2022-06-13

The development and implementation of a computer adaptive progress test across European countries

Rice, N

10.1016/j.caeai.2022.100083
Computers and Education: Artificial Intelligence
Elsevier

The development and implementation of a computer adaptive progress test across European countries

Neil Rice, Carlos Fernando Collares, Jolanta Kisielewska, Thomas Gale

This is a pre-proof version of the manuscript published in 'Computers and Education: Artificial Intelligence'

Abstract

Longitudinal progress testing promotes self-directed deep learning across a full spectrum of knowledge, enabling early detection of underperformance and opportunities for remediation. Computer adaptive testing (CAT), where the difficulty of a test dynamically adjusts according to a test taker's ability, has benefits in a progress testing context, but significant resource and experience is required to develop appropriate test materials. This study describes how a transnational consortium from eight medical schools in five countries across Europe was formed to develop a computer adaptive progress test applicable across international curricula. 1,212 students from more than 40 nationalities took part in the study, of whom more than 70% were not native English speakers, though nearly all reported competence in English. A content map for an international assessment blueprint was agreed and a substantial bank of 1127 English language progress test items were successfully calibrated after pilot testing to form the computer adaptive progress test (CA-PT) item bank. Results from the CA-PT pilot showed reliable convergence to stable estimates of ability, low standard errors of measurement and high test reliability for all participants. This study shows that an international collaborative consortium approach enables effective development of progress testing resources appropriate for computer adaptive testing, with potential for application across international borders and in populations where English is not the native language. Pooling resources internationally facilitates comparison and development of appropriate assessment blueprints and the efficient generation of high-quality assessment items.

Keywords:

- *Cooperative/collaborative learning*
- *Cross-cultural projects*
- *Data science applications in education*
- *Distributed learning environments*
- *Learning communities*

1. Introduction

In 2018, a consortium of medical schools affiliated to the European Board of Medical Assessors (EBMA) secured an ERASMUS+ research grant (Strategic Partnerships – Call 2018) to develop an Online Adaptive International Progress Test (OAIPT) applicable across national borders and curricula. The first phase of the project involved establishing and equipping a consortium to develop a calibrated item bank suitable for computer adaptive progress testing (CA-PT). The second phase of the project involved developing and implementing a CA-PT algorithm appropriate for progress testing, operationalizing this in an accessible online test delivery platform, and delivering the test to international participants.

1.1 Background

An increasingly mobile international healthcare workforce requires quality assurance of medical training programmes across international borders to promote equality for healthcare workers and to maintain excellence in patient care (Crossley et al., 2002; Weggemans et al., 2017; Huang et al., 2019). However, the semi-autonomous administration of many medical schools implies substantial variability in assessment systems used and little comparability of graduate outcomes, even within countries with regulatory authorities (Wijnen-Meijer et al., 2013; Devine et al., 2015), making international comparisons complex.

Progress testing, where learners are repeatedly tested from the same comprehensive domain of knowledge, has been shown to promote deep learning styles, aid knowledge retention and offer opportunities for early identification of learners requiring support, with outcomes feasible for international comparisons (Blake et al., 1996; van der Vleuten et al.,

1996; Verhoeven et al., 2005; Freeman et al., 2010; Chen et al., 2015). However, developing good quality items that assess applied medical knowledge requires significant resource (Schuwirth et al., 2010). Regional progress test consortia have emerged in some countries aiming to pool resources to reduce the burden of test item generation at an institutional level and ultimately to improve standards of assessment across members (Wrigley et al., 2012). Establishing international progress test consortia could be beneficial for equipping schools in graduating doctors prepared and confident to work in international healthcare settings.

There is evidence that progress testing consortia have been successful and enduring, both when organised at faculty level for assessment quality control and standard setting, for example in The Netherlands (Schuwirth et al., 2010; Tio et al., 2016), or when driven by students for formative learning purposes as in the Progress Test Medizin in Germany and Austria (Nouns and Georg, 2010). Previous studies have also shown that the use of progress test material from one country can form a basis for developing items in another country (Swanson et al., 2010).

Computer adaptive testing (CAT) provides opportunities to tailor tests to individual learners using an algorithm to dynamically select the difficulty of items based on the learner's performance in previous items, converging on a learner's true knowledge level (Collares and Cecilio-Fernandes, 2019). There is evidence that CAT increases test reliability and improves learner motivation and engagement compared to paper-based tests (Martin and Lazendic, 2018), and CAT has been recommended for the measurement of progress (Shapiro et al.,

2015; Nelson et al., 2017). Whether these positive effects are generalizable to professional programmes of study such as medicine has not been investigated.

The success of CAT depends on a large item bank of calibrated items of varying difficulty covering, proportionally, all areas of the underlying test blueprint, with at least 1000 items (10 items per blueprint dimension as a rule of thumb) suggested for high-stake assessments (Wise and Kingsbury, 2000). To this end, an international collaborative framework for item bank development becomes extremely helpful (van der Vleuten et al., 2018). A large item bank is also required to counter the potential issue of item overexposure, which can be mitigated further by well-developed CAT item selection algorithms. Though CAT based on item response theory (IRT) was first proposed over 40 years ago (Lord, 1980), the viability of adaptive testing with large item banks was dependent on the speed with which testing systems could select relevant test items (Huang et al, 2009).

One of the fundamental considerations in developing any knowledge test is defining the universe of knowledge that is to be assessed (Haladyna and Rodriguez, 2013). Content validity is a requirement of every evaluation and is achieved when the test content is consistent with the learning objectives and the learning experiences (Webb, 2006). A systematic blueprinting approach during test development is widely accepted as facilitating validity (Bridge et al., 2003) and meeting learners experience with their curriculum (Coderre et al., 2009), but blueprinting an international medical assessment across curricula, culture and language is particularly challenging (Wrigley et al., 2012).

Generating high quality assessment items is a critical step to realise an item bank with adequate psychometric properties to enable CAT, but it is a complex task (Koller et al., 2017). Items need to cover a wide range of sampling characteristics such as content and difficulty while maintaining a high degree of discriminatory power. Poor quality items can lead to undesirable psychometric proprieties and impact on assessment integrity (Jozefowicz et al., 2002; Downing, 2005). Reviewing is essential to ensure that items respect the guidelines and achieve the high quality that is demanded (van der Vleuten et al., 1996; Wallach et al., 2006).

1.2 Research Objectives

The objectives of this study are to describe and evaluate the development and implementation of the project in terms of: i) the development and reaching of consensus on an international assessment blueprint; ii) the generation of high quality test items sampling this blueprint; iii) the establishment of a substantial calibrated item bank with stable psychometric properties; and iv) the development, delivery and evaluation of the CA-PT.

2. Methods

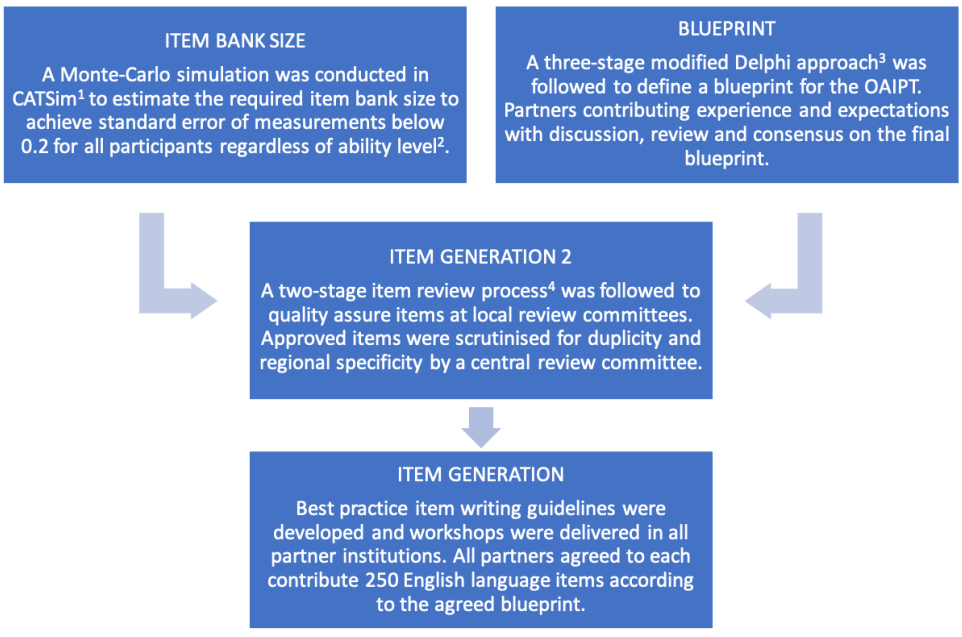
2.1 Consortium development

In August 2018, medical educators and assessment specialists from eight medical schools in five European countries who were institutional members of the EBMA formed the OAIPT consortium. Four schools had extensive experience with progress testing in their own undergraduate medical education programmes. The other four schools had little or no experience with progress testing.

2.2 OAIPT blueprint and item bank development

The consortium developed a blueprint and item-bank for the international progress test as outlined in figure 1.

Figure 1: OAIPT resource development



1. Weiss and Guyer, 2010; 2. Weiss, 2011; 3. Coderre et al., 2009; 4. van der Vleuten et al., 1996.

The consortium partners collaboratively agreed a terminology of disciplines and categories, the inclusion/exclusion of country specific content and the clarification of terms with regional specificities (e.g. equivalence of family medicine versus primary care). It was agreed that a test should consist of 125 items written in English covering 16 disciplines across 12 categories. A final blueprint design for the OAIPT was achieved. Following the successful delivery of item writing workshops at all eight partner institutions, a total of 1938 items were generated for the OAIPT item bank. Upon completion of the review processes 1422 items were considered acceptable for item calibration pilot testing. Figure 2 shows the agreed blueprint disciplines and categories for a single progress test, with the distribution of the number of items following the item writing and review processes.

Figure 2: OAIPT agreed blueprint and reviewed item bank distribution.

Category / Discipline	Respiratory system	Musculoskeletal system	Mental Health Care	Reproductive system	Blood, lymph, heart and circulation	Hormones and metabolism	Skin and connective tissue	Personal, social and prevention aspects	Digestive system	Kidneys and urinary tract	Nervous system and senses	Knowledge about skills	**Total**	**Accepted Items after Review**
Anatomy	1	1		1	2	1			1	1	1		**9**	**97**
Biochemistry, molecular and cellular biology and genetics	1	1		2	2	1			1	1	1	2	**12**	**115**
Surgery	1	2			1	1	1		1	2	1	1	**11**	**129**
Dermatology, ENT, ophthalmology	1						4	1			3	1	**10**	**100**
Epidemiology, methodology and statistics								3					**3**	**40**
Pharmacology	1	1			1	1			1	1		1	**7**	**87**
Physiology	1	1		1	2	1			1	1	1		**9**	**108**
Geriatrics			1		1			1					**3**	**47**
Obstetrics and gynaecology				2		1						1	**4**	**44**
Family medicine	1	1		1	2		1	1	1	1	1	3	**13**	**172**
Internal medicine	3	1			4	1	1	1	2	2	1	1	**17**	**194**
Paediatrics	1			1	1	1		1	1	1	1	1	**9**	**102**
Meta-medical aspects								1					**1**	**7**
Neurology		1									3		**4**	**46**
Pathology	1	1		1	1		1		1	1	1		**8**	**86**
Psychiatry and medical psychology			3					1					**4**	**38**
Social medicine								1					**1**	**10**
Total	**12**	**10**	**4**	**9**	**17**	**8**	**8**	**11**	**10**	**11**	**14**	**11**	**125**	
Accepted Items after Review	**117**	**121**	**56**	**86**	**214**	**105**	**92**	**110**	**143**	**110**	**154**	**114**		**1422**

2.3 Pilot testing for item bank calibration

The initial plan was for three non-adaptive pilot tests for item-bank calibration in the first year of the project but owing to the Covid-19 pandemic one pilot test had to be removed. As such, two non-adaptive pilot tests (pilot 1 and pilot 2) were administered in January 2020 (pilot 1) and January 2021 (pilot 2) at all partner institutions and data from these pilots were analysed by the project psychometricians. Items which could not be validated in pilot 1 were either retired or entered into the pool of items for pilot 2 tests. Data from the two pilot tests were then combined for item parameter calibration.

A non-equivalent groups with anchor test (NEAT) equating design was employed where common anchor items were embedded across test forms to enable linkage for concurrent calibration of items across test forms (Raykov, 2010; Von Davier, 2011). Using the NEAT design enables all the item parameters to have the same metrics across all groups, i.e. the item parameters in one sample are directly comparable with item parameters in all other samples. Eight separate test forms of 125 items were constructed each covering one full progress testing blueprint. A subset of 35 items representative of the blueprint was chosen to form a common anchor item set to be included in all test forms. Each test form also included two pairwise anchor sets of 20 items. This equating design allowed for the administration of 595 items in pilot 1, distributed across test forms as shown in table 1.

For pilot 1, students were recruited by each partner institution from all stages of medical study, and tests were scheduled and delivered online under locally invigilated exam conditions using the TestLife platform hosted by Maastricht University. Students were asked

to answer all items in the pilot tests and were given the opportunity (within the software) to make comments on any items that they wished to query during the test. Any items which were flagged with student comments during the pilot tests were reviewed by the local partner responsible for the item, and where necessary, either removed or reworked and administered again in pilot 2.

Since one of the planned calibration pilots had to be cancelled due to the Covid-19 pandemic, the equating design was revised between pilot 1 and pilot 2 to enable the inclusion of more items in the calibration study. Eight test forms of 125 items each were constructed for use in pilot 2. The number of items in the full anchor item sets (in all papers) was reduced, and anchor items were distributed across the test forms, with more unique items used in each test. Consequently, 715 items were administered across all pilot tests in pilot 2 (table 1).

Table 1: NEAT equating designs for allocation of items to pilot test forms

		Unique Items	Anchor Item Sets								Total items
			Common	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	
Pilot 1	Test 1	50	35		20	20						125
	Test 2	50	35	20			20					125
	Test 3	50	35	20				20				125
	Test 4	50	35		20				20			125
	Test 5	50	35			20				20		125
	Test 6	50	35				20				20	125
	Test 7	50	35					20			20	125
	Test 8	50	35						20	20		125

Pilot 2	Test 1	65	15		10	5	5	5	5	5	10	125
	Test 2	65	15	10		10	5	5	5	5	5	125
	Test 3	65	15	5	10		10	5	5	5	5	125
	Test 4	65	15	5	5	10		10	5	5	5	125
	Test 5	65	15	5	5	5	10		10	5	5	125
	Test 6	65	15	5	5	5	5	10		10	5	125
	Test 7	65	15	5	5	5	5	5	10		10	125
	Test 8	65	15	10	5	5	5	5	5	10		125

A further impact of the Covid-19 pandemic was that after pilot 1 tests could no longer be viably delivered using face-to-face invigilation as planned. Subsequent tests were instead delivered on the University of Minho's QuizOne® (http://quiz.one) testing platform which provides video proctoring functionality. Partners received training in remote proctoring and administered the pilot tests to their own students. In total 1219 items were administered across the two pilot tests including 91 items used in both pilots.

2.3.1 Calibration pilot testing analysis

One of the critical aspects in item-bank construction is that items should work in the same way for different populations. Parameter non-invariance is commonly called Differential Item Functioning (DIF) when we are considering pairwise comparisons. For group comparisons there are other methods available, such as multigroup confirmatory factor analysis and multiple indicators multiple causes analysis (MIMIC). For this study, we opted to use the

measurement alignment procedure, because of its approximate measurement invariance approach (Kim, 2017; van de Schoot, 2013), which conveniently allows for the use of more items while still testing for construct-irrelevant score variance (CIV). CIV was tested using a measurement alignment model in Mplus v8.4 (Muthén and Muthén, 2019).

A one parameter Rasch model under joint maximum likelihood (JML) estimation was conducted in both Winsteps (Linacre, 2021) and XCalibre (Thompson and Jieun, 2017) to estimate item parameters (Baghaei, 2008). Though it has been suggested that a RASCH model under marginal maximum likelihood (MML) generates more consistent estimations, JML has been shown to be a defensible choice and computationally efficient for large item response datasets (Chen et al., 2019), and at the time of the analyses allowed for replication between XCalibre and Winsteps (as Winsteps only offered JML estimation).

The Rasch model requires the underlying data to have certain properties one of which is that success or failure on any individual item should not depend on the test takers' performance on any other item (conditional or local independence). Item residual correlations were calculated using Winsteps to enable testing for local dependence.

The exclusion of items displaying local dependence improves compliance to the Rasch model assumption of unidimensionality: that all items measure a single latent construct. Unidimensionality was confirmed for both pilot studies using a principal component analysis of the residual item correlation matrix of the Rasch model. The large number of items in the analysis rendered using the widely accepted threshold for absolute unidimensionality of having an eigenvalue of the first residual contrast below 2.0 unfeasible. Instead, a test of

“essential unidimensionality” (Strout, 1990) was considered using the criteria that the first contrast should not explain more than 5% of the residual variance.

Items which could not be calibrated were flagged for remedial work and administration in p2, or where necessary removed from the item bank. In total, 1127 items (92.5% of items piloted) were successfully calibrated in phase 1 and formed the item bank for the phase 2 CA-PT pilot.

2.4 Computer Adaptive Progress Test Pilot

CAT requires the specification of five core components: the item bank; the starting point - an initial theta for individuals taking the test which may be fixed, randomized or evidence-based from data; the item selection algorithms – including blueprint content distribution, item exposure, and enemy items; the scoring algorithms – for example (weighted) maximum likelihood (MLE) estimation or Bayesian estimation; the termination criteria – such as a target standard error of measurement for an individual (Thompson & Weiss, 2011).

The Computer Adaptive Progress Testing (CA-PT) algorithms for item selection and theta estimation were developed using the open source R package catR (Magis & Barrada, 2017; R Core Team, 2020) which was embedded into the QuizOne testing platform. While catR enables selection of items from specific item sets, it does not currently facilitate item selection according to blueprints within an item bank. In order to sample items from across the progress test blueprint, we modified the catR algorithms to sequentially select items from subsets according to the pre-defined blueprint. Besides improving content validity this ensured that enemy items were not included.

Because we had no prior information on participants ability, we adopted a modified version of the starting rule. All participants in the pilot were presented with the same non-adaptive pool of 30 items (selected from items with good discrimination and intermediate difficulty parameters (b parameters in the range -1 to 1) at the start of their test to establish individual starting thetas for the CA-PT.

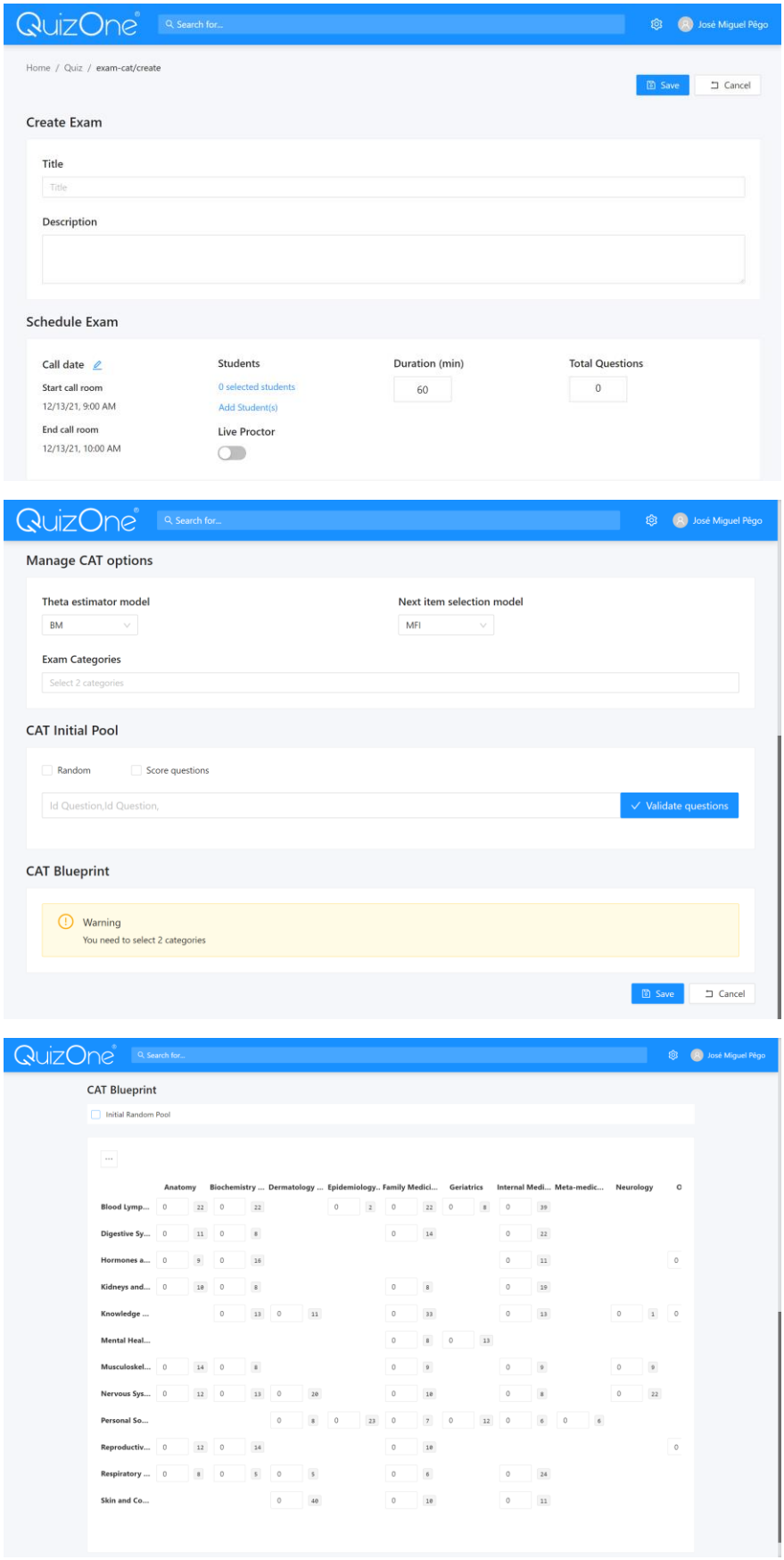
Weighted maximum likelihood estimation (WMLE) was used for theta estimation with the Maximum Fisher Information (MFI) employed for next item selection according to the minimum defined items for each unique dimension of the blueprint. WMLE overcomes to some extent the normality assumption constraint of MLE and is in principal more robust against non-normality, greatly reducing estimation bias (Warm, 1989). The WMLE with MFI method is the most computationally efficient in terms of speed of delivering the next item, and it has been shown that for longer tests there is no difference between the choice of item selection method in terms of the items the algorithm selects (Sulak and Kelecioğlu, 2019; Chen et al 2000).

The stopping rule for this pilot was a fixed test length (125 items) to ensure acceptable levels of measurement error for all participants regardless of their ability level and also to provide a full representation of the progress test blueprint.

The CA-PT pilot was administered on QuizOne using a CA-PT module specifically developed for the project. This module included an interface for the test design, a CA-PT engine and a test administration interface for students that was similar to a non-adaptive test. The

interface for the test design allows for the configuration of title, description, scheduling, theta estimator and next item selection models, blueprint categories, initial pool of items and the number of items required in a test by unique dimension of the blueprint. Figure 3 shows the interface for the CA-PT module developed in QuizOne.

Figure 3: The CA-PT design interface in QuizOne



2.5 Student recruitment, test administration and participant feedback

Ahead of each pilot test the consortium developed information materials (e.g. https://www.youtube.com/watch?v=-epx45BdKyc), test delivery guidelines and short videos for participants and test invigilators. These were updated to account for the different delivery methods between the pilots. Students who attended the pilot tests were invited to complete a feedback questionnaire about their test experience, to provide sociodemographic data, and to answer questions relating to their English language competence. Participants received individual feedback reports on their performance after each pilot test.

2.6 Student Participation

1,213 students participated in the study: 889 in the item calibration pilot tests (408 in pilot 1 and 481 in pilot 2) and 324 in the CA-PT pilot (table 2). The overall response rate to the post-test feedback questionnaires was 78%. One third of respondents declared English as their first language but more than 99% self-reported as proficient in their comprehension of English.

Table 2: Participants in the pilot studies by OAIPT partner country

Partner Country	**Finland**	**Netherlands**	**Poland**	**Portugal**	**UK**	**Total**
Pilot 1 Participants	59	156	47	66	80	**408**
Pilot 2 Participants	46	182	48	83	122	**481**
Pilot 3 (CA-PT) Participants	11	85	41	47	140	**324**
Total Participants	**116**	**423**	**136**	**196**	**342**	**1213**

3. **Results and Discussion**

3.1 Item Calibration Pilot Testing Results

In tests for local independence, item residual correlations above .70 were considered unacceptable, as item pairs with higher residual correlation interfere with the validity of test scores. In pilot 1 50 items displayed local dependence and another 15 items had zero variance (i.e. all students answered correctly or incorrectly) and these items were administered again in pilot 2. After pilot 2, 73 items displayed local dependence and five items had zero variance and were excluded from the final calibration. Most items removed from the calibration owing to local dependence or zero variance were unique items in tests with low participant numbers (n<25).

Anchor items were tested for measurement non-invariance across countries using a measurement alignment procedure. Data from partners in the same country (i.e. two schools in the UK, Netherlands and Poland) were combined to increase sample size for these analyses, as the measurement alignment procedure can lead to “false negatives” when the sample size per group is low, even when effect sizes are large. For pilot 1 we tested alignment using a fixed and a random factors approach separately. The results in terms of the indication of the non-invariant items and the mean scores for each country (group) were almost identical from both methods. For the second round we chose the random factors approach because the results are more generalisable and the groups are assumed to not be inherently different. The measurement alignment procedure explores the patterns of non-invariance across groups by searching the simplest possible data structure with the least number of items with the largest non-invariance which is similar to a rotation in an exploratory factor analysis (Kim et al., 2017).

The number of participants recruited to the study was lower than expected, but the flexible equating design employed enabled a successful and robust calibration of items despite the small sample size in some schools. Only 6% of items could not be calibrated owing to low pilot test participation rates in one or two partner institutions and less than 2% of piloted items were removed from the item bank owing to measurement alignment, local dependence or issues identified by test takers. However, the equating design did not enable all items to be tested for invariance and it may be that these items will actually show non-invariance in subsequent recalibrations.

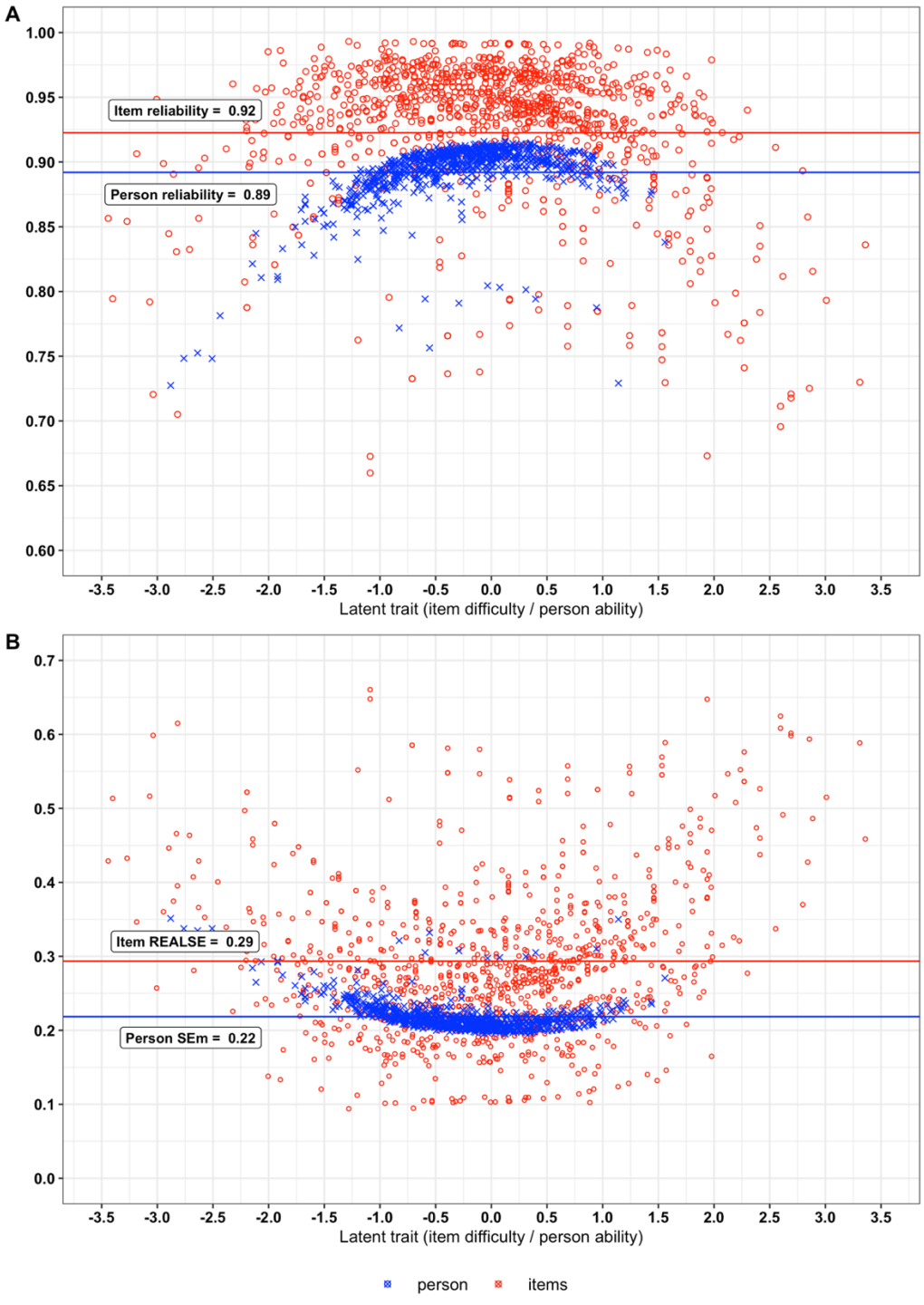
The reasons for non-invariance of some items in the item bank merits further investigation. If item non-invariance is a reflection of cultural differences for example, then there is little that can be done to make the item applicable internationally. However, if non-invariance is attributable to differences in curriculum coverage across partners, there are opportunities for partners to work collaboratively to develop teaching and enhance learning opportunities. In this way international testing may be a tool for educational development, which would be a very important deliverable from the item bank.

Ten items showed CIV in p1 and were reused in p2. With the increased sample size, only two items displayed significantly different difficulty parameters across the five participating countries in p2 and were excluded from the final calibration. Across both pilot tests participants commented on 235 items of which 16 were excluded from the item bank following review.

Essential but not absolute unidimensionality was evidenced in both pilot studies and in the final concurrent calibration using the Rasch model analysis of the 1127 items which remained after all exclusions. The first contrast (residual component) explained just 0.7% of the total score variance and none of the residual components explained more than 1% of the variance. However, measures (items and persons) only explained about 25% of the variance, somewhat lower than the expected from typical Rasch model analyses. It should therefore be noted that despite the acceptably low explained variance in the first contrast, multidimensionality of the data could not be explicitly ruled out. Cognitive diagnostic adaptive testing (see section 3.3 below) seeks to address this potential issue.

Item difficulty parameters ranged from -4.5 to 5.3 with a good distribution of items across the latent difficulty score range. Mean-square (MSQ) fit statistics were examined all 1127 items had acceptable infit MSQ between 0.7 and 1.4. The average person standard error of measurement inflated for misfit was acceptably low at 0.218 and was relatively stable across the central range of person abilities (figure 4b). Average conditional individual reliability was high at 0.892 and also relatively stable across the midrange of the latent scale of person ability (figure 4a).

Figure 4: Distribution of item and person reliability and standard errors across the latent scale



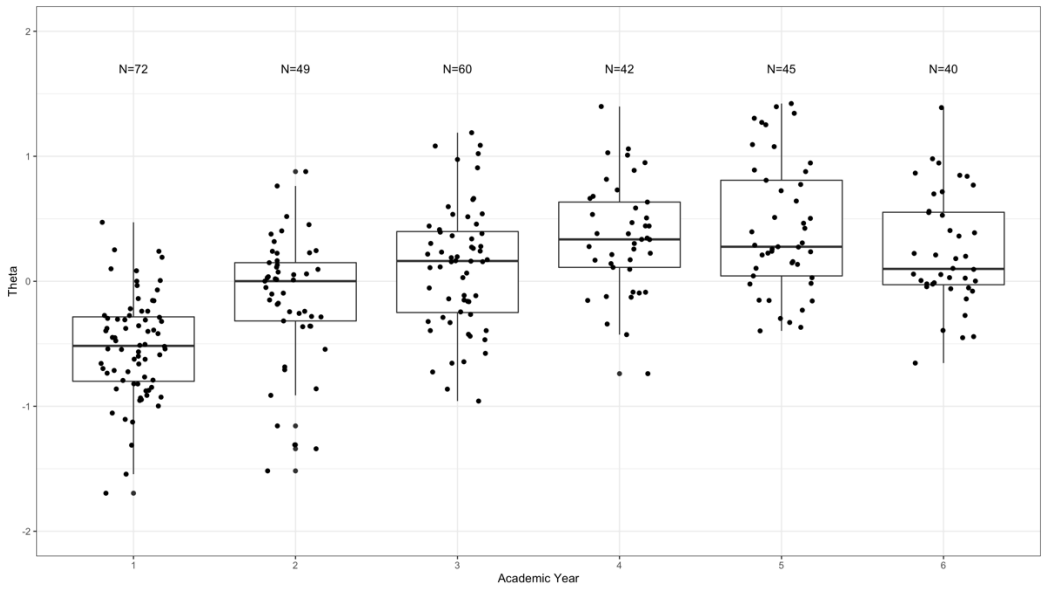
The pool of items developed and reviewed by the consortium are of good quality with stable psychometric properties. The fit of the items in the calibration model suggest that the estimates of the latent difficulty are a good representation of the response patterns. The

robust person standard errors of measurement inflated for misfit were small and did not vary greatly across the latent theta scale, which is an important and desirable feature of the calibration, providing evidence that the item bank produces reliable tests across populations.

3.2 CA-PT pilot testing results

The sample of participants in the CA-PT pilot, though smaller than expected, covered all stages of study. The distribution of scores by year was in line with previous progress testing data – showing diminishing rates of growth by seniority of students, with all theta person ability estimates between +/- 2 (figure 5).

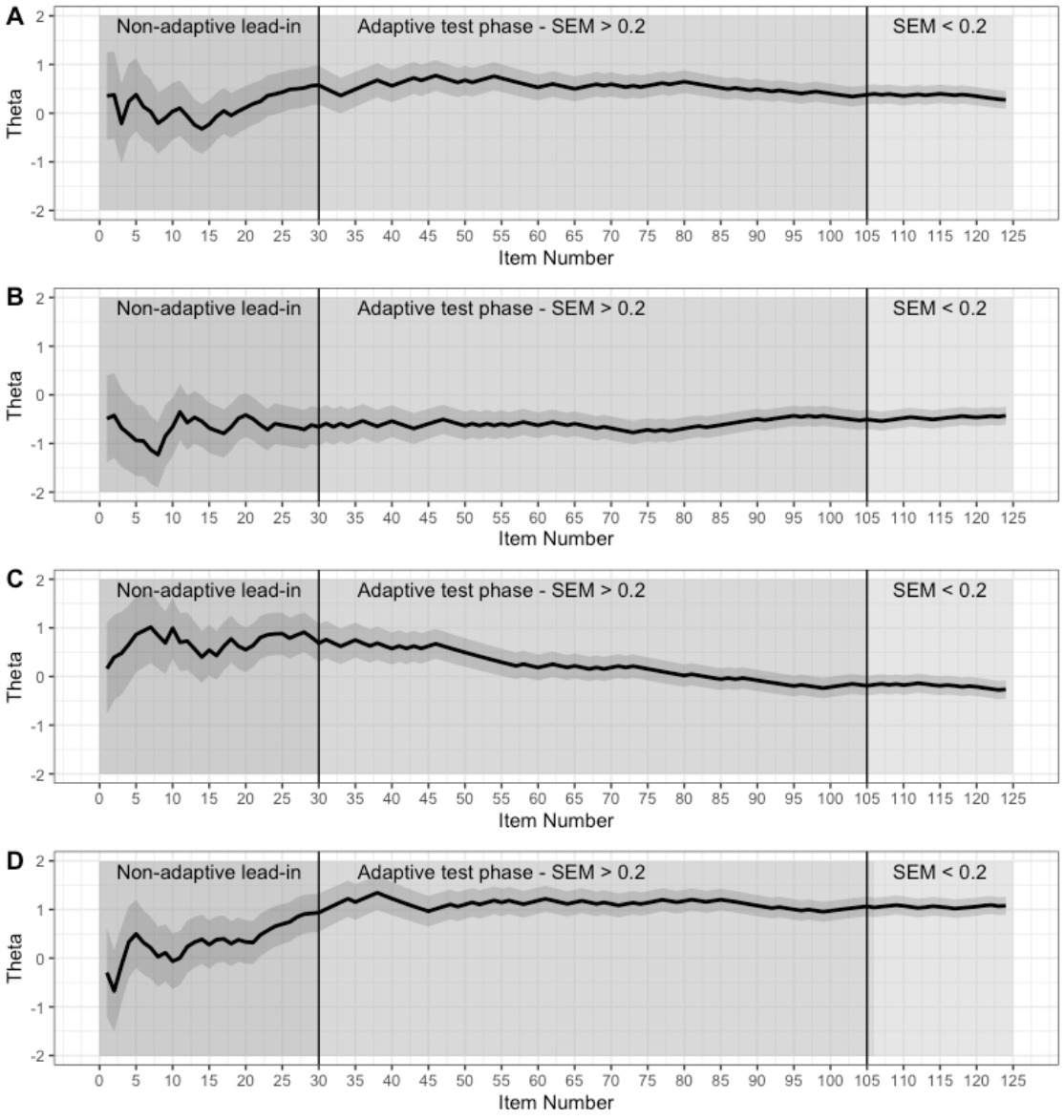
Figure 5: Distribution of CA-PT theta estimates by academic year of participants



We were able to successfully adapt and develop the open source catR item selection algorithms to select items according to our predefined progress testing assessment blueprint. These algorithms were built into the QuizOne test delivery platform and successfully implemented to deliver adaptive tests which showed convergence to

acceptably low standard errors of measurement for all study participants. Figure 6 shows data from four example participants with the line plot showing the convergence of participants' theta estimates and the shaded ribbon showing the associated SEM at each item. The vertical lines indicate where the adaptive phase of the test starts (item 31) and the point in the test where the person SEMs reduce to 0.2 for each student.

Figure 6: CA-PT convergence plots for four example participants



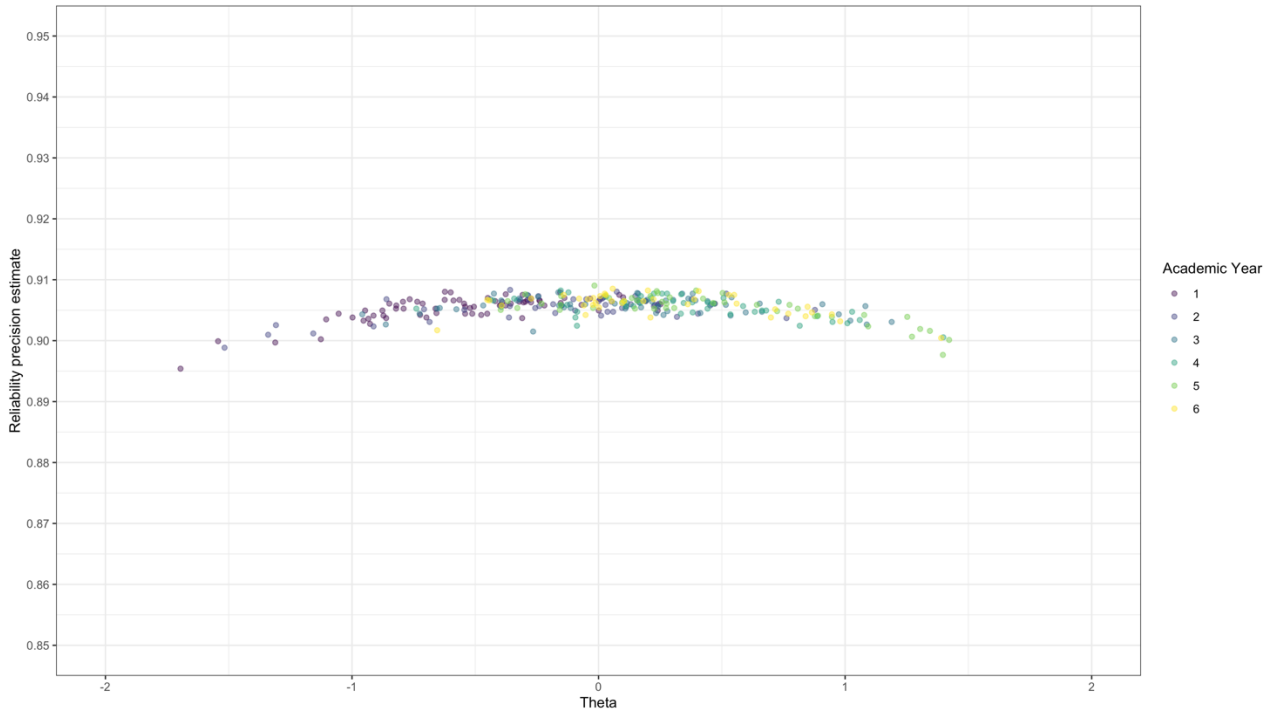
The 30 item non-adaptive lead-in enabled an estimation of the initial theta for the adaptive testing phase, but also promoted a balanced emotional reaction at the beginning of the test. The use of the non-adaptive lead-in and the fixed number of items stopping rule for this CA-PT pilot ensured a full coverage of the test blueprint whilst enabling the collection of comprehensive performance data to enable robust analysis of convergence to true thetas and SEMs for all participants. In a longitudinal CA-PT it is anticipated that the stopping rule could be determined with fewer items by the convergence to a stable low standard error of measurement for each participant, and this is often cited as a key advantage of adaptive testing. Indeed, the results of this CA-PT pilot study suggest that on average around 75 items are required to achieve convergence to a SEM of less than 0.2 (after an initial lead-in of 30 non-adaptive items) for this progress test (table 3; figure 6). However, using a variable test length stopping rule in practice can prove problematic for tests of knowledge covering a large blueprint as some candidates achieve a very low standard error with very few items administered using the adaptive algorithm, and hence, their tests do not adequately cover the testing domains required. The potential for students to receive different levels of testing can be problematic in terms of communicating performance and outcomes. This has particular implications in the context of assessment programmes built on the principles of assessment for learning, often programmes which adopt a progress testing framework.

Table 3: CA-PT pilot test results

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
Avg Theta @ CA-PT start (Item 31)	-.346	-.057	.086	.263	.406	.333
Avg SEM @ CA-PT start (Item 31)	.375	.378	.377	.377	.380	.377
Avg Theta @ CA-PT end	-.523	-.141	.116	.287	.421	.235
Avg SEM @ CA-PT end	.183	.183	.183	.184	.183	.183
Avg Item Number when SEM < 0.2	104.6	104.6	103.9	104.6	104.3	104.0

A major purported advantage of adaptive testing, especially applicable in the context of progress testing, is greatly improved individual test reliability of across the spectrum of ability. In non-adaptive progress tests of end-point knowledge, students in the earlier stages of study often cannot answer many items leading to tests with low reliability. Figure 7 shows the person reliability estimates from the CA-PT pilot by academic year of the study participants. It can be observed here that CA-PT test reliability is high (over .89) for individuals across the theta ability spectrum, representing significant improvements when compared with non-adaptive test reliability especially in the extremes of the theta spectrum (figure 4a).

Figure 7: CA-PT Person reliability by year of study



3.3 Study constraints and future directions

The study design planned for three potential rounds of pilot testing in the item calibration phase and two adaptive pilot tests. Owing to the Covid-19 pandemic, one full round of pilot testing and one stage of adaptive testing could not be realised, and the rates of participation in the pilots were much lower than expected in all schools. Required sample sizes of around 100 participants per school per pilot test were estimated to ensure good coverage of overlapping items in the equating design, and to appropriately power the measurement alignment models. Whilst local dependence and measurement alignment across countries in the study could be estimated and stable item parameters were established, it may be an overgeneralisation at this stage to suggest that this item pool is applicable for progress testing internationally or across populations with different native languages.

Being exploratory and developmental in nature, the study did not consider item exposure control parameters. If the CA-PT were to be fully deployed, item exposure analyses would an important consideration, for example to preserve the integrity of the items in the bank and to address the issue of item parameter drift over time.

A richer appraisal of the applicability of adaptive testing in a progress testing framework would include a longitudinal analysis of psychometric properties, but this was not the purpose of the project and was not possible in this study for reasons already identified. In future iterations of the adaptive progress test we aim to develop and implement our blueprint-based item selection algorithms using cognitive diagnostic modelling (CDM) to be able to tailor adaptive progress tests not just based on single time point item difficulty and

person ability, but also to take into account previous performance in blueprint domains. In this way, the adaptive progress test will have memory and will be able to oversample on content areas where students need to grow, whilst endorsing and challenging students in content areas where they have previously demonstrated mastery. When compared to scores and subscores based on a unidimensional IRT solution, CDM has been shown to provide more detailed information on students' strengths and weaknesses, which seems to be particularly applicable in developing tools of assessment for learning. (Kaplan et al, 2015).

3.4 Lessons Learned

To the best of our knowledge this is the first report of a transnational progress testing consortium involving multiple countries with different cultural backgrounds and native languages. Collaborative approaches to item development and the related potential improvements in item quality, standardisation and potential workload benefits are increasingly relevant in a pandemic impacted world. The main advantages of establishing consortia to develop large item banks are efficiency and complementarity. Notwithstanding the relative scarcity of psychometrician know-how, independently, schools would struggle to dedicate the significant resource required to develop, review, and calibrate a large enough set of items to build a CAT item bank. The creation of this consortium has demonstrated that, with the support of an established network and a collaboratively agreed framework, even inexperienced institutions are able to rapidly and efficiently adhere to a model to develop and deliver quality progress test items.

An exemplar of the benefits of a pooled resource consortium approach has been the successful reaction to the COVID-19 pandemic which has impacted medical education across Europe since early 2020 (Gill, 2020). In response to varying national lockdowns and restrictions, the expertise and resource within the consortium facilitated a move from face-to-face supervised administrations to remotely proctored online tests during the project. This change afforded partners greater flexibility in their local delivery.

It has been suggested that learners may be more accepting of formative progress testing in less rigid exam conditions (Karay et al., 2020), and it is anticipated that adaptive progress

testing will be well received by learners. This study has demonstrated that adaptive progress tests are reliable across the spectrum of participant ability, converging to stable theta ability estimates with low levels of measurement error in around a third the length of a traditional progress test. Subsequent research from this project will consider student satisfaction with and acceptability of the adaptive progress test.

4. Conclusions

This study has shown that adopting a collaborative institutional approach and appropriate methodology facilitates the rapid and effective implementation of computer adaptive progress testing, even across international borders and across populations where English is not the native language. Pooling resource and expertise encourages collaborative comparison and development of appropriate assessment blueprints and high-quality unbiased assessment items, even in schools with little or no experience of progress testing. Progress testing in this way could be a valuable tool to evaluate the medical knowledge of students and graduates from different countries facilitating planning and movement of health service workers internationally.

References

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transactions*, 22, 1145-1146.

Blake, J. M., Norman , G. R., Keane, D. R., Barber Mueller, C., Cunnington, J., & Didyk, N. (1996). Introducing progress testing in McMaster University's problem-based medical curriculum: Psychometric properties and effect on learning. *Academic Medicine,* 71(9), 1002–1007.

Bridge, P.D., Musial, J., Frank, R., Roe, T., & Sawilowsky, S. (2003). Measurement practices: methods for developing content-valid student examinations. *Medical Teacher,* 25(4), 414-421.

Chen, S.-Y., Ankenmann, R. D., & Chang, H. H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, 24(3), 241-255.

Chen, Y., Li, X. & Zhang, S. (2019). Joint Maximum Likelihood Estimation for High-Dimensional Exploratory Item Factor Analysis. *Psychometrika*, 84, 124–146.

Chen, Y., Henning, M., Yielder, J., Jones, R., Wearn, A., & Weller, J. (2015). Progress testing in the medical curriculum: students' approaches to learning and perceived stress. *BMC Medical Education,* 15(147).

Coderre, S., Woloschuk, W., & McLaughlin, K. (2009). Twelve tips for blueprinting. *Medical Teacher,* 31(4), 322-324.

Collares, C. F., & Cecilio-Fernandes, D. (2019). When I say ... computerised adaptive testing. *Medical Education,* 53(2), 115-116.

Crossley, J., Humphris, G., & Jolly, B. (2002). Assessing health professionals. *Medical Education,* 36(9), 800-804.

Devine, O. P., Harborne, A. C., & McManus, I. C. (2015). Assessment at UK medical schools varies substantially in volume, type and intensity and correlates with postgraduate attainment. *BMC Medical Education,* 15(146).

Downing, S. M. (2005). The Effects of Violating Standard Item Writing Principles on Tests and Students: The Consequences of Using Flawed Test Items on Achievement Examinations in Medical Education. *Advances in Health Sciences Education: Theory and Practice,* 10, 133–143.

Freeman, A., Van Der Vleuten, C., Nouns, Z., & Ricketts, C. (2010). Progress testing internationally. *Medical Teacher,* 32(6), 451-455.

Gill, D., Whitehead, C., & Wondimagegn, D. (2020). Challenges to medical education at a time of physical distancing. *Lancet,* 396(10244), 77-79.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and Validating Test Items*. Routledge.

Huang, P., Haywood, M., O'Sullivan, A., & Shulruf, B. (2019). A meta-analysis for comparing effective teaching in clinical education. *Medical Teacher,* 41(10), 1129-1142.

Huang, Y. M., Lin, Y. T., Cheng, S. T. (2009). An adaptive testing system for supporting versatile educational assessment. *Computers & Education,* 52(1), 35-67.

Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The Quality of In-house Medical School Examinations. *Academic Medicine,* 77(2), 156-161.

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New Item Selection Methods for Cognitive Diagnosis Computerized Adaptive Testing. *Applied Psychological Measurement*, 39(3), 167–188.

Karay, Y., Reiss, B., & Schauber, S. K. (2020). Progress testing anytime and anywhere – Does a mobile-learning approach enhance the utility of a large-scale formative assessment tool? *Medical Teacher,* 42(10), 1154-1162.

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement Invariance Testing with Many Groups: A Comparison of Five Approaches. *Educational and Psychological Studies Faculty Publications*, 194.

Koller, I., Levenson, M. R., & Glück, J. (2017). What do you think you are measuring? A mixed-methods procedure for assessing the content validity of test items and theory-based scaling. *Frontiers in Psychology,* 8(126).

Linacre, J. M. (2021). *Winsteps® Rasch measurement computer program*. Beaverton, Oregon: Winsteps.com.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. New Jersey: Lawrence Erlbaum.

Magis, D., & Barrada, J. R. (2017). Computerized Adaptive Testing with R: Recent Updates of the Package catR. *Journal of Statistical Software, Code Snippets*, 76(1), 1–19

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology,* 10(1), 27–45.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide. Eighth Edition*. Los Angeles, CA: Muthén & Muthén.

Nelson, P. M., Van Norman, E. R., Klingbeil, D. A., & Parker, D. C. (2017). Progress monitoring with computer adaptive assessments: the impact of data collection schedule on growth estimates. *Psychology in the Schools,* 54(5), 463–471.

Nouns, Z. M., & Georg, W. (2010). Progress testing in German speaking countries. *Medical Teacher,* 32(6), 467-470.

Raykov, T. (2010). Test Equating Under the NEAT Design: A Necessary Condition for Anchor Items. *Measurement: Interdisciplinary Research and Perspectives,* 8(1), 16-20.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schuwirth, L., Bosman, G., Henning, R., Rinkel, R., & Wenink, A. (2010). Collaboration on progress testing in medical schools in the Netherlands. *Medical Teacher,* 32(6), 476–479.

Shapiro, E. S., Dennis, M. S., & Fu, Q. (2015). Comparing computer adaptive and curriculum-based measures of math in progress monitoring. *School Psychology Quarterly,* 30(4), 470–487.

Strout, W. F. (1990). A new item response theory modelling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika,* 55, 293-325.

Sulak, S. & Kelecioğlu, H. (2019). Investigation of item selection methods according to test termination rules in CAT applications. *Journal of Measurement and Evaluation in Education and Psychology*, 10(3), 315-326.

Swanson, D. B., Holtzman, K. Z., Butler A., Langer, M. M., Nelson, M. V., Fuller, R. et al. (2010). Collaboration across the pond: The multi-school progress testing project. *Medical Teacher,* 32(6), 480-485.

Thompson, N. A., & Jieun, L. (2017). "Xcalibre 4". In: Van der Linden WJ ed. *Handbook of Item Response Theory.* Boca Raton: CRC Press. Routledge Handbooks Online.

Thompson, N. A., & Weiss, D. A. (2011). A Framework for the Development of Computerized Adaptive Tests. *Practical Assessment, Research & Evaluation*, 16(1).

Tio, R. A., Schutte, B., Meiboom, A. A., Greidanus, J., Dubois, E. A., & Bremers, A. J. (2016). The progress test of medicine: the Dutch experience. *Perspectives on medical education*, 5(1), 51-55.

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J. & Muthén, B. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*. 4, 770.

van der Vleuten, C. P. M., Verwijnen, G. M., & Wijnen, W. H. F. W. (1996). Fifteen years of experience with progress testing in a problem-based learning curriculum. *Medical Teacher,* 18(2), 103–110.

van der Vleuten, C., Freeman, A., & Collares, C. F. (2018). Progress test utopia. *Perspectives on Medical Education,* 7(2), 136–138.

Verhoeven, B. H., Snellen-Balendong, H. A., Hay, I. T., Boon, M. J., van der Linde, J. J., Blitz-Lindeque, R. J. I., et al. (2005). The versatility of progress testing assessed in an international context: A start for benchmarking global standardization? *Medical Teacher,* 27(6), 514–520.

Von Davier, A. (2011). *Statistical models for test equating, scaling, and linking*. New York, NY: Springer.

Wallach, P. M., Crespo, L. M., Holtzman, K. Z., Galbraith, R. M., & Swanson, D. B. (2006). Use of a Committee Review Process to Improve the Quality of Course Examinations. *Advances in Health Sciences Education: Theory and Practice,* 11, 61–68.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika,* 54, 427-450.

Webb, N. L. (2006). “Identifying content for student achievement tests”. In *Handbook of Test Development*. Downing, S. M., & Haladyna, T. M., eds. Mahwah, NJ: Lawrence Erlbaum Associates, 155–180.

Weggemans, M. M., Dijk, B., Dooijeweert, B., Veenendaal, A. G., & ten Cate, O. (2017). The postgraduate medical education pathway: an international comparison. *GMS Journal for Medical Education,* 34(5), Doc63.

Weiss, D. J. & Guyer, R. (2010). Manual for CATSim: Comprehensive simulation of computerized adaptive testing. St. Paul MN: Assessment Systems Corporation.

Weiss, J. (2011). Better Data From Better Measurements Using Computerized Adaptive Testing. *Journal of Methods and Measurement in the Social Sciences* 2(1), 1-27.

Wijnen-Meijer, M., Burdick, W., Alofs, L., Burgers, C., & ten Cate, O. (2013). Stages and transitions in medical education around the world: Clarifying structures and terminology. *Medical Teacher,* 35(4), 301-307.

Wise, S. L., & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicologica,* 21(1), 135-155.

Wrigley, W., van der Vleuten, C. P., Freeman, A., & Muijtjens, A. (2012). A systemic framework for the progress test: strengths, constraints and issues: AMEE Guide No. 71. *Medical Teacher,* 34(9), 683–697.