2021 BILC Conference

Rethinking the Rating Process:
Solution to the Threshold Performance Dilemma

Mary Jo Dibiase, USA; Jana Vasilj-Begović, Canada

When assessing constructed responses in advanced language proficiency tests, the extent of the raters' level of training and norming and their ability to foresee correct responses are critical for reliability and validity. These issues influence standard setting methods that establish meaningful candidate categories that embody the test's construct and purpose. This paper describes an approach to establishing cut scores in an integrated reading comprehension test via the skill of writing. *Retrodictive Modeling Approach* (RMA) –a mixed method approach– relies on raters' level of expertise and holistic ratings coupled with their qualitative analysis and scores. RMA yields a pattern for establishing a threshold performance level. Further research may lead to a wider use of the RMA in distinguished proficiency level testing.

2021 BILC Conference

1ˢᵗ Round

Mary Jo Dibiase, USA; Jana Vasilj-Begović, Canada

Results: Did the Native Readers test out at Level 4?

Clarification: Can you describe, in very basic terms, what was meant by *Holistic Evaluations* and *Analytical Scores*?

Follow-up: *A holistic rating alone does not seem to consistently predict the performance of the sample.* ─What was the correlation between holistic and analytical ratings?

The harsher the evaluator, the less impartial/reliable their rating was. ─Could you elaborate? Do you have any hypothesis as to why there seems to be a link between harsh evaluators and impartiality?

2021 BILC Conference

Computer Adaptive Language Testing:
From Idea to Implementation

Viktoriia Krykun, Ukraine

The Computer Adaptive Language Test (CALT) project began at the National Defence University of Ukraine in 2019 as the international scientific project. Further development was executed under the auspices of the NATO DEEP Ukraine. This report presents the CALT method and the CALT algorithm that consists of three blocks: "Starting point", "Item selection algorithm", and "Scoring and Termination criterion." The CALT algorithm has an adaptive ability, as it changes the complexity level, sequence, and the number of items based on the test taker's answers. The comparative analysis of the results of the CALT pilot and the Paper tests (in reading and listening per NATO STANAG 6001) justifies the effectiveness of the three-level CALT method. It permits us to determine the important benefits of CALT.

2021 BILC Conference

1ˢᵗ Round

Viktoriia Krykun, Ukraine

Results: With only 71% and 66% equivalence, the two formats are not equated. How will that be addressed?

The discrepancy between CALT and Paper test seems quite large. Could other factors be at play in addition to the unfamiliarity with CDT format?

Clarification: How many items per level were piloted? How many were accepted in the item bank after piloting? How often do you "feed" the item bank?

What is the algorithm requirement before delivering an item at a different level? Would it be useful to have other countries use the data base to help "teach" the algorithm? How is the algorithm calibrated?

What is the average time for an examinee to complete each test?

Is the software bought or developed? Is it managed/stored in-house or is there a license fee? Is there a demo available?

2021 BILC Conference

"I can't come to the words":
Assessment Guiding Flexible Training Solutions

Birgitte Grande, Hege Skilleås, Norway; Clayton Leishman, USA

The presenters provided examples and lessons learned from trying out the OPI+ format on Norwegian Joint Terminal Attack Controllers (JTACs), discussing how this particular form of assessment may guide flexible training and course design.

2021 BILC Conference

1st Round

Birgitte Grande, Hege Skilleås, Norway; Clayton Leishman, USA

Results: Since the OPI+ didn't affect the JTAC test results, how much does your organization plan to change the course syllabus for JTACs?

Given that there wasn't a significant effect on test results, can you elaborate on the benefits that could help justify expending manpower, training, and resources to adopt/develop OPI+ tests?

Elaboration: The OPI+, as envisioned by the US SOF community, focused on level 2 (and below) because they felt that a General Proficiency (GP) test did not reflect a candidate's linguistic capability requirements in their very specialized practitioner subjects. Conflating that with your study, could it be that, perhaps, the NATO JTAC standards, as expressed in GP terms, may simply be too high?

2021 BILC Conference

Localising a Foreign Training Resource:
The Finnish Defence Forces' English for UN Military Peacekeepers Manual

Laura Murto and Major Jarmo Vilo, Finland

The British Council and the Indonesian Armed Forces published their "English for UN Military Peacekeepers" training resources in 2014. This presentation discusses *why* and *how* the Finnish Defence Forces decided to localise it - rather than "simply" translate it.

2021 BILC Conference

1st Round

Laura Murto and Major Jarmo Vilo, Finland

- What is the shelf life of the hard copy of the manual? Are regular updates scheduled?

- We are amazed at how fast you went from project to practical outcome. Did this include an official proofreading and approval from stakeholders?

- Was the target audience involved in the project? If so, how and when?

2021 BILC Conference

2nd Round

Viktoriia Krykun, Ukraine

Elaboration: Some authors suggest that a limitation of CALT is that often the number of higher level items ends up being smaller than lower level items after the piloting, and that the items need to be piloted on a large number of test takers. What was your experience?

Statistics: Which statistical analysis did you incorporate in your software, classical or IRT?

Method: How many people participated in the development of this project?

Is online testing done remotely or on designated computers? If testing remotely, how were you able to ensure test content would not be compromised and to safeguard against cheating?

Was the algorithm developed specifically for this project or borrowed/adapted from another?

2021 BILC Conference

2nd Round

Birgitte Grande, Hege Skilleås, Norway; Clayton Leishman, USA

- Would it be better to design a specific English test for the JTACs similar to the ICAO English tests for pilots?

- Would you support the claim that STANAG 6001 descriptors should not be the ones used for developing tests for the JTACs? If yes, what could be used instead?

- Since the OPI+ integrates professional content domains in the Standard OPI; would adjusting the scaling of the scoring scheme in two parts give more accurate results?

2021 BILC Conference

2nd Round

Mary Jo Dibiase, USA; Jana Vasilj-Begović, Canada

- Could you explain which Rasch Model was used to analyze the data and why was it a good choice for the data you had?

- Could you elaborate on the limitations of the study, particularly on the area you plan to pursue for your next study?

2021 BILC Conference

2nd Round

Laura Murto and Major Jarmo Vilo, Finland

- Have you had the chance to collect feedback from the users of the manual or is it too early? If you have, what was it like?

- Is there a possibility to use the manual for blended / hybrid courses?