly, the need to more effectively engage with policy makers, employers, educators, language learners and test takers to better understand these implications has become ever more pressing. A series of 10-minute presentations by staff and associates of the Language Testing Research Centre (see titles and authors below) will reflect on how language testers engage with policy, the different ways in which test constructs mediate policy, the conflict between policy imperatives and test developers' intentions, and how test scores are interpreted and reshaped in practice by different test users. Two discussants and founding fathers of the Centre, Professors Joe Lo Bianco and Tim McNamara, will respond to these presentations and consider their contribution to current debates in the fields of language assessment, language policy and applied linguistics more generally.

Framing the LTRC's policy contribution in the languages arena

Cathie Elder

The LTRC, as a self-funding Centre, responds by necessity to policy shifts, whether these occur at broader societal or more local institutional and program levels. Its language testing activities also have the potential to influence policy, whether directly or indirectly. Focussing on examples of the Centre's work in languages education, this talk considers the diverse orientations adopted (i.e. to inform, enact or evaluate policy). I speculate about the policy impact in each case, highlighting complexities encountered and concluding with recommendations for policy responsible language testing.

The challenges of providing expert advice in policy contexts

Ute Knoch

Language testers can have various roles in relation to the use of language tests and policy. One such role may be to provide expert advice in policy formation or policy review contexts. Such instances are often not documented systematically, as advice may be provided in informal or closed meetings, or confidential documents, which are not available to the public. The challenges associated with providing advice are also rarely recorded. In this paper, I describe three instances in which the Language Testing Research Centre was invited to provide external policy advice. I specifically reflect on how we came to be invited to provide advice, what advice we were asked to provide, the complexity of providing advice in each instance, and whether the advice was taken up by the policy makers. The paper concludes with implications for training new language testers.

Navigating tensions between language testing intentions and policy imperatives: The case of the College English Test (CET) in China

Jason Fan & Yan Jin

In this talk, we demonstrate the conflicts or tensions between the intended uses of a language test and the actual uses, misuses and even abuses of the test in the real world, driven by various policy imperatives. We use the College English Test (CET), an English proficiency test targeting non-English major undergraduates in China's universities, as an example. Our analysis of the CET was guided by Knoch and Elder's (2013) framework for evaluating post-entry language assessment and revealed that the CET has been used for several purposes unintended by the test developer both in the educational community and the social domain, due to different policy imperatives. To address such issues, we argue that continuous professionalisation of language testing is instrumental in building an ethical milieu, which in turn can help to promote more responsible uses of test results.

Negotiating the boundaries of English: the role of tests and language testers in Australia's skilled migration policy space

Kellie Frost

While the importance of accounting for the use of tests as policy instruments is by now widely acknowledged, validation and fairness frameworks in language testing rest on core assumptions which, I argue, are often incommensurate with the ways in which policy intentions are articulated and pursued. This is especially the case in the context of Australia's skilled migration policy, where tests are embedded within a range of selection processes across various intersecting layers of policy, each determined by different stakeholders with different needs and intentions. In this paper, I will discuss the roles tests play in this complex policy setting, from policy maker perspectives, and from migrant perspectives as they respond to their entangled test-in-policy experiences. Highlighting the inevitable 'validity chaos' this policy context engenders for language testers, I conclude by arguing for a renewed criticality in language testing, where we expand our lens beyond evaluations of how well, if at all, test uses align with notions of best and ethical practice in our field, to engage with the wider discursive space within which problems of language and of policy are imagined, and to interrogate the ways in which we, as a discipline, are part of a material-discursive apparatus that produces idealised language users, workers, and citizens, together with the various inclusions and exclusions this entails.

Friday, 20th November (all times are in AEDT, GMT+11)

Papers

9.30AM

Impact of test-taker's background on score gain on IELTS

Okim Kang, Northern Arizona University, USA

Hyunkee Ahn, Seoul National University, South Korea

Kate Yaw, Northern Arizona University, USA

Soh Yon Chung, Seoul National University, South Korea

Much research has examined the relationship between test scores and academic performance for its predictive validation (e.g., Hill, Storch, & Lynch, 2000), or task difficulty and the rating process (Brown, 2006). Undoubtedly, individual differences or learner backgrounds (e.g., FL learning experience, language proficiency, or motivation) may be predictors in how successful a language learner will be at learning a new language (Brecht, Davidson, &Ginsberg, 1993). However, studies on how test-takers' individual characteristics predict their score gains from a longitudinal perspective have been rare. The current study investigated to what extent IELTS test performances (i.e., overall test scores and speaking section scores) changed over the period of 3 months. It further examined how learner background variables (e.g., hours of study, use of target language, and proficiency, program attendance, self-assessment) affected band score gains on the IELTS. Fifty-two Korean students, enrolled in IELTS preparation classes, participated. Participants' proficiency levels were determined by their in-house placement test scores (i.e., roughly 16 beginners, 17 intermediate, and 19 advanced). Once participants completed the pre-test survey, they took the pre-arranged official IELTS test. Participants' hours of study and target language use information was collected weekly. The post-survey and online interviews were conducted at the end of the 3-month period right after the official IELTS post-test. The primary analyses were regression models and linear mixed-effects models which treated students as random effects, learner background variables as covariates, and the IELTS performance gain scores as a dependent variable.

The results showed that students made various progress in English over the 3-month period with an average gain of slightly less than half a band (.3) and with the most score gain in the writing skill and the least score gain in the speaking skill. That is, the 12 weeks of intensive study did not make a huge difference to performance particularly in an EFL context although its change was still statistically significant with a small effect size. In addition, the mean gain score on Test 2 decreased as the students' proficiency increased, indicating that some type of temporary regression in the longitudinal process. In particular, hours of study and level of proficiency predicted the band score gains most potently. Together with the amount of target language, the background variables explained 34% of variance in the score gains. Findings suggest that a language proficiency gain does require an invested time commitment, possibly more than one often thinks. Self-reported target language use and language contact were not associated with enhanced performance among this group of participants. However, the degree of program attendance made a significant impact on the IELTS gains. Also, participants'

attitudinal and motivational factors played a role in their score gain process. Findings offer useful implications to the development of language testing and assessment as well as curriculum planning.

10.00 AM

Stakeholders' perceptions of IELTS speaking and writing tests and their impact on communication and achievement

Noriko Iwashita, The University of Queensland, Australia

With the rapid movement to globalisation, a good command of English as the international language of the world has become essential. In response to this trend, the current English language curriculum in Japan has been revised to focus on enhancing the four skills (i.e., listening, speaking, reading and writing). In 2013, the Japanese Ministry of Education proposed that all universities should accept the results of public standardised proficiency tests, such as Eiken, IELTS, and TOEFL. Although the number is increasing, compared to TOEIC and Eiken, the number of IELTS test-takers is still relatively small. However, with the government initiative, interests in IELTS and the number of test-takers are expected to grow.

In English language assessment research, there has been a growing emphasis on the importance of considering the views of stakeholders, such as students, teachers and administrators, to understand the social impact, current market focus and ethical implications of the test (e.g., Coleman et al., 2003; Hyatt, 2013). Consequently, many studies have investigated the impact IELTS has on IELTS preparation courses and academic success by collecting stakeholder perceptions. Most studies, however, focused on IELTS writing tests or IELTS test performance in general, but few studies have examined IELTS speaking tests. Students are required to possess good writing and speaking skills to succeed in their academic study and beyond, and also different skills, knowledge and strategies for successful communication in writing and speaking.

For that reason, the current study investigated Japanese stakeholder perceptions of IELTS writing and speaking tests, and their impact on communication and achievement in a given context. In particular, this study explored the level of familiarity with IELTS among high school teachers, university lecturers and first-year university students, and how IELTS is perceived among these stakeholders, with special attention paid to the impact of its inclusion in university entrance exams in the near future. The study allows further insights to be gained from these perceptions into the usefulness of IELTS in terms of students' readiness for L2 communication in an academic setting.

In the study, stakeholder perceptions of IELTS were assessed via a questionnaire survey (N=96) and semi-structured interviews (N=21). The preliminary findings show that while most teachers are familiar with IELTS, some student participants had never heard of IELTS until they participated in the survey. Further, the majority of the participants found both Speaking and Writing Tasks challenging and beyond what current high school students are required to do in the English curriculum. Nevertheless, they consider IELTS test tasks to assess written and oral communication skills adequately, but their views on its inclusion of IELTS in university entrance exam are mixed. The paper will discuss stakeholdersâ€™ insights into both the possible washback of the inclusion of four-skill-tests such as IELTS in high-stakes nation-wide exams (e.g., Sasaki, 2018) and the preparation of students for IELTS examinations and beyond.

12.30 PM

Expanding the interactional competence construct: Criteria from domain experts on interactional success in everyday life

David Wei Dai, Monash University, Australia

The assessment of interactional competence (IC) has drawn from findings in Conversation Analysis (CA), primarily focusing on the mechanics of interaction such as turn-taking and topic management (May et al., 2019). Though the incorporation of CA-informed IC markers broadens the traditional psycholinguistic Leveltian speaking conceptualisation (Levelt, 1989), IC, and communicative competence in general, encompasses a richer repertoire of skills, such as the volitional, attitudinal and other non-cognitive aspects of communication, which were components in Dell Hymes's original model of communicative competence (Hymes, 1972; McNamara, 1996; Sato & McNamara, 2018). This paper aims to make explicit the assessment criteria that L1-Chinese domain experts use to evaluate test-takers' IC. A data set from 22 test-takers' performance on a 9-item role play IC speaking test was utilised as stimuli to elicit domain experts' assessment criteria. The 22 test takers included 11 L1-Chinese speakers and 11 L2-Chinese speakers, which served to investigate if markers of linguistic competence were oriented to by the experts. The domain experts recruited were L1-Chinese speakers who had no linguistic or language teaching backgrounds but were everyday members of the Chinese society (Schutz, 1962), and hence had vast lived experience and expertise in the everyday language use domain as captured in the test. The experts listened to test-takers performances, written down features contributing to the success or lack of success in test-takers' management of interaction, and participated in focus-group interviews with the researcher to further elaborate their indigenous criteria of IC (Knoch & Macqueen, 2019). Both domain experts' written comments (47,162 words) and interview transcripts (229,806 words) were collated, thematically coded and collapsed into categories. The top five categories were selected as rating categories for this test: disaffiliation management, reasoning (logos), affiliative resources (pathos), morality (ethos) and role enactment. Though the five categories differ greatly from existing rating criteria frequently used in speaking assessment, they represent what sociologists consider everyday members' orientation towards and interpretation of the natural attitude in a shared life-world, where 'inter-action'□ between members of the life-world takes place (Garfinkel 1967; Schutz 1962). The five categories have been theorised in different forms in sociological and sociolinguistic theories such as positive and negative politeness strategies (Brown & Levinson, 1987), role categories and category-bound activities (Sacks, 1972), the Hymesean perspective of communicative competence (Hymes, 1972), moral order in social interaction (Habermas, 1981; Garfinkel 1967, Goffman, 1959; Wittgenstein, 1953) and tracing further back, the three modes of persuasion in Aristotelian rhetoric (Aristotle, 2006). Written comments and interview transcript excerpts from domain experts, coupled with test-takers' performance data were used to locate the five rating categories in empirical data, demonstrating how the sociolinguistic-interactional rating criteria (Roever & Kasper, 2018) were in keeping with theories on interaction and how they could be operationalised in IC assessment. Findings from this study have implications for IC construct definition and the extrapolation inference of IC assessment, as IC rating criteria should be congruent with what competent real-world everyday users of the language consider crucial to interactional success.

1.00 PM

The effect of public speaking anxiety on the public speaking performance of tertiary-level students

Tingting Liu, Sichuan International Studies University, China

Vahid Aryadoust, Nanyang Technological University, Singapore

Previous studies have shown that public speaking anxiety can lead to the avoidance of speaking situations and impairment of speaking performance (Glassman et al., 2016). Speakers' low self-image and past negative experience render speaker to self-focused while speaking and thus tax their working memory and disturb their mental process (Bodie, 2010). Additionally, public speaking anxiety can be intensified under assessment conditions when facing audiences of greater expertise such as the raters (Bodie, 2010). As such, it can be a potential source of construct-irrelevant variance and threaten test validity. However, there is a dearth of studies on public speaking assessment.

Thus, this study investigated the relationship between public speaking anxiety and public speaking performance under classroom assessment condition. 74 English learners from a foreign studies university delivered a 3-minute speech facing real audiences and answered Personal Report and Confidence as a Speaker questionnaire (PRCS, Paul, 1966). Their speeches were rated by two experienced raters using Public Speaking Competence Instrument (PSCI, Thomson & Rucker, 2002), which consists of five subscales (introduction, body, conclusion, delivery, global competence).

We subjected the ratings for all five subscales of PSCI and the total scores to Many-facet Rasch Measurement and investigated the impact of facets including student's speaking ability, rater severity and item difficulty. Only valid data with infit and outfit mean square indices lower than 1.5 were retained for correlation analysis. The following linear regression analysis showed that there was a moderate correlation between speaking anxiety and speakers' delivery (r = -0.324, p ï¼œ 0.005), global competence (r = -0.374, p ï¼œ0.001), and total score (r = -0.373, p ï¼œ 0.001).

It was proposed public speaking assessment among tertiary-level students should be learning-oriented and cognitive therapy such as cognitive preparation and video feedback be provided to test takers before and after assessment to enhance their speaking skills and reduce their speaking anxiety.

1.30 PM

A comprehensive review of research on automated written feedback

Huawei Shi, Yantai University, China; Nanyang Technological University, Singapore

Vahid Aryadoust, Nanyang Technological University, Singapore

Originally designed to generate summative scores for written essays in high-stakes tests, automated writing evaluation (AWE) systems were increasingly revised and readapted to contribute to classroom instruction and assessment by providing automated written feedback.

This research aims at informing such pedagogical practices and formative assessment with automated feedback generated by Automated Writing Evaluation (AWE) systems. We conducted a comprehensive review of previous research on automated written feedback within argument-based validation framework.

Scopus was used as the main database, and 77 papers were identified from 2010 to 2020. The main findings are: i) the domain description inference was totally ignored in studies of automated feedback, and most research focused on the utilization inference of automated feedback, indicating a strong tendency toward pedagogical purpose. In the meantime, the generalization, explanation, and extrapolation inferences of automate feedback was also overlooked areas, with very limited relevant research being identified; ii) For analysis methods, descriptive statistics, t-test, and qualitative analysis were the most frequently used methods across all studied inferences; iii) For areas of investigation, most studies yielded backing in related areas; nevertheless, rebuttals also were discovered in almost all studied inferences. It also should be noted that the threshold for the evaluation inference of automated feedback was not unanimously set, which might comprise this inference.

Based on these findings, this research suggests that prior to the utilization of automated feedback, more research should be done in regard to the other inferences. Specifically, in terms of the domain description inference, the target language use (TLU) domain of automated feedback should be explicitly discussed; for the evaluation inference, more research should be done to evaluate and improve the accuracy of automated feedback, which might greatly influence the effect of utilizing automated feedback in classroom environment; the construct of writing of interest in regard to automated feedback also deserves attention. Finally, for analysis methods, more experimental design is strongly recommended.

3.00 PM

Moderation as a mechanism for collecting nationally consistent EAL/D data inclusive of Indigenous EAL/D learners

Denise Angelo, The Australian National University, Australia

Catherine Hudson, The Australian National University, Australia

This paper proposes the mechanism of moderation for developing nationally consistent data on learners of English as an Additional Language/Dialect (EAL/D) in schools. In the contemporary education policy as numbers environment (Lingard, Creagh & Vass 2012), the lack of data on EAL/D learners impacts in multiple ways on this learner cohort and their teachers. Their invisibility in data sets accessed by education policy makers and services may wrongly suggest that EAL/D learning needs are of little consequence. This certainly runs counter to TESOL research and professional experience, and indeed to espoused inclusive education policies, and in the case of Indigenous EAL/D learners, the goals of the national Closing the Gap agenda. Yet, after some decades of EAL/D proficiency tool development we have local tools fit for local cohorts and local processes, so a one-size fits-all approach looks counter-productive. Instead, we propose a moderation, needs-based approach closely following a solution proposed by Sharma et al (2017) for a different education domain requiring nationally consistent data. We make the case that such an approach should take into account Indigenous EAL/D learners whose language backgrounds and mainstream curriculum learning contexts are often overlooked in EAL/D assessment policy and accountability proposals. This paper outlines the need for a representative body of EAL/D data of the full EAL/D learner cohort and the rationale for this moderation approach, along with a suggested process and moderation template. We hope it encourages productive discussion in the field.

3.30 PM

Examining learner perception toward test feedback in classroom-based speaking assessment using a validity framework

Rie Koizumi, Juntendo University, Japan

Akiyo Watanabe, Utsunomiya Minami High School, Japan

Makoto Fukazawa, University of Ryukyus, Japan

Chihiro Inoue, Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire, U.K.

Effective test feedback is essential for learners to be able to use test results for their future learning and for test developers and users to ensure the high validity of score interpretation and use (e.g., Sawaki & Koizumi, 2017). While the importance of formative test feedback has been emphasized in classroom assessment (e.g., Cheng & Fox, 2017), how test feedback that appears in score reports of second language (L2) tests affects students' perceptions and reactions does not seem to have been sufficiently examined (see Ducker et al., 2013, for an exception). Furthermore, test-takers' use of score reports can be utilized as evidence for validity of test interpretation and use, but there have been few cases in which such a use of score reports has been incorporated in a validity framework. This study reports on such an attempt, using Chapelle and Voss's (in press) argument-based validity framework, where learner perceptions toward the test feedback in a score report are investigated in a consequence implication inference. The warrant for the consequence implication inference is that test scores and score reports are used to improve learning. This warrant is based on two assumptions, the test scores and score reports (a) are clearly interpretable to test takers to obtain information for future learning and (b) positively impact learning. The current study focuses on the former. Backing for this assumption is collected through questionnaires from students.

A total of 116 students took six low-stakes speaking tests over one year in a required L2 English course at an upper secondary school in Japan. The tests were conducted in class, taking a variety of forms including a face-to-face teacher interview, a prepared presentation, and a pair or group discussion. Students' oral performances were double scored on the spot or after class using a recording. Score reports included descriptive feedback consisting of test scores, scoring rubrics, and study advice. After each test administration, score reports were given to the students and a teacher explained how to read the reports, the general tendency of students' speaking performances from the test, and advice for future learning. Students then answered a questionnaire composed of 5-point Likert scale questions and open-ended questions about how they perceived the test and score report. Preliminary analysis suggests that students perceived test feedback positively, students comprehended the score report along with the teacher's explanation well and were motivated to study English more and encouraged to study further based on the advice in the score report. The results are positive backing for the assumption that supports the warrant and the corresponding inference. Results of students; written responses thematically analyzed are to be reported in the presentation to explore why they responded positively, neutrally, or negatively and to consider possible revisions in the speaking test tasks, procedures, scoring, and provision of feedback. Implications are also discussed for future practices, in particular types of feedback students prefer in score reports in classroom-based speaking assessments.

Work-in-Progress session 3

10.30 AM

Developing a marking rubric for the integrated reading-writing test

Aynur Ismayilli Karakoc, Victoria University of Wellington, New Zealand

Integrated writing (IW) tasks are gaining popularity because they are claimed to be more authentic (Plakans & Gebril, 2012; Weigle & Parker, 2012). Usually, such test tasks include reading multiple sources or reading and listening to sources and writing an essay. One example of this test can be the integrated writing section in the TOEFL iBT test.

The original study is about the development and validation of an IW test task. The study consists of three phases: conducting needs analyses, designing the test and marking rubric, and validating the test. The test's purpose was to understand the first-year international students' academic preparedness at the reading-writing intensive courses at the Faculty of Humanities and Social Sciences, Faculty of Education, Faculty of Law, and Wellington Business School at the Victoria University of Wellington (VUW). The test asked the students to read two pieces of texts and write an essay. For this presentation, the design process of the marking rubric will be reported.

There are various approaches available in designing marking rubrics, including theory-driven, empirical, and data-driven methods. Although research is expanding on IW test tasks, the design process of marking rubrics has not been largely addressed. This research in progress study will report how the test rubric was designed based on the target language use domain (TLU) analyses (Bachman & Palmer, 2010).

The data came from three sources: artefacts, a questionnaire, and interviews. Artefacts included course outlines of 63 courses and marking rubrics used by course coordinators (faculty teachers). The survey included a Likert scale questionnaire asking about the teachers' requirements for first-year level essays covering reading for writing, source use integration, content, organization, and language features. Finally, the interviews with eight faculty teachers presented in-depth views about their expectations for the students' essays.

Based on these analyses, the draft of the rubric was sketched. Additionally, EAP (English for Academic Purposes) teachers' ideas were also sought regarding the rubric descriptors and students' essays. I am now working on the draft version of the rubric.

10.45 AM

The measurement of implicit linguistic knowledge: a meta-analysis

Viola Lan Wei, The University of Auckland, New Zealand

There is wide consensus that the ultimate and most highly prized goal of second language acquisition is spontaneous and unreflected L2 use, which is underlined by implicit linguistic knowledge (Sharwood Smith, 1981). Some people believe that implicit knowledge can only be acquired, rather than learnt, within the hypothesised critical period of language acquisition, while others maintain that explicitly learnt knowledge can be proceduralised into implicit knowledge given sufficient practice (Ellis, 2008). This debate on the theoretical assumptions of implicit knowledge is partly due to a lack of consistency in

how implicit knowledge is operationalised in empirical research, thus leading to difficulties in understanding how implicit knowledge is developed.

In the past decades, a considerable amount of studies that are relevant to the measurement of implicit knowledge have accumulated. Some of these studies aim at validating one or more proposed measures of implicit knowledge, while others utilise multiple measures for the purpose of obtaining some representation of implicit linguistic knowledge. For example, the most widely reported measures include elicited imitation tests, timed grammaticality judgement tests, and the word monitoring test, to name a few. However, there are considerate discrepancies in the way that implicit knowledge is conceptualised in different studies and the way that these instruments are designed and administered. These discrepancies could potentially lead to significant differences in how research findings are interpreted, although all of these studies share the same aim of accessing implicit linguistic knowledge.

Against this backdrop, a systematic review of past research findings is warranted. The present study set out to conduct a meta-analysis to understand how implicit knowledge has been measured to date in relation to the design features of different studies. Published research that have included measures of implicit knowledge from the past five decades will be included in this meta-analysis. Factors that will be under investigation include overall design of research, i.e., experimental vs. exploratory, and test modality, i.e., written vs. aural; timed vs. untimed; lingusitic vs. extra-linguistic. Results of this meta-analysis will provide information about the extent to which the test design features influence assessment scores of implicit linguistic knowledge. This knowledge will add to the current understanding of the validity of implicit knowledge measures, and will give implications for how these measures should be implemented in the future.

11.00 AM

Assessing speaking in a Post-Entry Language Assessment

Liz Kose, The University of Auckland, New Zealand

Spoken communication skills are gaining currency both in education and the workplace and are specifically mentioned in university graduate profiles in Australia and New Zealand. Speaking skills are also important to students with English as an additional language, and are often a source of anxiety, and unwillingness to participate in discussions or approach teaching staff for help. University Post Entry Language Assessments (PELAs), however, rarely assess speaking skills. The University of Auckland's PELA, the Diagnostic English Language Needs Assessment (DELNA), for example, includes academic writing, reading, and listening only. The lack of a speaking assessment means that the construct of academic language proficiency is underrepresented. Therefore, to support students deemed at risk of under achievement in a more comprehensive way, the inclusion of a speaking assessment is necessary.

This presentation reports on the design and small scale trial of a diagnostic speaking assessment which was developed as postgraduate coursework for a language testing course. The test is a semi direct, computer-mediated, human rated performance test. It comprises three parts, in which the tasks increase in complexity, moving from very familiar topics, such as personal information and opinions, to less familiar topics. One task requires an extended monologue with preparation time. To allow for detailed feedback with a finer grain than proficiency tests, multi trait analytic rating scales and detailed marksheets were developed. The trial was carried out with six learners at B2 level, and three raters, all

of whom were experienced teachers. Based on the ratings, participants were given written feedback and advice on language development.

The usefulness of the diagnostic speaking assessment was analysed using the assessment use argument designed for PELAs by Knoch and Elder (2013), and Bachman's (2010) assessment use argument. The small scale of this trial allows warrants to be discussed to support the evaluation inference in Knoch and Elder's framework, but the higher-level inferences require a larger sample, further trials and multiple test forms. Notwithstanding the limitations, the initial results were promising. Feedback from test takers and raters about the content and format of the assessment was positive, and diagnostic feedback was well received. However, the marksheets were not fully utilised by raters, and some descriptors on the rating scales were difficult to interpret. Revisions of the marksheet and rating scales are therefore needed, and multiple forms of the test tasks should be developed. As mentioned above, a larger trial is also required to make sound claims as to the usefulness of the assessment and its potential inclusion in a PELA programme such as DELNA.

11.15 AM

Investigating the pedagogical usefulness of Automated Writing Evaluation (AWE) System in academic writing instruction

Leila Zohali, The University of Melbourne, Australia

The impetus for the current study emerges from the concern that the past decade has seen the rapid development of automated writing evaluation (AWE) tools and their incorporation in many educational settings, but it is still difficult to determine the efficacy of automated feedback. Results from previous studies on the pedagogical effectiveness of AWE tools in writing classes (Riedel, Dexter, Scharber, & Doering, 2006; Saricaoglu, 2019; Schroeder, Grohe, & Pogue, 2008) seem to be far from conclusive. This is mainly due to the limited number of studies available, flaws and differences in the research designs (including variation in writing tasks used, contexts and participants). Furthermore, most of the research on automated feedback has focused on the writing product with little attention being paid to the process of revision and students' engagement with automated feedback (Stevenson & Phakiti, 2014). For this reason, scholars have called for well-designed studies focusing on the process of revision and engagement with AWE feedback (Storch, 2018; Stevenson & Phakiti, 2014).

The present study aims to investigate the efficacy of incorporating automated written corrective feedback, provided by CriterionÂ®, into language classes. In particular, it aims to (a) investigate the effectiveness of using AWE feedback on writing improvement in both revised and new subsequent drafts, (b) compare the effectiveness at two different proficiency levels, (c) compare students' engagement with teacher and automated corrective feedback at two different proficiency levels and, (d) investigate student and teacher perceptions of the automated feedback. Aimed to be an exploratory research project, this study adopts a mixed-method approach with a sample of about 53 participants, including 48 English as a second language (ESL) learners and five teachers at different language schools.

The quantitative part of the study, which is students' writing improvement using Criterionïƒ', will be captured by using Complexity, accuracy and fluency (CAF) measures in pre, post, and delayed pot-tests and new and revised drafts in three sessions. Linguistic accuracy is measured as error counts in the writing tasks. Syntactic complexity is calculated as the number of clauses per T-units (C/T). Fluency

measure is excluded because it deals with the length of the essay (number of words), yet, the participants are expected to produce essays of the same length in pre, post and delayed post-tests.

The qualitative part will go deeper to gain insight about how students interact with automated feedback and how Criterionïƒ' feedback is perceived by using think-aloud protocols (TAPs), retrospective individual interviews and individual end of course interviews. Student engagement with teacher and CriterionÂ® feedback will be coded by drawing on Fredrick, Blumenfeld and Parisâ€™ (2004) model of engagement, which was later applied to written corrective feedback (CF) by Ellis (2010). Data from TAPs and retrospective interviews will be coded and analyzed thematically based on the themes emerged from the pilot study and informed by research questions. Students' and teachers' perceptions will be captured by end-of-course interviews and will be analyzed qualitatively.

In this session I will present my initial results and ask for some feedback from audience.

Key words: automated writing evaluation (AWE), accuracy improvement, engagement, perception

11.30 AM

Integrating Post-Entry Language Assessments into the Curriculum. What works well?

Tracy Ware, Edith Cowan University, Australia

Andrew Kelly, Edith Cowan University, Australia

Edith Cowan University (ECU) provides a comprehensive approach to the development of the communication skills of commencing students through the implementation of a short diagnostic assessment of student writing, the Post-Entry Language Assessment (PELA). This mandatory task for all coursework students provides a mechanism for early identification of 'at-risk' students who may struggle with the demands of tertiary study. A main aim of the PELA is for students to receive specific individualised feedback on their English language proficiency in order to make informed decisions about developing their respective language skills. To be truly diagnostic, the link between the feedback, intervention and support is crucial.

This paper presents the results of the larger scale application of an initiative trialled at ECU in 2019; that is, the integration of a PELA into early online low stakes assessments in two units. As a university-wide PELA, this second phase explores the expansion of the integrated PELA into a range of early assessment tasks over several schools, which were marked using different tools. This work in progress presents a comparison of PELA completion rates, insight into the marking experience and perceived value of incorporating the PELA this way through qualitative feedback, which was collected from the markers, lecturers and students involved. It demonstrates how, by embedding the PELA into an existing assessment task, students are able to reflect on the feedback provided allowing the PELA to act as a formative feedback tool rather than just a standalone diagnostic assessment. Finally, this paper offers unique insights into which kind of assessment tasks and marking methods were considered to be the most effective in providing relevant feedback to students, and also offering a sustainable marking experience for both the PELA markers and the Unit lecturers.

11.45 AM

A sociocultural analysis of teacher assessment literacy development: The promises and pitfalls of taking a Vygotskian perspective on assessment

Xuan Minh Ngo, The University of Queensland, Australia

Teacher assessment literacy has recently received considerable interest in the language assessment community as evidenced by the rising number of related publications, particularly those in the special issues of Language Testing (2013) and Papers in Language Testing and Assessment (2017). This growing body of literature has demonstrated that teachers are often poorly prepared to perform their assessment duties and that teachers tend to acquire their assessment literacy on the job (Crusan, Plakans, & Gebril, 2016; DeLuca & Johnson, 2017; Lam, 2015; Lam, 2019; Vogt & Tsagari, 2014; Xu & Brown, 2016). Nevertheless, how teachers' experiential learning of assessment unfolds remains relatively under-studied (DeLuca & Johnson, 2017; Xu & Brown, 2016; Yan, Zhang, & Fan, 2018). It is this crucial gap that my PhD project aims to bridge. Since recent studies (Inbar-Lourie, 2017; Yan et al., 2018; Lam, 2019; Xu & Brown, 2016) suggest that teacher assessment literacy development is a complex and context-bound process involving numerous sociocultural factors, I have adopted Vygotskyâ€™s sociocultural theory, particularly the genetic method and the concept of perezhivanie (lived experience) as the project's theoretical framework. Guided by these two sociocultural theory tools, my project has sought to investigate significant lived experiences (perezhivaniya) that mediated the participating teachers' on-the-job learning of assessment, the interaction of environmental and personal factors in those significant experiences (perezhivaniya), and the often overlooked emotional dimension of the assessment literacy development process. Regarding methods, the study has employed the multiple case study design that involves multiple sources of data (interviews, classroom observation, stimulated recall, and think-aloud protocol) collected from six novice Vietnamese EFL teachers in two phases with each lasting three months (the first phase in 2019 and the second scheduled for 2021). The data will first be analysed simultaneously with data collection as recommended in Merriam and Tisdell (2016) and then re-examined when the entire data gathering process has concluded. As such, the analysis process will start inductively and become progressively deductive, and within-case analysis will be conducted first, followed by cross-case examination. Once completed, this project will add to the limited literature that examines teacher assessment literacy development from a sociocultural perspective and offer crucial implications for education of both pre-service and in-service language teachers. In this work-in-progress presentation, after a quick overview of the entire project, I will showcase the promise of a sociocultural perspective on teacher assessment literacy with some preliminary findings from the first round of data conducted in late 2019, discuss some challenges of framing teacher assessment literacy from a sociocultural perspective, and elaborate on my initial attempt to address those challenges.

Teacher sharing session

4.00 PM

Assessing learners of Indigenous languages - balancing perspectives

Cathy Bow, Charles Darwin University, Australia

Susy Macqueen, Australian National University, Australia

Language assessment in the context of Indigenous Australian languages has focused on assessing the English language proficiency of Indigenous students (speakers of Aboriginal English, contact languages or traditional Indigenous languages). The assessment of learners of Indigenous languages is a relatively recent development. In this effort, it is important to incorporate the perspectives of language owners, both for their linguistic expertise, as well as for the values that underpin assessment. In this presentation, I report on the process of assessing non-Indigenous learners studying an Indigenous Australian language (Bininj Kunwok) in a university context. The course was supervised and co-designed with language owners, who were also involved in the assessment process as raters of speaking assessments. It was in this process that the institutional and Indigenous perspectives appeared to diverge. While the assessment of additional languages tends to prioritise constructs such as the development of fluency and accuracy on 'typical' proficiency trajectories, the Indigenous assessors tended to prioritise engagement with their language as a key criterion. As a result, the approach of the Indigenous assessors was highly encouraging, with a reluctance to be in any way critical of the learners' linguistic efforts. This case study introduces some of the incongruity between the values underpinning the assessment of endangered languages by Indigenous assessors and the expectations of institutions where assessment values have developed in the context of widely-spoken standard languages. The experience offers a basis for discussions about how to incorporate Indigenous language ownersâ€™ perspectives and those of English-medium tertiary institutions in ways that respect the needs of both.

4.15 PM

From structure to substance: scaffolding the development of argumentation in EFL academic writing

Fan Chen, Nanjing University of Information Science & Technology, China

Toulmin's model of argument (2003), which features six elements of argumentation (i.e. claim, data, warrant, backing, qualifier and rebuttal), has been widely used as a rubric for assessing the quality of argumentation. Although the inclusion of essential argumentative elements, namely the surface structure, is key to good argumentation, the quality of reasoning, namely the substance, also determines the persuasiveness of an essay yet receives less research attention. With a focus on both substance and structure, the present study explored how the use of an analytical rubric helped to improve the quality of argumentation. In the present study, 28 first-year EAP students wrote argumentative essays following a modified Toulmin model, and their first drafts underwent a peer-then-teacher evaluation before being revised accordingly. A scoring rubric was developed to assess both the structure and substance of the argumentation. Statistical analyses showed that the first draft grades given by peers and teachers were significantly correlated in multiple areas. A comparison of the first and final drafts showed that both structure and substance improved but not always in tandem. Three case studies further revealed how students revised their essays with the help of peer and teacher feedback.

4.30 PM

Upgrading critical evaluation and skill development through Group Assessment

Maha Hassan, Teaching ESL HUB, Cairo, Egypt

The modern trend of teaching usually encourages teachers to use critical thinking as part and parcel of their work in class. Student involvement in the learning process enriches their class experience and get them more interested in learning the language. This idea of critical thinking can be used at its best through student-self assessment and peer assessment. Peer assessment is a powerful meta-cognitive tool. It engages students in the learning process and develops their capacity to reflect on and critically evaluate their own learning and skill development. It supports the development of critical thinking, interpersonal and other skills, as well as enhancing understanding within the field of knowledge of a discipline.â€ Not only individual peer assessment is that important, but group assessment as well.

To apply peer or group assessment, students need 'a set assessment criteria'. This adds a lot to student confidence, bearing responsibility as well as transferring some ownership of the assessment process to theme.. aiding your students to develop judgment skills, critiquing abilities and self-awareness.

During my presentation I'd like to speak about peer and group assessment that I applied with my classes through a number of activities and projects. I will give the details of the activities used, analyse them and give the attendees the final outcome of the experience and how they can apply similar ones in their classes at different levels.

4.45 PM

Turbocharge your peer feedback with summative criteria

Gaby Lawson, M.U.E.L.C, Monash College, Australia

The proposed teacher-sharing session will look at the use of summative assessment criteria to improve learning outcomes in EAP peer review tasks. Many students seem to lack the skills or motivation to engage in peer review. However, many EAP courses now have peer review activities in their curricula so teachers need to increase their confidence in delivering it and boost the learning outcomes. The aim of the session is to help teachers empower and motivate students with peer review by understanding why and how to integrate summative criteria into the process in a learner-centred way. If teachers introduce part of the summative criteria for a task, students can negotiate the meanings of the criteria. They can then use their interpretations of the criteria to give each other peer feedback. This empowers learners to use criteria, helps them understand what they need to do to pass the course. This also encourages a Growth Mindset by putting the focus on meeting criteria rather than being a 'good' or 'bad' student. The proposed session will be 10 minutes long with 5 minutes for questions. The presentation will be given on Google Slides and a Google Doc for teachers to access during the presentation.