International Journal of Humanities Social Sciences and Education (IJHSSE)
Volume 2, Issue 8, August 2015, PP 23-33
ISSN 2349-0373 (Print) & ISSN 2349-0381 (Online)
www.arcjournals.org

Validity of Progress Testing in Healthcare Education

Anastasios Plessas

Peninsula Dental School
University of Plymouth, Plymouth, UK
anastasios.plessas@plymouth.ac.uk

Abstract: *Progress testing is a novel form of assessment introduced fairly recently in the field of medical and healthcare education. It is a comprehensive test which assesses knowledge across all the content areas of medicine reflecting the end objectives of the curriculum. It offers several advantages compared to traditional assessments. All assessments in the field of education need to be valid and fit for purpose. Test validity is the extent to which a test accurately measures what it purports to measure. The aim of this review is to discuss the validity evidence of this modern medical knowledge assessment. A comprehensive literature search was conducted using several medical and educational databases (Medline via EBSCO, British Educational Index, Dentistry and Oral Science Source, Education abstracts and ERIC). Forty four (44) relevant papers were retrieved from the above search. Content validity is assured by a carefully designed blueprint, high quality items written by content experts and reviewed for quality control. Feedback from and to the students, supports the response process of the validity of the test whilst comprehensive psychometric characteristics of the test provide even further internal structure-related evidence. The construct validity is supported by the increase of the mean scores of the test according to the year and the relationship to other relevant tests and licensing examinations. The validity of progress testing seems to be supported by several sources of evidence. However, ongoing training, recourses investment and high commitment is required on behalf of the involved institutions to sustain the high reliability and validity of the test.*

Keywords: *medical education, medical students, assessment, validity, reliability.*

1. Introduction

The introduction of problem based learning (PBL) as a new philosophy in medical and dental education brought the need for new methods to assess knowledge consistent with the PBL main principals of student-directed, deep and life-long learning. Therefore, medical schools which introduced a problem based curriculum adopted the progress testing as a method of testing factual medical knowledge. Progress test of applied medical knowledge was pioneered independently by both Maastricht and Missouri universities as early as the late 1970s (Vleuten, Verwijnen, & Wijnen, 1996). Recently progress test has gained popularity amongst many institutions and adopted by other schools, namely McMaster University and Peninsula College of Medicine and Dentistry. International collaborations have also been established and constantly flourishing (A. Freeman, Van Der Vleuten, Nouns, & Ricketts, 2010).

Progress test is a comprehensive test sampling knowledge across all content areas of medicine reflecting the end objectives of the curriculum. The test is periodically given to all medical students in the curriculum regardless of their year of training (Vleuten et al., 1996). The format does not allow the students to prepare themselves specifically for the test and develop pre-test revision strategies, therefore preventing temporary memorization of facts and surface test-driven learning (Blake et al., 1996; Vleuten et al., 1996). Thus, this longitudinal integrated assessment approach is plausible to have a positive effect on student learning behaviour by discouraging binge learning (L. W. T. Schuwirth & van der Vleuten, 2012).

It is not surprising why progress test has gained popularity amongst different institutions. It offers several advantages compared to traditional assessments. Given that assessment drives learning, testing at regular intervals over the course of an educational programme helps monitor the progress of students (Ali, Coombes, Kay, & al., 2015; Blake et al., 1996; L. W. T. Schuwirth & van der Vleuten, 2012). It also facilitates assessment of functional knowledge (Blake et al., 1996; L. W. T. Schuwirth & van der Vleuten, 2012; Vleuten et al., 1996). In addition, it offers opportunities for feedback; it can identify learners in need of remediation early in the programme which can subsequently lead to improvement of performance in successive years (Ali et al., 2015; Blake et al., 1996; L. W. T. Schuwirth & van der Vleuten, 2012; Vleuten et al., 1996). Nonetheless, progress test can be a successful assessment method regardless if the schools employ problem based learning or not. Verhoeven et al. found no systematic differences on total test scores between PBL and non-PBL students from two Dutch medical schools. It has also been suggested that progress test methodology provides a versatile instrument that can be used to assess medical schools across the world implying the feasibility of national and international progress testing collaborations (Verhoeven, Snellen-Balendong, Hay, & al., 2005). A recent guide published by the Association for Medical Education in Europe (AMEE) describes a systemic framework for the progress test encouraging the establishment of potential or new progress testing in medical education programmes (Wrigley, van der Vleuten, Freeman, & Muijtjens, 2012).

In this piece, a review of the progress test literature in the undergraduate medical and dental education through the lenses of the validity framework (Education., 1999) as rigorously described and interpreted in Downing's 2003 paper:'' (Downing, 2003) will be attempted. Test validity is the extent to which a test accurately measures what it purports to measure. Validity refers to the evidence presented to support or refute the meaning or interpretation assigned to assessment results (Downing, 2003). All assessments require validity evidence and nearly all topics in assessment involve validity in some way (Downing, 2003).

2. METHODS

A thorough literature search was conducted using several medical and educational databases (Medline via EBSCO, British Educational Index, Dentistry and Oral Science Source, Education abstracts and ERIC) and search terms such as medicine, medical education and progress test. A hand-search of the literature of included full articles was also performed. The search strategy and search terms used along with a flowchart depicting the selection process of the papers included are presented in Fig.1 and Fig.2 respectively..

3. RESULTS AND DISCUSSION

Table 1 summarizes the main characteristics of the different progress tests identified in the literature. In the following paragraphs the several stages of the development and validation of a progress test as described in the AMEE guide (Wrigley et al., 2012) will be discussed as an effort to identify different sources of evidence which may support or refute the hypothesis of construct validity of the progress test as an assessment method in undergraduate medicine and dentistry. It has to be noted that the assessment itself is never said to be 'valid' or 'invalid', rather assessment scores have more or less evidence to support the proposed interpretations (Downing, 2003), namely whether the assessment measures what it purports to measure (L. W. Schuwirth & van der Vleuten, 2011). Construct validity has multiple facets. It refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests (Downing, 2003).Evidence from five sources, as set by the APA validity framework will be sought, namely content, response process, internal structure, relationship to other variables and consequences (Downing, 2003).

3.1. Blueprinting

The blueprint is the basic and fundamental requirement on which the progress test relies for the valid and reliable construction of its content (Wrigley et al., 2012).The blueprint ensures that each test will be a representative and balanced sampling of the same content (Wrigley et al., 2012). The blueprint will be reflecting the end objectives of the curriculum, knowledge will be expected to have been acquired by a newly qualified doctor or dentist (Ali et al., 2015).The Peninsula Medical and Dental School progress test blueprint is informed by the GMC and GDC standards (Ali et al., 2015; A. C. Freeman & Ricketts, 2010) whilst the progress test in Maastricht and the Dutch collaborative progress test are informed by the Dutch National Blueprint for the Medical Curriculum (L. Schuwirth, Bosman, Henning, Rinkel, & Wenink, 2010). This latter blueprint has been shared or adapted by several international individual or collaborative progress tests in South Africa (Verhoeven et al., 2005),Indonesia (Findyartini, Werdhani, Iryani, & al., 2014), Mozambique (Aarts, Steidel, Manuel, & Driessen, 2010) and Germany (Nouns & Georg, 2010). Besides, the content of the National Board of Medical Examiners USMLE (United States Medical Licencing Examination) has informed the blueprint of a US progress test in Florida (Johnson, Khalil, Peppler, Davey, & Kibble, 2014), the collaborative US-UK multi-school progress testing project (MSPT) (Swanson, Holtzman, Butler, & al., 2010), and a progress test employed in the KSAU-HS medical school in Saudia Arabia (Al Alwan, Al-Moamary, Al-Attas, & al., 2011).

3.2. Test format, Item Writing and Quality Control

Several question formats have been used in different progress tests. As it can be observed in table 1 most of the progress tests employ multiple choice questions (single best answer). The Dutch collaborative progress test was originally comprised by 250 True-false questions but from 2005 onwards this has been changed to 200 single-best option multiple choice questions (L. Schuwirth et al., 2010). The single best answer option provides more reliable scores and a lower guessing probability enhancing the validity of the assessment (Wrigley et al., 2012). Utrecht University employs short answer question based upon 40 clinical cases (Rademakers, ten Cate, & Bär, 2005). Rademakers et al. demonstrated that with a lower number of students and questions, a short answer progress test is also reliable and feasible as high internal consistencies and reliability was achieved (Cronbach's alpha 0.85-0.87) (Rademakers et al., 2005).

In the Dutch consortium the number of options selected for each item varies between two and five, with most having three or four options while the progress test at Peninsula has consistently used 5-option items (A. C. Freeman & Ricketts, 2010; L. Schuwirth et al., 2010; Wrigley et al., 2012). It can be argued that trying to write questions with five possible answers may force the item writers to include less appropriate alternatives that are easily recognised as being incorrect will have a negative impact on the construct validity of the test as it will interfere with the test difficulty (Wrigley et al., 2012). However, research carried out at Peninsula medical School has demonstrated that using consistently a 5-option test provides a constant subtracted mark of –0.25 for an incorrect answer thereby removing the need to regularly alter the rubric in programmes that mark automatically (A. C. Freeman & Ricketts, 2010; Wrigley et al., 2012).

The quality of questions and the presence of independent experts and trained item writers are sources of content-related validity evidence (Downing, 2003). Content experts have been utilised in the majority of the tests namely the US SEPT (Spaced Education Progress Testing) (Kerfoot, Shaffer, McMahon, & al., 2011) and MSPT (multi-school progress testing) (Swanson et al., 2010). In the Dutch cross institutional collaboration, eight committee members with backgrounds in basic, clinical and behavioural science are involved in the item writing and quality control process (van der Vleuten et al., 2004). Similarly, in the Peninsula Medical School, all the test items must be accompanied by a literature reference, supporting even further the content related validity of the assessment (A. C. Freeman & Ricketts, 2010; van der Vleuten et al., 2004). Item quality is enhanced by item-writing training followed by extensive item review, both with regard to the accuracy of the item content as well as an assessment of the extent to which each item fits with the purpose of test (Albanese & Case, 2015).

Quality control procedures to review the performance of the test items and remove the poorly performing ones is a source of response process-related validity evidence (Downing, 2003; Wrigley et al., 2012). The Maastricht cross institutional collaboration (van der Vleuten et al., 2004) and Peninsula Dental School (Ali et al., 2015) employ a rigorous two quality control cycles before and after the test administration. A demographic analysis ensures that the test is fair and does not discriminate against particular groups (Ali et al., 2015) adding some consequential evidence of validity into the assessment (Downing, 2003). Significant source of construct validity is the involvement of the students in the decision making and quality control of the test items. The majority of the tests offer this feedback opportunity to students, namely Maastricht (van der Vleuten et al., 2004), McMaster University (Canada) (Blake et al., 1996), Germany (Nouns & Georg, 2010), Peninsula (Ali et al., 2015; Bennett, Freeman, Coombes, Kay, & Ricketts, 2010; A. Freeman et al., 2010; A. C. Freeman & Ricketts, 2010) and Mozambique (Aarts et al., 2010). Nonetheless, providing the students with enough detailed information about the test before the examination takes place and with detailed reports and explanations of their scores comprises another source of the consequence related construct validity evidence (Albanese & Case, 2015; Findyartini et al., 2014; Finucane, Flannery, Keane, & Norman, 2010; A. Freeman et al., 2010; Nouns & Georg, 2010; L. Schuwirth et al., 2010; Swanson et al., 2010; van der Vleuten et al., 2004)

Furthermore, investigating the difficulty and discrimination of the test items comprises internal structure related validity evidence (Albanese & Case, 2015; Downing, 2003; Wrigley et al., 2012). The mean item difficulties in the multi-school progress test (MSPT) ranged between 0.78 and 0.80 (Swanson et al., 2010). The San Paulo progress test found no difference in the mean degree of difficulty over the years of testing (Al Alwan et al., 2011), whilst in the Indonesian collaborative progress test (cPT) to assure the quality, items with difficulty index of 0.3-0.7 and discrimination index of >0.25 were incorporated (Findyartini et al., 2014). However, administering different test forms to examinees on repeated occasions represents the concern that the forms might differ in difficulty (Langer & Swanson, 2010). A psychometric statistical process has been suggested, called equating, which addresses this issues and can control the differences in difficulty between forms so that scores can be used interchangeably (Langer & Swanson, 2010).

3.3. Test Administration

From the table 1 becomes clear that there is also a considerable variation in the number of items and the frequency of the test administration. Progress test designers use different testing frequencies (typically two, three or four tests per year) and different test sizes (number of items varying between 100 and 250) (C. Ricketts, Freeman, Pagliuca, Coombes, & Archer, 2010). Short tests may under-represent the content of the blueprint challenging the content validity of the test (Downing, 2002).Similarly reducing the number of tests in a year may decrease the total sampling opportunities and hence validity (C. Ricketts et al., 2010). Ricketts et al. used the generalizability theory and reported the standard errors of measurement (SEM) as a trade-off between number of items per test and number of tests per year. The lower the SEM, the more reliable the results of the testing are. Namely, the SEM for a progress test of 200 items delivered twice a year was 3.02 (Germany), for a progress test of 200 items four times a year 2.45 (Netherlands group), and for 125 items 4 times a year 3.00 (Peninsula). (C. Ricketts et al., 2010)

The synchronicity of the test and its feasibility is a valuable source of response process-related validity (Wrigley et al., 2012). When synchronicity is not feasible due to space limitation for example, a software like a ‘’secure browser’’ ’may ensure the validity of the test. Such software is used by the US/UK collaborative multi-school progress test where the students take the web-based test in waves. The software locks down the workstation so that the students cannot copy the test materials, consult online references or send e-mails to others (Swanson et al., 2010).

Reliability is an important aspect of an assessment’s validity evidence. Reliability refers to the reproducibility of the scores on the assessment (Downing, 2003). The most commonly reported estimate for reliability in progress testing is either the alpha internal consistency estimates (such as Cronbach’s alpha) or test retest reliability across repeated administrations (Albanese & Case, 2015). The test–retest reliability of the McMaster exam ranged between .53 and .64 (Blake et al., 1996) whilst he internal consistency alpha values of the Maastricht exam have ranged between .70 and .80 (Van Der Vleuten, 1996). In the dental progress test the average reliability (Cronbach’s a) over 42 test/cohort combination was 0.753 (Ali et al., 2015). The higher the Cronbach alpha, the stronger is the evidence for the validity of the test. Table 2 summaries the different progress testing reliability scores reported in the literature. As it can be observed the majority of the tests meet the 0.8 reliability coefficient required for high stakes assessment (A. Freeman et al., 2010).

3.4. Scoring, Standard Setting and Benchmarking

The use of the ‘‘don’t know’’ option is included in the progress tests of the Dutch, German and Canadian collaborations and at Peninsula to reduce the frequency of guessing as well as to reduce the influence of guessing on the score (Wrigley et al., 2012). Key scoring issues surrounding progress testing are whether or not to correct for guessing (formula scoring), and how to handle the ‘‘don’t know’’ option if it is used (Albanese M, 2008). Accuracy of final scores is a source of response process-related validity evidence (Downing, 2003). There is some evidence suggesting that when the test is taken under formula scoring the number of correct reliabilities is higher (A. M. Muijtjens, Mameren, Hoogenboom, Evers, & van der Vleuten, 1999). Furthermore, the results of a recent study by Wade et al. provide empirical support for the use of methods to control for guessing enhancing the construct validity of the tests which employ these methods (Wade et al., 2012). On the other hand, the ‘‘don’t know’’ option has been criticised as it may introduce measurement error by discriminating against those students with risk-taking behaviour (Albanese M, 2008; Wrigley et al., 2012). However, mathematical analysis has suggested that this effect is small compared with the measurement error of guessing (Espinosa & Gardeazabal, 2010).

Applying pass-fails decisions accurately is an important source of construct validity (response process and consequences related); therefore, the higher the stakes of the test the stronger the requirement becomes for the standard setting. An Angoff process is considered to be a defensible method of standard setting. Verhoven et al. has investigated the reliability and credibility for an Angoff standard setting procedure involving graduated students. He concluded that this may be an appropriate, acceptable and feasible method for standard setting (Verhoeven et al., 1999; Verhoeven, Verwijnen, Muijtjens, Scherpbier, & van der Vleuten, 2002). Similarly, Rickets et al. have shown the usefulness of triangulating standard-setting data across a number of internal sources involving student test results and an external source of data from newly qualified doctors (C. Ricketts, Freeman, & Coombes, 2009). The Dutch, German and Peninsula progress tests however prefer norm referenced versus criterion referenced methods. This is justified by reliable evidence supporting that because of the variation in progress test difficulty, using an absolute cut-off score is more precarious than norm-referenced scores (A. M. M. Muijtjens, Hoogenboom, Verwijnen, & van der Vleuten, 1998). Namely, with norm referencing the failure rate was found reasonably constant, whilst with absolute referencing the failure rate varied from 2 to 47% for different tests (A. M. M. Muijtjens et al., 1998). Finally when it comes to cross-institutional benchmarking where collaborations have developed, if longitudinal data are available the use of cumulative deviation has been shown to be the most appropriate method as it suppresses the noise of systematic differences (A. M. M. Muijtjens, Schuwirth, Cohen-Schotanus, Thoben, & van der Vleuten, 2008; Schauber & Nouns, 2010).

3.5. Educational Impact for Students’ Learning

As progress test is testing the growth of knowledge within the curriculum, evidence which demonstrate this growth comprises strong sources of construct validity evidence. This steady growth of knowledge is reflected through an increase of the mean test scores across the years and the presence of significant differences of the mean scores between years (Ali et al., 2015; Findyartini et al., 2014; A. C. Freeman & Ricketts, 2010; Tomic, Martins, Lotufo, & Bensenor, 2005; van der Vleuten et al., 2004). For example, the mean scores on the CBSE (Comprehensive Basic Science Examination, US) increased in a relatively linear fashion across the test administrations with scores at each time point being significantly different from the preceding time point (Johnson et al., 2014). Additionally, evidence of knowledge growth can be also reflected through the concomitant reduction of the ‘‘don’t know’’ responses as it was shown to be the case in the dental progress test (Ali et al., 2015). De Champlain et al. quantified the knowledge gain between the first and the final administration of the test to be of the magnitude of 2 standard deviations (+2SD) (De Champlain, Cuddy, Scoles, & al., 2010). Adding up to the validity evidence, a study by Kerfoot et al. investigating the educational effect of ‘‘Spaced Education Progress Testing (SEPT)’’, suggested that cycled reviews generated a 170% increase in learning retention relative to baseline (Kerfoot et al., 2011).

The evidence of construct validity can be also supported by the correlation of the progress test scores to the students’ cumulative GPA (Al Alwan et al., 2011; Findyartini et al., 2014). Namely, Alwan et al. Found that the correlations are higher in senior students compared to junior students and ranged from 0.38 to 0.77 (Al Alwan et al., 2011). Besides, in another study by Boshuizen et al. the Progress Test and the Clinical Reasoning Test revealed the same pattern of increasing scores over the years, and had a high inter-correlation (Boshuizen, van der Vleuten, Schmidt, & Machiels-Bongaerts, 1997). Progress-test performance has also been correlated significantly with the US Medical College Admission Test (MCAT) (Kerfoot et al., 2011). Significant correlations have been observed between progress tests and different licensing examinations such as the German National Licensing Exams

(Nouns & Georg, 2010), the licensing examination of the Medical Council of Canada(Blake et al., 1996) and the US Medical licencing Examination (USMLE) Step 1 (Johnson et al., 2014; Kerfoot et al., 2011) and Step 2 (Kerfoot et al., 2011). Nonetheless, progress testing can identify poor performing students as it has been shown by Kerfoot at al. in whose study progress-test correctly identified those second-year students who scored below the mean on Step 1 with 75% sensitivity, 77% specificity, and 41% positive predictive value (Kerfoot et al., 2011).

Finally, securing fairness in the progress testing supports even further the validity of the test. Studies have shown no significant differences between performance of males and females students (Findyartini et al., 2014; Tomic et al., 2005). Rickets et al. concluded that properly-designed progress test of items do not systematically discriminate against medical students with specific learning disabilities (Chris Ricketts, Brice, & Coombes, 2010).

3.6. Informing the Curriculum

Progress test scores are an important source of information for item authors, teachers in the programme, faculty and the overview committee (Wrigley et al., 2012). Progress test can be used as a diagnostic tool for the curriculum (Al Alwan et al., 2011; Findyartini et al., 2014) and show how the different patterns of learning exist in different curriculum areas (Coombes, Ricketts, Freeman, & Stratford, 2010). In areas where all the students perform poorly, the curriculum can be revised (Aarts et al., 2010) and subsequently the effect of the changes can be monitored through future progress testing scores (Coombes et al., 2010). All the above comprise valuable sources of consequence-related validity evidence (Downing, 2003). In the case of institutional collaborations the progress testing can find grounds for programme and curriculum comparisons (A. M. M. Muijtjens, Schuwirth, Cohen-Schotanus, & van der Vleuten, 2007; L. Schuwirth et al., 2010). However, these inter-curriculum comparisons may be of threat to validity if certain issues are not taken into consideration. Muijtjens, et al. observed that in a cross-institutional collaborative progress test the students obtained better results on items produced at their own schools (A. M. M. Muijtjens et al., 2007). Thus, progress test items were subject to origin bias. To address such issues, all the participating schools should contribute equal numbers of test items (A. M. M. Muijtjens et al., 2007). Similarly, sharing test materials has been shown to be viable but efforts should be made to eliminate the introduction of any translation and cultural bias which may compromise the validity and fairness of the test (Verhoeven et al., 2005).

4. Figures and Tables

Search Strategy		
S1	Medicine	3,266,932
S2	Medical	3,355,188
S3	Dentistry	234,317
S4	Dental	454,518
S5	S1 OR S2 OR S3 OR S4	6,207,230
S6	‘‘Progress Test’’	1,816
S7	‘‘Progress Testing’’	1,613
S8	S6 OR S7	3,301
S9	Validity	176,219
S10	Reliability	144,772
S11	Psychometric*	83,044
S12	Construct	129,503
S13	Content	662,236
S14	‘‘Educational Impact’’	5,771
S15	Utility	143,713
S16	Feasibility	135,090
S17	Blueprint*	6,151
S18	‘‘Standard Setting’’	7,576
S19	Benchmarking	16,318
S20	S9 OR S10 OR S11 OR S12 OR S13 OR S14 OR S15 OR S16 OR S17 OR S18 OR S19	1,342,892
S21	S5 AND S8 AND S20	94

Fig1. *Search strategy.*

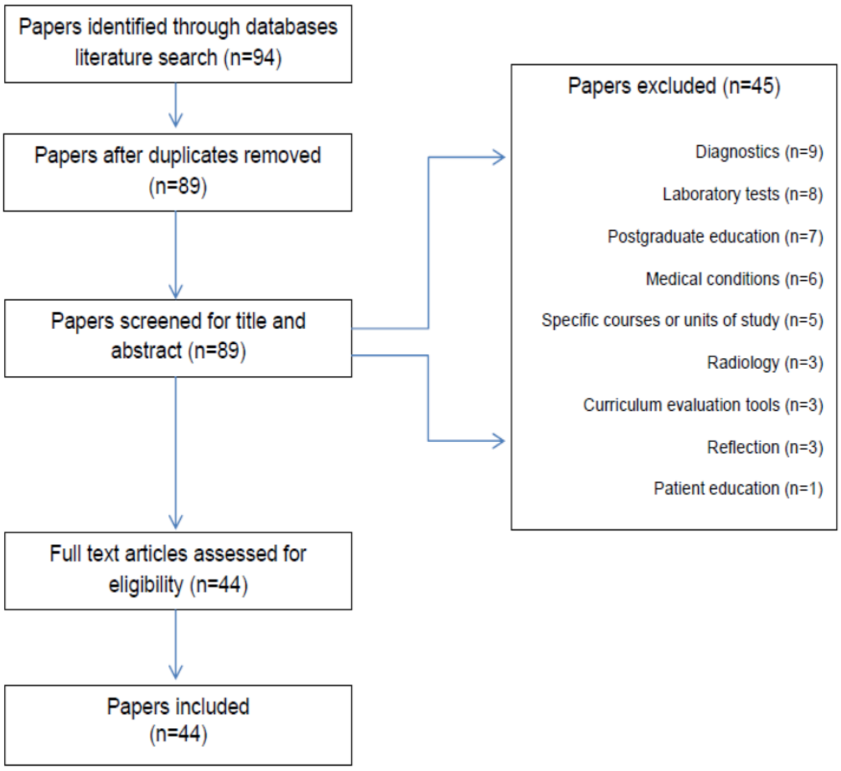
Papers identified through databases literature search (n=94)
Papers after duplicates removed (n=89)
Papers screened for title and abstract (n=89)
Full text articles assessed for eligibility (n=44)
Papers included (n=44)
Papers excluded (n=45)
Diagnostics (n=9)
Laboratory tests (n=8)
Postgraduate education (n=7)
Medical conditions (n=6)
Specific courses or units of study (n=5)
Radiology (n=3)
Curriculum evaluation tools (n=3)
Reflection (n=3)
Patient education (n=1)

Fig2. *Flowchart. Selection process of studies/papers*

Table1. *Main characteristics of progress tests internationally.*

School/ Country	Year introduced	Purpose	Delivery Method	Delivery Format	Frequency N/year	Blueprint	N items	Time of exam	Correction for guessing (Y/N)	Score aggregation (Y/N)
Netherlands Group (collaboration)	1999	Summative	Paper-based	MCQs	4	Whole of medical knowledge	200	4hours	Y	Y
McMaster (Canada)	1992	Summative & Formative	online	MCQs	3	Whole of medical knowledge	180	3 hours	Y	N
Germany	1999	Formative	Paper-based	MCQs	2	Whole of medical knowledge	200	3 hours	Y	N
NBME UK	2008	Summative & Formative	Web-based	MCQs	2	Whole of medical knowledge	120	3 hours	N	N
NBME US	2006	Summative & Formative	online	MCQs	4	Whole of medical knowledge	230	5 hours	N	N
Manchester	1997	Summative	Paper-based	MCQs	2	Whole of medical knowledge	125	2.5 hours	N	N
PU PMSD Medical	2002	Summative	Paper-based/ online	MCQs	4	Whole of medical knowledge GMC	125	3 hours	Y	Y
PU PMDS Dental	2007	Summative (Y3,4,5) Formative (Y1&2)		MCQs	4	Dental knowledge for a newly qualified dentist GDC	100	3 hours	Y	Y
Finland		Formative	Paper-based	T/F	3	Whole of medical knowledge (discipline based structure)	224	3 hours	Y	N
Sao Paulo (Brazil)	2001	Formative	Paper-based	MCQs	2	Whole of medical knowledge	100	-	N	N
Indonesia	2008	Formative	Paper-based	MCQs	2	Whole of medical knowledge	120	-	N	N
Mozambique	2001	Summative and Formative	Paper-based	MCQs	4	Whole of medical knowledge	200	4 hours	Y	
Saudi Arabia	2007	Summative and Formative	Paper-based	MCQs	2	Whole of medical knowledge	180	-	N	N

Table2. *Average reliability of progress tests internationally.*

Author (Year)	Institution	Average Reliability (Cronbach's alpha)
Aarts et al. (2010)	University of Mozambique	**0.83**
Ali et al. (2015)	Peninsula Dental School (UK)	**0.753**
Boshuizen et al. (1997)	University of Limburg (the Netherlands)	**0.9**
Findyartini et al. (2015)	Universitas Indonesia (FM UI), , Universitas Andalas (FM UNAND), Universitas Sebelas Maret (FM UNS), Indonesia	**0.90** **0.85** **0.86**
Johnson et al. (2014)	University of Central Florida Florida	**0.73**
Kerfoot et al. (2011)	Multi-Institutional US Study	**0.87**
Nouns & Georg (2010)	13 medical Schools in Germany and Austria	**0.96**
Rademakers et.al (2005)	University of Utrecht (The Netherlands)	**0.85**
Swanson et al. (2010)	MSPT (NBME US, St George's University, Leeds School of Medicine, Barts and The London School of Medicine, Queen's University (UK)	**0.82**

5. Conclusion

In summary, the validity of progress testing seems to be supported by several sources of evidence. Content validity is assured by a carefully designed blueprint, high quality items written by content experts and reviewed for quality control. Feedback from and to students supports the response process of the validity of the test whilst comprehensive psychometric characteristics of the test provide even further internal structure-related evidence. The construct validity is also supported by the increase of the mean scores of the test according to the year and the relationship to other relevant tests and licencing examinations. Lastly, consequence related evidence was also found in the literature to significantly support the validity of the test. Some threats to validity exist when cross-institutional collaborations and comparisons are attempted. The AMEE generic systemic framework provides an analysis of the basic requirements to minimise those threats and emphasises the need for quality control procedures, ongoing training and evaluation, effort and recourses investment and commitment to sustain high reliability and validity of the test (Wrigley et al., 2012). To conclude, it is well know that assessment drives learning (Wass, Van der Vleuten, Shatzer, & Jones, 2001). The cycle of testing, giving feedback, students using that feedback to direct learning and then retesting is inherent in progress testing (C. Ricketts et al., 2010) and therefore progress testing can facilitate learning more efficiently than frequent revision could ever do (Wood, 2009).

References

[1] Aarts, R., Steidel, K., Manuel, B. A. F., & Driessen, E. W. (2010). Progress testing in resource-poor countries: a case from Mozambique. Medical Teacher, 32(6), 461-463.

[2] Al Alwan, I., Al-Moamary, M., Al-Attas, N., & al., e. (2011). The progress test as a diagnostic tool for a new PBL curriculum. Educ Health (Abingdon), 24(3), 493.

[3] al., H. V. G. M. V. A. J. J. A. S. R. S. G. e. (1998).

[4] Albanese M. (2008). Benchmarking progress tests for cross-institutional comparisons: every road makes a difference and all of them have bumps (Vol. 42, pp. 4-7). United Kingdom.

[5] Albanese, M., & Case, S. M. (2015). Progress testing: critical analysis and suggested practices. Adv Health Sci Educ Theory Pract. doi: 10.1007/s10459-015-9587-z

[6] Ali, K., Coombes, L., Kay, E., & al., e. (2015). Progress testing in undergraduate dental education: the Peninsula experience and future opportunities. European Journal of Dental Education. doi: 10.1111/eje.12149

[7] Bennett, J., Freeman, A., Coombes, L., Kay, L., & Ricketts, C. (2010). Adaptation of medical progress testing to a dental setting. Medical Teacher, 32(6), 500-502. doi: 10.3109/0142159x.2010.486057

[8] Blake, J. M., Norman, G. R., Keane, D. R., Mueller, C. B., 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.

[9] Boshuizen, H. P., van der Vleuten, C. P., Schmidt, H. G., & Machiels-Bongaerts, M. (1997). Measuring knowledge and clinical reasoning skills in a problem-based curriculum. Medical Education, 31(2), 115-121.

[10] Coombes, L., Ricketts, C., Freeman, A., & Stratford, J. (2010). Beyond assessment: feedback for individuals and institutions based on the progress test. Medical Teacher, 32(6), 486-490. doi: 10.3109/0142159x.2010.485652

[11] De Champlain, A. F., Cuddy, M. M., Scoles, P. V., & al., e. (2010). Progress testing in clinical science education: results of a pilot project between the National Board of Medical Examiners and a US Medical School. Medical Teacher, 32(6), 503-508. doi: 10.3109/01421590903514655

[12] Downing, S. M. (2002). Threats to the validity of locally developed multiple-choice tests in medical education: construct-irrelevant variance and construct underrepresentation. Adv Health Sci Educ Theory Pract, 7(3), 235-241.

[13] Downing, S. M. (2003). Validity: on meaningful interpretation of assessment data. Medical Education, 37(9), 830-837.

[14] Education., A. E. R. A. A. P. A. N. C. o. M. i. (1999). Validity. Stand. Educ. Psychol. Test. Washington DC: AERA; 1999. p. 9–17.

[15] Espinosa, M. P., & Gardeazabal, J. (2010). Optimal correction for guessing in multiple-choice tests. Journal of Math Psych, 54(5), 415-425. doi: http://dx.doi.org/10.1016/j.jmp.2010.06.001

[16] Findyartini, A., Werdhani, R. A., Iryani, D., & al., e. (2014). Collaborative progress test (cPT) in three medical schools in Indonesia: The validity, reliability and its use as a curriculum evaluation tool. Medical Teacher, 1-8.

[17] Finucane, P., Flannery, D., Keane, D., & Norman, G. (2010). Cross-institutional progress testing: feasibility and value to a new medical school. Medical Education, 44(2), 184-186. doi: 10.1111/j.1365-2923.2009.03567.x

[18] Freeman, A., Van Der Vleuten, C., Nouns, Z., & Ricketts, C. (2010). Progress testing internationally. Medical Teacher, 32(6), 451-455. doi: 10.3109/0142159x.2010.485231

[19] Freeman, A. C., & Ricketts, C. (2010). Choosing and designing knowledge assessments: experience at a new medical school. Medical Teacher, 32(7), 578-581. doi: 10.3109/01421591003614858

[20] Johnson, T. R., Khalil, M. K., Peppler, R. D., Davey, D. D., & Kibble, J. D. (2014). Use of the NBME Comprehensive Basic Science Examination as a progress test in the preclerkship curriculum of a new medical school. Adv in Phys Educ, 38(4), 315-320. doi: 10.1152/advan.00047.2014

[21] Kerfoot, B. P., Shaffer, K., McMahon, G. T., & al., e. (2011). Online "spaced education progress-testing" of students to confront two upcoming challenges to medical schools. Academic Medicine, 86(3), 300-306. doi: 10.1097/ACM.0b013e3182087bef

[22] Langer, M. M., & Swanson, D. B. (2010). Practical considerations in equating progress tests. Medical Teacher, 32(6), 509-512. doi: 10.3109/0142159x.2010.485654

[23] Muijtjens, A. M., Mameren, H. V., Hoogenboom, R. J., Evers, J. L., & van der Vleuten, C. P. (1999). The effect of a 'don't know' option on test scores: number-right and formula scoring compared. Medical Education, 33(4), 267-275.

[24] Muijtjens, A. M. M., Hoogenboom, R. J. I., Verwijnen, G. M., & van der Vleuten, C. P. M. (1998). Relative or Absolute Standards in Assessing Medical Knowledge Using Progress Tests. Adv Health Sci Educ, 3(2), 81-87. doi: 10.1023/A:1009728423412

[25] Muijtjens, A. M. M., Schuwirth, L. W. T., Cohen-Schotanus, J., Thoben, A. J. N., & van der Vleuten, C. P. M. (2008). Benchmarking by cross-institutional comparison of student achievement in a progress test. Medical Education, 42(1), 82-88.

[26] Muijtjens, A. M. M., Schuwirth, L. W. T., Cohen-Schotanus, J., & van der Vleuten, C. P. M. (2007). Origin bias of test items compromises the validity and fairness of curriculum comparisons. Medical Education, 41(12), 1217-1223.

[27] Nouns, Z. M., & Georg, W. (2010). Progress testing in German speaking countries. Medical Teacher, 32(6), 467-470. doi: 10.3109/0142159x.2010.485656

[28] Rademakers, J., ten Cate, T. J., & Bär, P. (2005). Progress testing with short answer questions. Medical Teacher, 27(7), 578-582.

[29] Ricketts, C., Brice, J., & Coombes, L. (2010). Are Multiple Choice Tests Fair to Medical Students with Specific Learning Disabilities? Adv Health Sci Educ, 15(2), 265-275.

[30] Ricketts, C., Freeman, A., Pagliuca, G., Coombes, L., & Archer, J. (2010). Difficult decisions for progress testing: how much and how often? Medical Teacher, 32(6), 513-515. doi: 10.3109/0142159X.2010.485651

[31] Ricketts, C., Freeman, A. C., & Coombes, L. R. (2009). Standard setting for progress tests: combining external and internal standards. Medical Education, 43(6), 589-593. doi: 10.1111/j.1365-2923.2009.03372.x

[32] Schauber, S., & Nouns, Z. M. (2010). Using the cumulative deviation method for cross-institutional benchmarking in the Berlin progress test. Medical Teacher, 32(6), 471-475. doi: 10.3109/0142159x.2010.485653

[33] Schuwirth, L., Bosman, G., Henning, R. H., Rinkel, R., & Wenink, A. C. (2010). Collaboration on progress testing in medical schools in the Netherlands. Medical Teacher, 32(6), 476-479. doi: 10.3109/0142159x.2010.485658

[34] Schuwirth, L. W., & van der Vleuten, C. P. (2011). General overview of the theories used in assessment: AMEE Guide No. 57. Medical Teacher, 33(10), 783-797. doi: 10.3109/0142159x.2011.611022

[35] Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2012). The use of progress testing. Persp Med Educ, 1(1), 24-30. doi: 10.1007/s40037-012-0007-2

[36] Swanson, D. B., Holtzman, K. Z., Butler, A., & al., e. (2010). Collaboration across the pond: the multi-school progress testing project. Medical Teacher, 32(6), 480-485. doi: 10.3109/0142159x.2010.485655

[37] Tomic, E. R., Martins, M. A., Lotufo, P. A., & Bensenor, I. M. (2005). Progress testing: evaluation of four years of application in the school of medicine, University of Sao Paulo. Clinics (Sao Paulo), 60(5), 389-396. doi: /S1807-59322005000500007

[38] Van Der Vleuten, C. P. (1996). The assessment of professional competence: Developments, research and practical implications. Adv Health Sci Educ Theory Pract, 1(1), 41-67. doi: 10.1007/bf00596229

[39] Van der Vleuten, C. P., Schuwirth, L. W., Muijtjens, A. M., Thoben, A. J., Cohen-Schotanus, J., & van Boven, C. P. (2004). Cross institutional collaboration in assessment: a case on progress testing. Medical Teacher, 26(8), 719-725. doi: 10.1080/01421590400016464

[40] Verhoeven, B. H., Snellen-Balendong, H. A., Hay, I. T., & al., e. (2005). The versatility of progress testing assessed in an international context: a start for benchmarking global standardization? Medical Teacher, 27(6), 514-520. doi: 10.1080/01421590500136238

[41] Verhoeven, B. H., Van der Steeg, A. F. W., Scherpbier, A. J. J., Muijtjens, A. M. M., Verwijnen, G. M., & van der Vleuten, C. P. M. (1999). Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges. Medical Education, 33(11), 832-837.

[42] Verhoeven, B. H., Verwijnen, G. M., Muijtjens, A. M. M., Scherpbier, A. J. J. A., & van der Vleuten, C. P. M. (2002). Panel expertise for an Angoff standard setting procedure in progress testing: item writers compared to recently graduated students. Medical Education, 36(9), 860-867.

[43] Vleuten, C. P. M. V. D., 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), 103-109. doi: 10.3109/01421599609034142

[44] Wade, L., Harrison, C., Hollands, J., Mattick, K., Ricketts, C., & Wass, V. (2012). Student perceptions of the progress test in two settings and the implications for test deployment. Adv Health Sci Educ Theory Pract, 17(4), 573-583. doi: 10.1007/s10459-011-9334-z

[45] Wass, V., Van der Vleuten, C., Shatzer, J., & Jones, R. (2001). Assessment of clinical competence. Lancet, 357(9260), 945-949. doi: 10.1016/s0140-6736(00)04221-5

[46] Wood, T. (2009). Assessment not only drives learning, it may also help learning. Medical Education, 43(1), 5-6. doi: 10.1111/j.1365-2923.2008.03237.x

[47] Wrigley, W., van der Vleuten, C. P. M., 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. doi: 10.3109/0142159X.2012.704437

[48] Polard D., Hussy S. and Verma K. G., Biochemical aspects of air pollution induced injury symptoms of some common ornamental road side plants, Int. J. Res. BioSciences. 4(2), 50 (2000).

AUTHOR'S BIOGRAPHY

Anastasios Plessas is an NIHR Academic Clinical Fellow in Peninsula Dental School. His current role involves teaching and research within the Dental School. He was awarded a distinction in his Master's Degree in Primary Dental Care from the University of Glasgow in 2014, and he recently completed a Postgraduate Certificate in Clinical Education at the University of Plymouth.