Computer Adaptive Language Tests (CALT)

Aurore Bargat, University of Illinois, United States

The European Conference on Language Learning 2022
Official Conference Proceedings

Abstract
Creating a test to adequately assess reading, speaking, listening, and writing proficiency in a foreign language has many challenges. Traditionally, such tests have been paper-based or done by an evaluator in a face-to-face mode. The increasing use of technology in language education has recently shifted the way assessment can be performed. This paper will develop the concept of Computer Adaptive Language Tests (CALT). The definition and characteristics of an effective CALT will be presented alongside an evaluation of currently available CALT platforms. Finally, advantages and challenges of such a language assessment option will be discussed.

Keywords: Computer Adaptive Language Tests, CALT, Adaptive Testing, Language Assessment, Educational Technology

iafor
The International Academic Forum
www.iafor.org

Introduction

Paper-based tests have been the norm for many years. For language testing specifically, creating a test to adequately assess reading, speaking, listening, and writing proficiency in a foreign language has many challenges. Switching those tests to computer-based tests makes it even harder due to the nature of language teaching and learning. When dealing with a large number of students, creating a level-appropriate and relevant test takes quite some time and correcting those tests is extremely time-consuming. The idea presented in this article is to use technology to benefit both teachers and students during language assessment. Therefore, this article will focus on the integration of technology in summative language assessments and take a closer look at the concept of computer adaptive language testing.

Definition of Assessment

Assessment provides "observable evidence of learning, determines student progress and demonstrates understanding of the curriculum" (Oldfield et al., 2012, p. 3). More precisely, relevant assessments of language proficiency provide clear criteria to the students, support personalized learning, focus on student development, and ensure that feedback leads to students' improvement (JISC, 2010). In fact, to support and engage learners, language teachers should engage the interest of learners through a variety of assessment strategies, offer alternative assignment formats, foster peer review, use technology relevantly (such as using e-portfolios to promote students' reflection and self-assessment), and design assignments in ways that encourage original thoughts and minimize opportunities for cheating (JISC, 2010). To measure students' language proficiency at the end of a module, course, or semester, summative assessments need to elicit, demonstrate and analyze what knowledge and skills learners have accumulated. Standardized multiple-choice questions are omnipresent in such assessments, resulting in "an over-reliance on simple, highly structured problems that tap fact retrieval" (Pellegrino and Quallmalz, 2010, p. 122).

In recent years, interest has risen to use computer adaptive language tests (CALT) in the context of language proficiency assessment in order to perform summative evaluation of learning outcomes (Guzmán and Conejo, 2005). Given the advantages of individual and time-independent language testing, computer adaptive language tests will prove to be a positive development in the field of language assessment (Brown, 1997).

Origins and Characteristics of CALT

In the 1960s and 1970s, the U.S. Department of Defense perceived the potential benefits of adaptive testing and supported extensive theoretical research in computer adaptive language testing (Wainer, 1990). Early attempts to build adaptive tests by the U.S. Army, Navy, and Air Force were often unsuccessful and expensive (Dunkel, 1999). In the late 1980s, computer adaptive language tests grew exponentially thanks to the work of foreign language researchers at the Defense Language Institute and at universities throughout the United States, Britain, the Netherlands, and other countries (Dunkel, 1999).

Computer adaptive language testing is the use of a computer to select and present test items to test-takers according to the estimated level of their language ability (Dunkel, 1999).

The main idea is to have the computer perform like a real-life examiner who would adapt the level of the questions as the test progresses (Dunkel, 1999). As Wainer (1990) mentioned, we learn the most about an examinee's ability when we accurately direct our questions at the current level of the examinee's ability.

A computer adaptive language assessment should reliably adjust to the examinee and generate new questions based on the student's answers. To do so, the computer uses a complex algorithm to score each answer. Based on the score of that item, the computer algorithm adjusts the level of the next item to the next level. Correct answers increase the difficulty and incorrect answers lower the difficulty of the next question. Today, in the field of language assessment, most CALT evaluate students' reading and listening skills. However, it is still difficult to use computer adaptive tests to accurately measure speaking and writing skills because of their intrinsic variability nature. New practices are extending the use and purpose of technology-enhanced assessment, which can now include results management and processing, learning analytics, and tools that enable instant formative feedback and collaboration on feedback processes (Beevers et al., 2011).

Within CALT, there is a real need for identifying and specifying the exact purpose of each assessment. This is important because the purpose of a computer adaptive language test can be any of the following (Green et al., 1995):

- Identify that a learner has met the specific objectives of a language course.
- Indicate a learner's level of achievement in a specific skill (e.g., listening comprehension, reading comprehension, or grammar knowledge).
- Identify specific areas in which a student needs specific assistance (e.g., knowledge and use of specific grammatical points or recognition of specific idioms and vocabulary items).
- Diagnose a learner's strengths and weaknesses.
- Detect if learners have met minimum course requirements.

Case Studies

One of the first examples of CALT was created by the Brigham Young University in 1986 to test reading and listening abilities. Based on its trials and success in establishing the reading and listening levels of students at that university, the first studies of computer adaptive language tests were finally possible. Madsen's study (1991) indicates that in CALT, fewer items are necessary than in paper-based tests to determine the students' level, and the testing time is shorter. In this specific example, an average of 23 items were needed to adequately test the students in about 27 minutes. The comparable conventional reading test used in the study required 60 items and 40 minutes (Brown, 1997). This proves that CALT can save time during language assessment.

In another study, Bergstrom, Lunz, and Gershon (1992) analyzed the responses of 225 students on a medical technology examination and found that altering the difficulty of items slightly raises the number of items that are necessary to test students with adequate precision.

With very little studies available in the realm of CALT and the lack of diversity in terms of languages analyzed, it would be highly relevant to conduct more studies related to CALT. It seems that those limits can only allow us to analyze a few of the current (and still scarce) examples of CALT available on the market today. The six CALT tools analyzed below have been randomly distributed and do not reflect any preference or endorsement from the author.

1. Linguaskill is an adaptive language test developed by Cambridge. It has been trialed for 40 languages from 50 countries to ensure accuracy and reliability. The reading and listening parts are adaptive because the level of the questions varies based on the test-taker answers. Therefore, every test-taker receives a unique version of the test tailored to their ability level (Cambridge English, 2018). For the written part of the test, an automarker, using a complex algorithm, automatically corrects the text entered by the test-taker. The automarker technology works by extracting and analyzing features from the test-takers' written samples such as sentence structure, errors, word combinations, vocabulary used, grammar accuracy, punctuation rules, and text cohesion (Cambridge English, 2018).

2. The Avant STAMP Pro is a computer adaptive language proficiency test evaluating speaking, writing, listening, and reading skills. It provides assessment in 16 languages (including Arabic, English, French, Japanese, Korean, Mandarin, Russian, Spanish). A sample test for the French language shows the authentic reading excerpt on the left side and the multiple-choice question on the right side. Most CALT evaluating students' reading comprehension use multiple-choice questions as they are the easiest type of question for the computer to analyze and grade. One of the drawbacks that can be noticed in this test is the overall use of English for the instructions and for parts of the test itself. Providing an authentic environment would allow for a better measurement of the actual language competencies of the test-takers. Clear and simple instructions in the target language are recommended as well as the use of authentic reading and listening material.

3. The ACTFL Listening and Reading Computer Adaptive Test® (ACTFL L&Rcat) is a computer adaptive language assessment designed to test the English listening and reading proficiency of the test-taker. This test customizes the texts and passages based on the test□taker's own reading and listening ability to create a unique testing experience (Language Testing International, 2018). The ACTFL L&Rcat questions are based on a range of informal and formal material on general, social, professional, and academic topics, such as daily interactions, announcements, emails, instructions, newspaper articles, technical reports, literary texts, discussions, lectures, broadcasts, etc. These are real□world examples of spoken and written language that surpass the scope of traditional fixed tests (Language Testing International, 2018). One could wish that languages other than English could be evaluated using this tool.

4. The Duolingo English Test is an English language proficiency test using artificial intelligence and machine learning to automatically create thousands of criterion-referenced items, assess each item, and grade test-takers' answers. When Duolingo started designing this test, the company's objective was to create a language proficiency test with greater efficiency, better security, lower cost, and universal access (Brenzel and Settles, 2017). Regarding test security, it is said that test-takers would have to take the test about 1,000 times before seeing the same test item again. Also, this test does not rely on multiple-choice questions but focuses on interactive items such as

listening transcription and speaking exercises. For the speaking part of the test, the computer uses not only speech recognition, but also intonation, rhythm, and stress analysis (Brenzel and Settles, 2017). For this tool too, one could wish that languages other than English could be evaluated.

5. WebCAPE is a multiple-choice assessment that uses adaptive technology to accurately measure reading, grammar, and vocabulary skills. The test is available in English, Spanish, French, German, Italian, Mandarin, and Russian. The company developing it, Emmersion, considers that problems related to cheating and test security are virtually nonexistent in this test because each test is unique to one test-taker. Once the algorithm has reached a point where it can no longer increase or decrease the difficulty of the questions, the test terminates, and the test-taker receives a final score (Emmersion, 2020). Unfortunately, writing, listening, and speaking skills are not evaluated by this tool.

6. TrueNorth is an automated speaking assessment using artificial intelligence to measure speaking ability. This test has three sections that collect three distinct types of data. It begins with a language background survey where the software collects information such as the test-taker's overall language experience and previous exposure to the language being assessed (Emmersion, 2020). The first part of the actual assessment measures how efficiently the test-taker's brain can process language information. It is used to determine the difficulty level for the questions asked in the second part of the test. The second part of the assessment follows a question-answer format, collecting spontaneous responses from the test-takers. For each question, a test-taker has 30 seconds to read the prompt and prepare to respond, and then 60 seconds to respond (Emmersion, 2020). After both parts are finished, the system analyzes all of the data to determine a final score.

After looking at those six examples of CALT, one can still wonder if there is a relationship between item difficulty and test length and what difficulty level would be ideal to precisely measure students' language proficiency level (Brown, 1997).

Advantages of CALT

In a computer adaptive language test, each test-taker takes a unique test that is tailored to their ability level. Questions with low information value about the test-taker's proficiency are avoided, which allows a higher precision across a wider range of ability levels (Carlson, 1994). Some of the advantages of CALT include the following points (Angus and Watson, 2009; Carlson, 1994; JISC, 2010; McNamara, 1991; Pellegrino and Quellmalz, 2010; Wainer, 1990):

- promotes personalization
- improves the appropriateness, effectiveness, and consistency of the assessment
- provides efficient assessment processes that have pedagogic benefits
- allows test-takers to work at their own pace
- test-takers are challenged by test items at an appropriate level and are not discouraged or annoyed by items that are far above or below their ability level
- can be scored immediately, providing instantaneous feedback to the test-taker
- can include text, graphics, photographs, and even full-motion video clips

- can help identify individuals who meet specific performance standards
- provides opportunities to design richer and more interactive assessment and feedback
- supports the diverse needs of learners
- improves learner engagement
- can capture wider skills and attributes not easily assessed otherwise
- provides accurate results with opportunities to combine human and computer marking
- provides accurate, timely, and accessible evidence on the effectiveness of curriculum design and delivery
- increases efficiency and reduce teachers' workloads
- improves assessment validity and reliability
- accommodations can easily be made for visually or auditory impaired test-takers

Challenges and Critiques

First, summative assessments have received significant criticism over the last few years when it comes to accurately measuring the proficiency level of language learners. It is often considered a simple recall of knowledge previously learned that measures students' ability to attain a specific level (Gee and Shaffer, 2010). Additionally, it is seen to offer little in terms of evaluating the actual knowledge or transferable skills that the students would use in the world outside school (Gee and Shaffer, 2010).

The use of technology could potentially solve this issue, but the implementation of new technology tools within the realm of language assessments still brings new challenges. The first would be that computer adaptive language tests do not provide an accurate determination of the grammatical and lexical abilities of students (JISC, 2010). For instance, multiple-choice questions are still often used and simply require candidates to select an answer from predetermined options rather than construct their own answers (JISC, 2010). A badly designed multiple-choice test, resulting in the students guessing the right answer, would negatively impact the viability of CALT.

According to Whitelock and Watt (2008), the following items can be a source of concerns when it comes to implementing computer adaptive language tests:

- technology-enhanced assessment practices are not spread evenly across subjects (still scarce for languages)
- concerns about plagiarism detection
- inability to handle open-ended questions
- difficulties in scalability and transferability of practices
- concerns over reliability and validity of assessment (how to ensure all students receive equivalent tests if questions are selected at random from a question bank)
- user identity verification and security issues
- lack of teacher training on how to use this technology and how to get the most out of the analytics provided by the computer software
- cost of investment in training, support, infrastructure (having enough on-site computers if on-site testing is required)
- use an appropriate scoring method (raw scores, weighted raw scores, or scaled scores)

Language assessment has its own set of specificities that add a few more challenges when it comes to creating a relevant computer adaptive test. For example, in the case of a listening proficiency assessment, items in the pool need to include a variety of listening tasks, such as comprehension of the main ideas, recognition and recall of key details, as well as identification of specific words and phrases (Green et al., 1995).

According to Whitelock (2010), assessment should embrace the "potential to move away from the assessment of individual skills to implement a social constructivist view of learning" (p. 2). Following this idea, some skeptics of the transformative potential of technology tools for learning cite how social and educational identities and inequalities will not necessarily shift through the use of new tools. Hughes (2009) argues that "enthusiasm for the new technologies should be tempered with an appreciation that identity and belonging to learning communities are complex and often unresolved issues for learners" (pp. 291-292). Indeed, studies have shown that the shift in format from paper to digital can create physical and psychological issues for the test-taker (like eyestrain and anxiety) and that familiarity (or not) with technology can greatly impact test results (Fritts and Marszalek, 2010; Lu et al. 2016; Madsen, 1991).

The last challenge, that is sometimes overlooked by teachers and test-takers alike, is the technical difficulties of creating such computer-based assessment tools. The use of technology in computer adaptive language tests can encompass a wide variety of devices from recording equipment, statistical programs, and databases, to programs capable of language recognition (Chapelle, 2008). All those parts need to be relevant and properly set for the assessment to be reliable. Things like the user interface generation, the test engine, and the random yet relevant selection of questions also need to be properly set in order for the test to be viable. One of the requirements of CALT is to provide a flexible and configurable architecture that can be instantiated in a structured way for different use cases (Oppl et al., 2017). As long as this cannot be resolved efficiently, a reliable and viable computer adaptive language test will never be able to accurately determine a test-taker proficiency level in all four modalities (writing, listening, speaking, and reading).

Conclusions

After defining the concept of CALT and providing six relevant and current examples, the advantages and challenges of CALT as well as critiques and literature gaps were discussed.

As García Laborda (2007) states, the benefits of computer adaptive language tests should overcome its drawbacks, as it can be faster, more efficient, and less costly than paper-based tests. CALT can also facilitate the difficult task of level evaluation, rapid correction, feedback, and reporting (García Laborda, 2007. Further research should be done on the accuracy and reliability of CALT. It is also vital to better understand how digital technologies can support fairer and more equitable assessment methods for language learners. Finally, it would be relevant to get more insights in the debate around ethics and the challenges to protect users' data when test-takers are taking a computer adaptive language test, either on-site or at home.

References

Angus, S.D., & Watson, J. (2009). Does regular online testing enhance student learning in the numerical sciences? Robust evidence from a large data set. *British Journal of Educational Technology*, *40*(2), 255-272.

Beevers, C. et al. (2011). What can e-assessment do for learning and teaching? Part 1 of a draft of current and emerging practice review by the e-Assessment Association expert panel. *International Journal of e-Assessment*, *1*(2).

Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the difficulty level in computer adaptive tests. *Applied Measurement in Education*, *5*, 137-149.

Brenzel, J., & Settles, B. (2017, September 28). The Duolingo English Test — Design, Validity, and Value. Duolingo English Test Whitepaper. *Duolingo*. https://s3.amazonaws.com/duolingo-papers/other/DET_ShortPaper.pdf

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology*, *1*(1), 44-59.

Carlson, R. (1994). Computer-adaptive testing: A shift in the evaluation paradigm. *Journal of Educational Technology Systems*, *22*(3), 213- 224.

Chapelle C.A. (2008) Utilizing Technology in Language Assessment. In: Hornberger N.H. (eds) *Encyclopedia of Language and Education*. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30424-3_172

Dunkel, P. A. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning & Technology*, *2*(2), 77–93. http://dx.doi.org/10125/25044

Fritts, B. E., & Marszalek, J. M. (2010). Computerized adaptive testing, anxiety levels, and gender differences. *Social Psychology of Education*, *13*(3), 441–458. https://doi.org/10.1007/s11218-010-9113-3

García Laborda, J. (2007). On the net: Introducing standardized EFL/ESL exams. *Language Learning & Technology*, *11*(2), 3–9.

Gee, J.P., & Shaffer, D.W. (2010). Looking where the light is bad: Video games and the future of assessment. *Edge: The latest information for the education practitioner*, *6*(1), 3-19.

Green, B., Kingsbury, G., Lloyd, B., Mills, C., Plake, B., Skaggs, G., Stevenson, J., Zara, T., & Schwartz, J. (1995). *Guidelines for computerized-adaptive test development and use in education.* Washington, DC: American Council on Education, Credit by Examination Program.

Guzmán, E., & Conejo, R. (2005). Self-assessment in a feasible, adaptive web-based testing system. *IEEE Transactions on Education*, *48*(4), 688–695.

Hughes, G. (2009). Social software: new opportunities for challenging social inequalities in learning? *Learning, Media and Technology*, *34*(4), 291-305.

JISC. (2010). *Effective Assessment in a Digital Age. A guide to technology-enhanced assessment and feedback.* https://facultyinnovate.utexas.edu/sites/default/files/digiassass_eada.pdf

Lu, H., Hu, Y., Gao, J., & Kinshuk. (2016). The effects of computer self-efficacy, training satisfaction and test anxiety on attitude and performance in computerized adaptive testing. *Computers & Education*, *100*, 45–55. https://doi.org/https://doi.org/10.1016/j.compedu.2016.04.012

Madsen, H. S. (1991). Computer-adaptive testing of listening and reading comprehension. *Computer-assisted language learning and testing: Research issues and practice*. 237-257. New York, NY: Newbury House

McNamara, T. (1991). Test dimensionality: IRT analysis of an ESP listening test. *Language Testing*, *8*(2), 139-159.

Oldfield, A., Broadfoot, P., Sutherland, R., & Timmis, S. (2012). *Assessment in a Digital Age: A research review*. University of Bristol. https://www.bristol.ac.uk/media-library/sites/education/documents/researchreview.pdf

Oppl, S., Reisinger, F., Eckmaier, A., & Helm, C. (2017). A flexible online platform for computerized adaptive testing. *International Journal of Educational Technology in Higher Education*, *14*(2). https://doi.org/10.1186/s41239-017-0039-0

Pellegrino, J. W., & Quellmalz, E.S. (2010). Perspectives on the Integration of Technology and Assessment. *Journal of Research on Technology in Education*, *43*(2), 119-134.

Wainer, H. (1990). *Computer adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum.

Whitelock, D. (2010). Activating Assessment for Learning: are we on the way with Web 2.0? In Lee, M.J.W. and McLoughlin, C. (Eds.) *Web 2.0-Based-E-Learning: Applying Social Informatics for Tertiary Teaching.* IGI Global. 319–342.

Whitelock, D., & Watt, S. (2008). Reframing e-assessment: adopting new media and adapting old frameworks. *Learning, Media and Technology, 33*(3), 151-154.

Contact email: abargat2@illinois.edu