Improving Test Validity and Accessibility with Digital-First Assessments

Naomi Care & Bryan Maddox
Assessment MicroAnalytics Ltd
Naomi@microanalytics.co.uk • Bryan@microanalytics.co.uk

Abstract

Digital-first high stakes assessments invite us to rethink the standards, user experience, and vocabulary of assessment validity. We explore the disruptive potential of digital-first high stakes assessments to set higher standards in test performance and validity. In this paper we describe the new lexicon that digital-first assessments have introduced into high stakes digital assessments such as *personalisation, user experience,* and *accessibility*. The features of digital high stakes assessments have the potential to reduce sources of construct-irrelevant variation and improve test validity. In conclusion we argue that the distinctive features of digital high stakes assessment challenge our understanding of good assessment design setting new standards for assessment performance and validity.

Keywords

High-stakes assessment, digital-first assessment, user experience, validity, inclusive assessment.

About the authors

Naomi Care has an academic background in education, anthropology, and psychology. As an Analyst with Assessment MicroAnalytics, she leads work in the area of digital ethnography, and on neurodiversity, disability and chronic illness in assessment.

Bryan Maddox is Executive Director of Assessment MicroAnalytics, Professor in Educational Assessment at the University of East Anglia, and visiting professor at the Centre for Educational Measurement at the University of Oslo. Bryan has conducted assessment research in France, Luxembourg, Mongolia, Nepal, Senegal, Slovenia, the United States, and the UK.

Contents

Introduction

The world of educational assessment is experiencing an unprecedented upheaval in both design and delivery. The recent advances in computer-based assessment have provided an unparalleled opportunity to rethink the standards, user experience, and vocabulary of assessment validity. Assessments are not simply ways to evaluate knowledge, they are the means to allow test-takers to flourish in their everyday lives. In terms of language assessments, these provide a means to study and reside in the country of the test-takers' choice. With advances in digital assessment, there are new capabilities to interact with test-takers, create virtual scenarios to test their knowledge using engaging items, and remove the administrative barriers afforded by paper-based exams, which have been accentuated during the Covid-19 pandemic. As Mislevy and Haertel (2006) asked, "How can we use these new capabilities to tackle assessment problems we face today?" (p. 6).

In this paper we describe the new lexicon that digital assessments have introduced into high stakes assessment, such as personalisation, user experience, and accessibility, to explore the opportunities provided by computer-based assessment. Using the Duolingo English Language Test (DET) as an example of innovative assessment design, we argue that the features of high-stakes digital assessments have the potential to reduce sources of construct-irrelevant variation and improve test validity.

What are the challenges of personalisation for validity?

The DET assesses English language proficiency and aims to lower barriers to access (Settles et al, 2020). It is delivered via the internet from the users' home and can be taken at any time on any day. The DET is personalised to the individual user; by using computer-adaptive test (CAT) design, the test chooses items that are the most appropriate level for the test-taker, ensuring that the questions are neither too difficult nor too easy. Such personalisation has become possible due to the advent of digital assessment. With the digital transition comes a new challenge to consider the implications for test validity, robustness, and fairness in the wake of significantly different test designs. This segment will explore the various challenges of personalisation for validity, with a particular emphasis on the departure from standardisation. In short: can a personalised assessment be a valid one?

Large-scale, high stakes paper-based assessments (PBAs) require standardisation to ensure they are robust and valid. Standardisation refers to following predetermined criteria and standards, including ensuring the test items in the assessment accurately represent the required difficulty level (Gerritsen-van Leeuwenkamp, Joosten-ten Brinke and Kester, 2017). For PBAs such predetermined criteria are essential to guarantee that the test measures what it purports to measure. For example, if pupils could take their PBA at a time of their choosing then answers could easily be leaked, undermining the validity of that test. Instead, all students are required to take the same test, at the same time, in roughly the same environment, with the promise that their tests will be subject to the same scoring processes.

It has been argued that standardisation of assessment comes with clear advantages; Kane (2013) suggests that it promotes fairness by ensuring all test takers complete the same items under the same conditions. Furthermore, standardisation allows for the effective control of random error, though does have the disadvantage of introducing possible systematic errors. Kane (2013) offers the example of providing a standardised time limit to all candidates. This controls any random errors and unfairness caused by candidates being rushed or receiving extended time during the assessment, whilst also limiting systemic errors by ensuring the time limit is sufficient for all. Indeed, since the introduction of widespread

educational testing in the 20th Century, standardisation has become the hallmark of fair and accurate assessment (Sireci, 2020). Whilst PBAs remain the standard method of assessment, it is hard to envisage a situation where standardisation will not be necessary.

Nonetheless, with the gradual introduction of computer-based assessments (CBAs) as a meaningful alternative to PBAs, it is time to question the relationship between standardisation, validity, and fairness. Standardisation introduces two implicit assumptions. First, that the test is being taken by a homogenous population that approaches the test with similar needs and abilities. Second, that it is possible to create a largely equivalent testing situation where procedures, protocols and the testing environment are similar enough to assure the validity of the test is unaffected. Here these two assumptions will be unpacked to question the need for standardisation in the design of CBAs. It will be argued that, rather than being a hallmark of robust assessment, standardisation introduces construct irrelevant variance and delivers an exam that, instead of being fair, risks excluding students from the meaningful assessment of their abilities.

Fairness and standardisation

Students form a vastly heterogeneous population. They bring with them differences in socio-economic resources, cultural beliefs and understanding, neurodivergence, health, experiences and expectations of the institutions supplying the tests, and different geographies. This heterogeneity results in standardisation working for some, but never for all. For example, a PB standardised test that requires students to sit in a large, noisy exam hall may well work for many of the students that sit it. However, students who experience high levels of test anxiety or have a form of neurodiversity such as Autism Spectrum Disorder may find the testing situation far less than ideal (Daly, Chamberlain and Spalding, 2011; Chown et al., 2018). DeCaro et al. (2011), found that the performance environment influences attention and skill execution, with some students "choking under pressure" as their concentration is directed away from the task at hand, and towards the outcome of the task. As such, the standardised delivery of the test, even with conventional accommodations, can lead to some students becoming unable to adequately demonstrate their knowledge and skills.

The legacies of standardisation stretch far beyond the testing situation. A standardized assessment means that each student is required to answer the same questions, some questions will be below the student's aptitude whilst others may exceed their skill-set. Such an assessment results in a loss of granularity; students become unable to demonstrate their actual skill-sets as they wade through questions that are not attuned to their ability. In order to retain granularity, testing organisations create longer tests to include a wider range of questions. Again, this leads to an unsatisfactory burden on students. Many students, particularly those with disabilities and neurodiversity, struggle disproportionately with fatigue, strained working memory and stressed executive functioning that comes from sitting an exam of 3 hours or more (Daley and Birchwood, 2010; Douglas et al., 2016; Cameron et al., 2019). As a result, the test fails to test a student on their skills and knowledge, but instead introduces the contaminants of stamina and sustained attention into the testing situation. Such contaminants are forms of construct irrelevant variance (Messick, 1990); the result being the exclusion of a subset of students who are unable to meet the additional demands of extended performance. In most cases, a shorter test is more accessible than conventional accommodations such as extra time. As such, there are potential advantages to being able to personalise a test to reduce both the length and energy spent on irrelevant items. This then raises the question: is it possible for shorter tests to be as valid as longer standardised assessments?

The opportunities afforded by CAT

Given the transition from paper-based assessment to digital assessment and the increasing popularity of CAT as a testing framework, questions pertaining to reliability and validity do not fall solely on the DET, but also on a vast number of organisations who are beginning to utilise CAT in a number of high stakes assessments (Georgiadou, Triantafillou and Economides, 2006). As such, it is worth exploring whether a CAT can be as reliable and robust as a fixed assessment.

Whereas traditional standardised assessments aim to accommodate the "average" responder, CAT chooses questions that meet the ability of the test taker based on their previous responses (Martin and Lazendic, 2018). The questions given should not be too easy, nor too difficult, but instead scaffold and stretch the test taker at their current level. A result of the CAT format is that tests are far shorter. In the case of the DET, the test can be completed in less than an hour, whereas traditional English language Tests such as the PTE Academic, TOEFL iBT and IELTS take up to 3 hours to complete. Proponents of CAT have argued that the targeting afforded by CAT leads to less error, through greater granularity and discrimination (Stone and Davey, 2011). Kingsbury and Hauser, (2004) compared fixed and adaptive forms of 4th and 8th grade reading and mathematics tests and found that the granularity of information at the extreme ends of ability distribution was three times greater in the CAT setting. Such findings have been echoed by a number of researchers who have found that CAT enables greater measurement precision and improved granularity for high and low achievers (Thurlow et al., 2010; Stone and Davey, 2011; Martin and Lazendic, 2018). The research literature indicates CAT provides the opportunity to create a shorter assessment with fewer test items whilst enhancing granularity and measurement precision. As such, though the use of CAT does not in of itself create a valid test, the extant literature presents a robust argument for the validity of CAT as an assessment tool.

Does delightful user design support a more valid test?

The Duolingo English Test's test taker experience (TTX) is a holistic perspective which includes the development of "delightful UX design" (Burstein et al., 2021 p.5). User Experience (UX) studies, though established in the fields of web and game design, have received relatively little attention within the realm of educational assessment. However, this is slowly beginning to change, as academics and test designers begin to explore the impact of good or poor UX on the academic outcomes of test-users (Krisnawati et al., 2019). UX can be understood as a "person's responses and perceptions that resulted from the use and/or anticipated use of a system, product or service" (Nagalingam and Ibrahim, 2015 p.424). As such, user experience includes hedonic elements, such as enjoyment, engagement and motivation, as well as more tangible interface elements such as navigation, item design and aesthetics (Allam, Hussin and Dahlan, 2013). TTX goes beyond UX to include the full test-taker experience, from item design to test administration to score reporting processes, as well as test-readiness materials and support (Burstein et al., 2021). In terms of test design, good TTX is essential to minimise forms of construct irrelevant variance (CIV). Below, a case will be made for including UX in assessment validity arguments, with the premise being that an enjoyable, well-designed test is also a more valid test.

Within assessment design, there are underlying assumptions concerning how a candidate will engage with an assessment. Assumptions include candidates displaying their ability using established procedural processes and knowledge elements. Kane and Mislevy (2017) suggest that process-model interpretations can be used to validate assessments when candidates perform in ways that align with a specified model.

The soundness of such interpretations can be evaluated using a range of data points, including log data, response times or in-situ evaluations such as eye-tracking, think-aloud and gesture analysis (Maddox, 2017). Such evaluations are essential given that humans do not necessarily perform in expected and unitary ways. During a testing situation candidates may experience an array of affective responses including boredom and engagement, they may become frustrated with the assessment interface, give up, arrive at the correct answer using unexpected strategies or fail to adequately read the question. Unexpected user responses can challenge score interpretations and, as a result, undermine the validity of a test (Kane and Mislevy, 2017). As such, there is an imperative for test designers to consider the importance of TTX and the wider test taker experience to mitigate, as far as possible, against unexpected responses (Burstein et al., 2021).

Cronbach (1949) noted that one of the primary assumptions made about test takers is that they are motivated to take the exam and perform to their best ability and, furthermore, that this assumption is likely to be flouted in some cases. Indeed, test takers have been shown to experience a wide range of emotions, some of which may hinder their ability to perform on the day (Goetz et al., 2007). Pekrun (2006) devised an integrative framework to examine the antecedents and effects of emotions experienced in achievement settings, such as examinations. Here he noted that students experience a broad range of emotions including frustration, hopelessness, anxiety, pride and hope when faced with academic tasks. Furthermore, he noted that affective responses alter academic outcomes, for example, boredom has been shown to lead to poorer academic outcomes, poor study habits and disengagement from the situation as a means to escape (Pekrun et al., 2010; Tze, Daniels and Klassen, 2016). In this case, the design of an assessment can directly affect the experience of candidates. In a comparison between a game-based assessment and a conventional assessment, researchers found that all performance groups reported increased enjoyment when using the game-based assessment as opposed to the conventional, with less boredom and more enjoyment (Lehman, Jackson and Forsyth, 2019). Other researchers have found that an enjoyable test leads to increased engagement and effort, ultimately leading to test-takers being able to perform at an optimum level during their test (DiCerbo, 2017; Klerk and Kato, 2017). In this case, boredom and frustration can be seen, to borrow Messick's (1990) language, as a contaminant to the testing situation as, instead of testing knowledge, the test begins to assess how well a candidate can concentrate in the face of tedium. As such, creating an enjoyable test directly supports the validity of an assessment by lessening boredom and disengagement and, thus, allowing the test taker to display their knowledge adequately.

A more comfortable testing experience can be created by perfecting different elements of UX design such as navigation, flow, and aesthetics. Each of these can impact on the experience and responses of the test taker. For example, poor or confusing navigation can quickly lead to frustration, disengagement and disorientation (Webster and Ahuja, 2006). In a high-stakes testing situation, an unclear interface can cause undue stress as the test taker aims to move on and save precious time. UX failures, such as text size that is too small, unclear images, or a cluttered interface can introduce construct-irrelevant variation. For example, if the test taker cannot see an image they are required to describe in an English Language exam, they will not pass that test item. Even the aesthetics of a test can influence how users respond, with rapid guessing rates being linked to the attractiveness of a test (Penk, Pöhlmann and Roppelt, 2014). However, computer-based assessments provide affordances that allow test designers to mitigate such issues well. New item types in the DET are piloted prior to use in order to test their viability by minimising UX-related CIV and unexpected test-taker responses.

What does computer-based testing mean for accessibility?

Historically, students with disabilities, learning difficulties and neurodiversity have faced barriers both in accessing suitable education, and being able to display their knowledge in assessments. Statistically, students with disabilities face a significant achievement gap compared to their peers throughout their education (Gilmour, Fuchs and Wehby, 2019). In terms of computer-assisted assessments, poor implementation of access arrangements or accommodations have led to a "digital divide" between students with and without disabilities (Konur, 2007 p.207). In the United States, "for disabled people who use assistive technology or need adaptations (such as increased text-size), there has been no guarantee that digital services will work as both the web and digital tools have proliferated" (Lewthwaite and James, 2020 p. 1360). Indeed, though there are laws in the UK and Europe to ensure that online assessments are fully accessible to students with disabilities, these laws have not been enforced, resulting in inequity in many online assessments (Lewthwaite and James, 2020). The transition from paper-based testing to computer-based testing does not automatically equate to a more accessible assessment.

However, digital first assessment does offer an opportunity to create equitable assessments when accessibility is built into the design of the assessment (Nganji and Brayshaw, 2017). This can be done in two ways; by designing a test that is more accessible to all, and by ensuring test takers are able to personalise an exam in a way that accommodates their disability. In terms of design, there are several features of the DET that lend well to supporting students with disabilities. As discussed above, a CAT examination has a number of benefits for all test takers, including shorter testing time and targeted questioning. The majority of disabilities and learning difficulties come with an element of fatigue (Gilmour, Fuchs and Wehby, 2019). Attention-Deficit Disorder (ADD) and associated conditions not only lead to difficulties in concentrating, but also associated fatigue when prolonged concentration is demanded (Gillberg, 2014; Lovett and Nelson, 2021), whereas test takers with dyslexia are likely to experience fatigue after long reading tasks (Mortimore and Crozier, 2006). Providing extra-time in assessments as an adjustment does not always support the test taker, but instead prolongs the amount of time they need to concentrate and can exacerbate fatigue (Lovett, 2020). However, shorter tests can support students and allow them to adequately display their skills and knowledge (Stone and Davey, 2011). Similarly, creating a test that can be taken at home means that the barriers often presented to test-takers with physical disabilities, chronic pain or mental health difficulties are lessened, as they can take the test in a place that is comfortable for them, with the equipment that they need and at a time where their pain and energy levels are better controlled. In these ways, the DET is accessible by design; the test is shorter than legacy English proficiency exams, and can be taken at home (Burstein et al., 2021).

Nonetheless, further measures can be taken to support test-takers with disabilities. Stone and Davey (2011) note that, even with a CAT design, barriers to test takers can remain. For example, innovative item types need to be compatible with adaptive technology (Nganji and Brayshaw, 2017). There is also a need to allow multiple options for selecting responses (e.g. mouse, keyboard, touchscreen) (Stone and Davey, 2011). However, CBAs lend themselves well to allowing for such accommodations. Furthermore, CBAs provide the opportunity to personalise a test to accommodate all users, for example, by allowing them to choose the most appropriate text size and overlay options that minimises eye-strain. The DET is already commended for their accessible approach. However, by utilising their expertise in innovative, digital first design, the DET has the opportunity to create an assessment that allows all students to flourish.

Conclusion

This paper highlights key considerations for education leaders in evaluating the scope for digital first design to reduce access inequalities, the validity of CAT design, and the scope for a more delightful test-taking experience. The transition to digital first assessments has created new opportunities to enhance not just the testing experience, but the robustness and validity of assessments. The affordances of digital first assessments mean that we need not continue to accommodate the legacies of paper-based testing. In particular, the opportunity to provide personalised assessment using CAT design creates opportunities to create shorter, enjoyable, accessible, and robust assessments. Whilst standardisation can be exclusionary by design, CAT ensures that each test taker is able to demonstrate their knowledge and abilities in a shorter time, at a time of day that works well for them, and potentially, in the comfort of their own home. The disruptive potential of the transition to digital first assessment brings new opportunities to re-examine and reconsider the old architecture of assessment, not just by introducing a new lexicon and updating the ways in which we evidence validity, but also by setting higher standards and aspirations. A valid assessment does not need to be a tedious assessment; indeed, improving TTX can enhance the validity of the test whilst making it a more comfortable experience for the test taker. Most importantly, by embracing personalisation and TTX it is possible to make assessments accessible to those who have historically struggled to demonstrate their skills due to the barriers placed by standardisation. It should no longer be acceptable to include exclusionary ‘contaminants’ in assessment design. This means removing some of the practical and pragmatic constraints that are legacies of the paper mode of design and delivery and rethinking what assessment validity means, and how it is evidenced in the digital era.

References

Allam, A., Hussin, A. and Dahlan, H. (2013) 'User Experience: Challenges and Opportunities', Journal Of Information Systems Research And Innovation, 3(1).

Burstein, J. et al. (2021) 'A Theoretical Assessment Ecosystem for a Digital-First Assessment—The Duolingo English Test', Duolingo Research Report DRR-21-04 [Preprint]. Available at: englishtest.duolingo.com/research.

Cameron, H. et al. (2019) 'Equality law obligations in higher education: reasonable adjustments under the Equality Act 2010 in assessment of students with unseen disabilities', Legal Studies, 39(2), pp. 204–229. doi:10.1017/lst.2018.31.

Cardwell, R., La Flair, G. and Settles, B. (2021) 'Duolingo English Test: Technical Manual'. Duolingo English Test. Available at: https://s3.amazonaws.com/duolingo-papers/other/Duolingo%20English%20Test%20-%20Technical%20Manual%202019.pdf (Accessed: 16 September 2021).

Chown, N. et al. (2018) 'The "High Achievers" project: an assessment of the support for students with autism attending UK universities', Journal of Further and Higher Education, 42(6), pp. 837–854. doi:10.1080/0309877X.2017.1323191.

Cronbach, L.J. (1949) Essentials of psychological testing. Oxford, England: Harper (Essentials of psychological testing), pp. xiii, 475.

Daley, D. and Birchwood, J. (2010) 'ADHD and academic performance: why does ADHD impact on academic performance and what can be done to support ADHD children in the classroom?', Child: Care, Health and Development, 36(4), pp. 455–464. doi:https://doi.org/10.1111/j.1365-2214.2009.01046.x.

Daly, A.L., Chamberlain, S. and Spalding, V. (2011) 'Test anxiety, heart rate and performance in A-level French speaking mock exams: An exploratory study', Educational Research, 53(3), pp. 321–330. doi:10.1080/00131881.2011.598660.

DeCaro, M.S. et al. (2011) 'Choking under pressure: Multiple routes to skill failure', Journal of Experimental Psychology: General, 140(3), pp. 390–406. doi:10.1037/a0023466.

DiCerbo, K.E. (2017) 'Building the Evidentiary Argument in Game-Based Assessment', Journal of Applied Testing Technology, 18(S1), pp. 7–18.

Douglas, G. et al. (2016) 'Including Pupils with Special Educational Needs and Disability in National Assessment: Comparison of Three Country Case Studies through an Inclusive Assessment Framework', International Journal of Disability, Development & Education, 63(1), pp. 98–121. doi:10.1080/1034912X.2015.1111306.

Georgiadou, E., Triantafillou, E. and Economides, A.A. (2006) 'Evaluation parameters for computer-adaptive testing', British Journal of Educational Technology, 37(2), pp. 261–278. doi:10.1111/j.1467-8535.2005.00525.x.

Gerritsen-van Leeuwenkamp, K.J., Joosten-ten Brinke, D. and Kester, L. (2017) 'Assessment quality in tertiary education: An integrative literature review', Studies in Educational Evaluation, 55, pp. 94–116. doi:10.1016/j.stueduc.2017.08.001.

Gillberg, C. (2014) ADHD and its many associated problems.

Gilmour, A.F., Fuchs, D. and Wehby, J.H. (2019) 'Are Students With Disabilities Accessing the Curriculum? A Meta-Analysis of the Reading Achievement Gap Between Students With and Without Disabilities', Exceptional Children, 85(3), pp. 329–346. doi:10.1177/0014402918795830.

Goetz, T. et al. (2007) 'Emotional experiences during test taking: Does cognitive ability make a difference?', Learning and Individual Differences, 17(1), pp. 3–16. doi:10.1016/j.lindif.2006.12.002.

Kane, M. and Mislevy, R. (2017) 'Validating Score Interpretations Based on Response Processes', in Validation of Score Meaning for the Next Generation of Assessments. Routledge.

Kingsbury, G.G. and Hauser, C. (2004) Computerized Adaptive Testing and 'No Child Left Behind', Northwest Evaluation Association. Northwest Evaluation Association. Available at: https://eric.ed.gov/?id=ED491245 (Accessed: 17 September 2021).

Klerk, S. de and Kato, P.M. (2017) 'The Future Value of Serious Games for Assessment: Where Do We Go Now?', Journal of Applied Testing Technology, 18(S1), pp. 32–37.

Konur, O. (2007) 'Computer-assisted teaching and assessment of disabled students in higher education: the interface between academic standards and disability rights', Journal of Computer Assisted Learning, 23(3), pp. 207–219. doi:https://doi.org/10.1111/j.1365-2729.2006.00208.x.

Krisnawati et al. (2019) 'First Time User Experience Assessment on Web based Online Examination', in 2019 International Conference on Information and Communications Technology (ICOIACT). 2019 International Conference on Information and Communications Technology (ICOIACT), pp. 829–834. doi:10.1109/ICOIACT46704.2019.8938550.

Lehman, B., Jackson, G.T. and Forsyth, C. (2019) 'A (Mis)Match Analysis: Examining the Alignment between Test Taker Performance in Conventional and Game-Based Assessments', Journal of Applied Testing Technology, 20(S1), pp. 17–34.

Lewthwaite, S. and James, A. (2020) 'Accessible at last?: what do new European digital accessibility laws mean for disabled people in the UK?', Disability & Society, 35(8), pp. 1360–1365. doi:10.1080/09687599.2020.1717446.

Lovett, B.J. (2020) 'Extended Time Testing Accommodations for Students with Disabilities: Impact on Score Meaning and Construct Representation', in Integrating Timing Considerations to Improve Testing Practices. Routledge.

Lovett, B.J. and Nelson, J.M. (2021) 'Systematic Review: Educational Accommodations for Children and Adolescents With Attention-Deficit/Hyperactivity Disorder', Journal of the American Academy of Child & Adolescent Psychiatry, 60(4), pp. 448–457. doi:10.1016/j.jaac.2020.07.891.

Maddox, B. (2017) 'Talk and Gesture as Process Data', Measurement: Interdisciplinary Research and Perspectives, 15(3–4), pp. 113–127. doi:10.1080/15366367.2017.1392821.

Martin, A.J. and Lazendic, G. (2018) 'Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience', Journal of Educational Psychology, 110(1), pp. 27–45. doi:10.1037/edu0000205.

Messick, S. (1990) Validity of Test Interpretation and Use. Available at: https://eric.ed.gov/?id=ed395031 (Accessed: 7 February 2021).

Mislevy, R.J. and Haertel, G.D. (2006) 'Implications of Evidence-Centered Design for Educational Testing', Educational Measurement: Issues and Practice, 25(4), pp. 6–20. doi:10.1111/j.1745-3992.2006.00075.x.

Mortimore, T. and Crozier, W.R. (2006) 'Dyslexia and difficulties with study skills in higher education', Studies in Higher Education, 31(2), pp. 235–251. doi:10.1080/03075070600572173.

Nagalingam, V. and Ibrahim, R. (2015) 'User Experience of Educational Games: A Review of the Elements', Procedia Computer Science, 72, pp. 423–433. doi:10.1016/j.procs.2015.12.123.

Nganji, J.T. and Brayshaw, M. (2017) 'Disability-aware adaptive and personalised learning for students with multiple disabilities', The International Journal of Information and Learning Technology, 34(4), pp. 307–321. doi:10.1108/IJILT-08-2016-0027.

Pekrun, R. (2006) 'The Control-Value Theory of Achievement Emotions: Assumptions, Corollaries, and Implications for Educational Research and Practice', Educational Psychology Review, 18(4), pp. 315–341. doi:10.1007/s10648-006-9029-9.

Pekrun, R. et al. (2010) 'Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion', Journal of Educational Psychology, 102(3), pp. 531–549. doi:10.1037/a0019243.

Penk, C., Pöhlmann, C. and Roppelt, A. (2014) 'The role of test-taking motivation for students' performance in low-stakes assessments: an investigation of school-track-specific differences', Large-scale Assessments in Education, 1(2), pp. 1–17. doi:10.1186/s40536-014-0005-4.

Sireci, S.G. (2020) 'Standardization and UNDERSTANDardization in Educational Assessment', Educational Measurement: Issues and Practice, 39(3), pp. 100–105. doi:10.1111/emip.12377.

Settles, B. et al (2020) 'Machine-learning Driven Language Assessment', Transactions of the Association for Computational Linguistics, 8, pp. 247–63. https://doi.org/10.1162/tacl_a_00310.

Stone, E. and Davey, T. (2011) 'Computer-Adaptive Testing for Students with Disabilities: A Review of the Literature', ETS Research Report Series, 2011(2), pp. i–24. doi:10.1002/j.2333-8504.2011.tb02268.x.

Thurlow, M. et al. (2010) Computer-Based Testing: Practices and Considerations. Synthesis Report 78, National Center on Educational Outcomes, University of Minnesota. National Center on Educational Outcomes. Available at: https://eric.ed.gov/?id=ED512613 (Accessed: 17 September 2021).

Tze, V.M.C., Daniels, L.M. and Klassen, R.M. (2016) 'Evaluating the Relationship Between Boredom and Academic Outcomes: A Meta-Analysis', Educational Psychology Review, 28(1), pp. 119–144. doi:10.1007/s10648-015-9301-y.

Wagner, E. (2020) 'Duolingo English Test, Revised Version July 2019', Language Assessment Quarterly, 17(3), pp. 300–315. doi:10.1080/15434303.2020.1771343.

Wagner, E. and Kunnan, A.J. (2015) 'The Duolingo English Test', Language Assessment Quarterly, 12(3), pp. 320–331. doi:10.1080/15434303.2015.1061530.

Webster, J. and Ahuja, J.S. (2006) 'Enhancing the Design of Web Navigation Systems: The Influence of User Disorientation on Engagement and Performance', MIS Quarterly, 30(3), pp. 661–678. doi:10.2307/25148744.