Reliability and validity of methods to assess undergraduate healthcare student performance in pharmacology: comparison of open book versus time-limited closed book examinations.

Practitioner Research
In Higher Education

Issue 15(1) pages 14-23

David Bell, Vikki O'Neill and Vivienne Crawford, Queen's University Belfast, Northern Ireland

Abstract
We compared the influence of open-book extended duration versus closed book time-limited format on reliability and validity of written assessments of pharmacology learning outcomes within our medical and dental courses. Our dental cohort undertake a mid-year test (30xfree-response short answer to a question, SAQ) and end-of-year paper (4xSAQ, 1xessay, 1xcase) in pharmacology. For our first year medical cohort, pharmacology is integrated within a larger course, contributing 20xclinical vignette questions (to select the single best answer (SBA) to each question from a choice of 5 plausible answers) to a mid-year test and 3-5xSAQ to an end-of-year paper. Our experience indicates that SAQ are as reliable as SBA for closed-book time-limited assessments; reliability correlates with number of questions employed. We have found good correlation between mid-year and end-of-year performance (predictive validity), between questions (factorial validity) and between pharmacology and other subjects within the assessment (concurrent validity). Adoption of open-book extended duration assessments resulted in only modest reduction in reliability and validity.

Key words
Validity; reliability; single best answer; short notes questions; Covid-19 pandemic.

Introduction
A range of strategies exist to assess the academic performance of healthcare students enrolled on professional courses using traditional closed-book, time-limited written examinations including: single best answer (SBA) or only correct answer selected by the candidate from a choice of multiple answers (MCQ) to a question, (very) short answer free written response to a question (SAQ), longer essay-style questions and structured problems and clinical case studies (Fallatah et al., 2015; Hift, 2014; Sam et al., 2016; Wilkinson and Shaw, 2015). There is a lack of consensus whether candidates tend to score more highly in SBA / MCQ assessments than in free written response assessments (Preston et al, 2020; Sam et al, 2016; Sullivan, 2011; Wilkinson and Shaw, 2015); the former have been criticised for encouraging cueing and superficial learning (Holzinger et al., 2020), reflecting student perception that SBA/MCQ assessments are easier and require less effort to be invested in learning (Holzinger et al., 2020; Jaenicke et al., 2020; Preston et al., 2020).

Reliability is defined as the extent to which an assessment method or instrument measures consistently the performance of the candidate (Andreatta and Gruppen, 2009; Downing, 2003; Fallatah et al.; 2015; Sullivan, 2011). Reliability is measured by analysing the correlation between the answers to multiple questions. For this purpose, Cronbach's alpha can be applied to short free-response answers, whereas Kuder-Richardson 20 [KR20] represents a special case of Cronbach's alpha as applied to dichotomous answers and is more suitable for measuring reliability of SBA assessments. Values tending to 1 indicate high reliability.

Validity defines how well the assessment tool employed actually measures the underlying outcome of

Citation
Bell, D., O'Neill, V. and Crawford. V. (2023) 'Reliability and validity of methods to assess undergraduate healthcare student performance in pharmacology: comparison of open book versus time-limited closed book examinations', *PRHE Journal,* 15(1), pp. 14-23.

interest. There are various facets to validity (Andreatta and Gruppen, 2009; Downing, 2003; Fallatah et al.; 2015; Patil et al, 2015; Sam et al 2016; Sullivan, 2011). Content validity establishes that assessment strategies sample the breadth of the curriculum, often employing sampling grids. Construct validity ensures that assessment strategies are mapped to professional learning outcomes at an appropriate level (for example, blueprinting to General Medical or General Dental Curriculum outcomes for graduates- GDC 2015; GMC, 2018). Predictive (Criterion) validity determines the extent to which an assessment outcome predicts performance in another future assessment. Concurrent validity addresses the extent to which performance in the assessment correlates with performance in another assessment by the same cohort of candidates. Finally, factorial validity measures the extent of correlation of different discrete factors within the whole assessment.

UK medical schools are increasingly drawing on SBA/MCQ style-assessments. Experience indicates that the assessment strategy adopted will likely impact how students approach their learning (Preston et al, 2020; Witt et al., 2022). Although students may prefer SBA/MCQ format and find such assessments easier these might not necessarily test deeper understanding and facilitate acquisition of long-term knowledge to the same degree as other strategies (Holzinger et al., 2020; Witt et al., 2022). There is a lack of consensus in the educational literature however as to whether SBA/MCQ assessments are more reliable and convenient than free-response SAQ, but at the cost of validity (Hift, 2014; Holzinger et al., 2020; Patil et al., 2015). The first objective therefore was to analyse the reliability and validity of our approaches to assess achievement of learning outcomes for basic and clinical pharmacology content within the undergraduate medical and dental curricula.

It has been argued that open-book extended duration assessments encourage deeper engagement and application of knowledge relative to simple recall of memorised factual information in closed book time-limited assessments. Open-book assessments may also generate less student anxiety but are more susceptible to cheating (Spiegel and Nivette, 2023). The second objective therefore was to compare the impact of open-book, extended duration format versus traditional time-limited closed-book format on the reliability and validity of our assessment strategies. The COVID-19 pandemic necessitated that many assessments adapted to open-book, extended duration format, providing further incentive and opportunity to address this objective.

Methods

Organisational context and study cohorts

Medicine at QUB

At The Queen's University of Belfast has an intake of >260 students annually [261+5.0, mean + sd, n=8 years]. Prior to 2017, an introduction to pharmacology and therapeutics was incorporated together with pathology into a single semester module in spring of Year 1 of 5 (20 CATS): Principles of Disease and Treatment. Assessment comprised an end of semester written examination (weighted 100%, 2h, short answers x10, of which 5 were pharmacology and 5 were pathology). Since 2017, introduction to pharmacology and therapeutics has been part of larger full-year first year module (40 CATS). Assessment comprises 2x 40 SBA mid-year class tests each weighted 15% (2h, 20 SBA pharmacology and therapeutics and 20 SBA pathology; 40 SBA genetics and biochemistry) together with an end of year examination weighted 70% (2h, short answers x10, of which pharmacology x3, pathology x3, genetics and biochemistry x4).

Dentistry at QUB

There is an intake of 55-60 students annually [58+5.6, mean + sd, n=7 years]. Pharmacology for Dentistry is a full year module delivered in Year 2 of 5. Assessment of this comprises a mid-year class test (1h, open-ended very short answer questions x30, weighted 10%), end of year examination (2h,

short answers x4 , essay x1, case study x1, weighted 80%) and a group presentation on pharmacological management of a dental condition (15min, weighted 10%) .
Our assessment strategies map to respective professional learning outcomes (General Dental Council, 2015; General Medical Council, 2018), ensuring construct validity. Free response answers are marked by one internal examiner but 25% of scripts are double-marked by a second internal examiner prior to review by the external examiner: inter-marker variability is <5%. Detailed model answers and marking schemes indicating allocation of marks are provided for all questions for use by the examiners to facilitate standard setting and reduce potential for inter-marker variability. A representative selection of these drawn from past papers is also provided to students for the purposes of self-assessment and formative feedback in advance of the summative examination.

Example questions

Dental class test: List TWO advantages of adding a vasoconstrictor to a local anaesthetic preparation
Dental short notes question: Write short notes on mode of action, therapeutic use and adverse effects of (1) ibuprofen; (2) nystatin
Dental essay question: Discuss the pharmacological management of diabetes mellitus and its implications for dental practice

Dental case study: in the case below a series of questions is based around the pharmacology of the drugs taken, management of the dental condition and any acute medical emergencies arising during dental practice. Answer <u>all parts</u> of the following case history:
Mr Brown is a 68 year old life-long smoker. His medical history includes chronic obstructive pulmonary disease. He had been using a salbutamol inhaler as required for relief of his breathlessness, particularly on exertion.
(a) Describe the mechanism of action by which salbutamol provides symptomatic relief.
This has helped, but more recently his symptoms had been getting worse and more frequent and his doctor had commenced him on a regular (preventer) Seretide inhaler containing salmeterol plus fluticasone.
(b) Explain how the properties of salmeterol differ from salbutamol.
(c) Name the class of drug that fluticasone belongs to and describe the mechanism of action.

(d) Name a side-effect affecting the oral cavity that is associated with inhaled fluticasone.

(e) Describe two practical measures that Mr Brown could take when using his Seretide inhaler to minimise the risk of the side-effect identified in (d) from occurring.

(f) If despite the measures recommended in (e) the side-effect identified in (d) did occur, what drug could be prescribed to treat the condition?

(g) Describe the mechanism of action of drug identified in (f)

A routine dental examination reveals halitosis and periodontitis, with red, swollen and recessed gums, significant calculus and some loosening of several teeth. During the root surface instrumentation to clean below the gum-line, Mr Black becomes breathless.

(h) How would you manage the acute breathless episode?

As an adjunct to mechanical debridement to remove calculus Mr Brown's dentist decided to prescribe an antibiotic therapy.

(i) Name an antibiotic that is indicated for the treatment of periodontitis and justify your choice.

(j) Describe the mechanism of action of the drug identified in (i)

(k) List any TWO steps Mr Brown could take to improve his oral hygiene and prevent progression of his periodontitis?

Despite the use of the regular preventer inhaler, Mr Brown is still experiencing persistent breathlessness and his doctor decided to step up his treatment with addition of a third preventer drug, tiotropium, prior to his next dental appointment.

(l) What class of drug does tiotropium belong to and what is the mechanism of action?

(m) Name a side-effect affecting the oral cavity associated with inhaled tiotropium?

Medicine short notes question: Compare and contrast the mode of action, clinical indications and adverse effects of apixaban and warfarin.

Medicine sample SBA/MCQ question:

A 73 year old woman is taking atorvastatin. What is the key mode of action of this drug?
(a) Activation of HMG Co-A reductase enzyme
(b) Activation of PPARa receptors
(c) Inhibition of HMG Co-A reductase enzyme
(d) Reduced absorption of cholesterol from the intestine
(e) Reduced LDL receptor expression

Adaptation to open book extended duration format during the Covid-19 pandemic 2020-2022
Dentistry: the mid-year class test was held in January 2020 in the traditional closed book format before restrictions were introduced. The end of year paper held in May 2020 was delivered remotely in open-book format with extended duration of 24 hours for completion. The mid-year class test in January 2021 was also delivered remotely, in open-book format and was of 90 minutes duration rather than the usual 60 minutes. The end of year paper held in May 2021 was delivered remotely in open-book format, but time available for completion was scaled back from 24 hours to 3 hours. In January and May 2022 the class test and paper both reverted to closed-book in-person completion within a traditional examination venue and returned to pre-covid duration of 1 hour and 2 hours, respectively.

Medicine: the first mid-year class test was held in December 2019 in the traditional closed-book format before restrictions were introduced; the second mid-year class test scheduled for March 2020 was cancelled. Questions pertaining to Biochemistry and Genetics were omitted from the end of year short notes examination as the related learning outcomes had been examined in the December class test. Candidates received instead a paper in May 2020 which contained 5x pathology and 5x pharmacology and therapeutics short notes questions, which was delivered remotely in open-book format of an extended (24 hour) duration.

Data-sets and Analysis:

In this study, we analysed data on the performance of our dental students in closed book assessments (held prior to Covid-19, January 2015-January 2020), open-book assessments (delivered during the pandemic, May 2020-May 2021) and on return to closed book assessments (post-pandemic, January 2022-May 2022). We have also included data on the performance of our medical students in closed book assessments (held prior to Covid-19, May 2013-December 2019) and open book assessments (held during the pandemic, May 2020). [In September 2020, QUB Medical School introduced a new case-based learning curriculum and a strategy of assessing all subjects and specialties in an integrated fashion through regular progress testing (Heeneman et al, 2015) so we have not been able to include any comparable data beyond May 2020 in the present study]. Statistical analysis was undertaken using GraphPad Prism (Version 5) and SPSS (Version 23) to generate reliability and correlation coefficients. This research study relates to taught courses at The Queen's University of Belfast and received ethical clearance within the arrangements provided by the University for taught programmes.

Results

Dental Curriculum

In closed book, time-limited assessments held between January 2015 and January 2020 performance was 20.39+1.16 out of 30 (68%), mean + sd., n=6 in the mid-year class test and 51.46+2.8 out of 80 (64%), mean + sd., n=6 in the end of year examination (standard set cut score was 38.5+2.2 out of 80, mean +sd., n=6). Cronbach's alpha was 0.82+0.06 and 0.74+0.09, mean + sd., n=6 for the mid-year class test and end of year examination, respectively, evidencing reliability. No significant differences were noted between subsections /question type within the end of year examination in regards to reliability.

There were positive correlations between subsections of the paper indicating factorial validity: short notes v essay Pearson's r= 0.42+0.16; short notes v case study Pearson's r=0.44+0.22; essay v case study Pearson's r=0.42+0.09; mean + sd., n=6, each $p<0.001$. Performance in the mid-year test correlated ($p<0.001$) with performance in the end of year paper (Pearson's r=0.63+0.05, mean + sd., n=6) indicating predictive validity. Neither correlated with performance in the group presentation. As illustrated by data obtained for the 2019 cohort, the mark for the Pharmacology for Dentistry module correlated ($p<0.001$) with those for the Physiology for Dentistry (Pearson's r= 0.65) and Disease Mechanisms for Dentistry (Pearson's r= 0.71) modules, evidencing concurrent validity.

In 2021, in which the mid-year class test was delivered remotely in extended open-book format, performance was higher relative to that observed in previous years: 25.74+5.3 out of 30 (86%), mean + sd., 60 students. In 2022, the mid-year class test reverted to pre-Covid-19 in-person, closed-book short-duration format and performance reverted to pre-pandemic levels (19.74+5.3 out of 30 (66%) mean + sd., 56 students). Performance in the remotely-delivered open book end of year examination in 2020, which was of 24 hours duration, was also inflated: 68.97+6.44 out of 80 (86%), mean + sd., n=57 students. Grade inflation was less evident in 2021 when duration of the open book remotely delivered examination was reduced from 24 to 3 hours: 60.45+7.07 out of 80 (76%), mean + sd., n =60 students. In 2022, on reverting to a closed-book in-person time-limited examination, performance returned to (and indeed was lower than) pre-pandemic levels: 44.95+9.04 out of 80 (56%), mean + sd., n= 56 students.

Reliability of the class test was reduced on switching to remote open-book extended duration assessment in 2021 (Cronbach's alpha = 0.63) but returned to pre-pandemic levels with reinstatement of the in-person closed-book time-limited assessment in 2022 (Cronbach's alpha = 0.88). Similarly, reliability of the end of year paper was reduced by open-book remote delivery (24 hour assessment 2020, Cronbach's alpha = 0.65; 3 hour assessment 2021, Cronbach's alpha =0.68) but returned to pre-

pandemic levels with reintroduction of a time-limited in-person closed-book examination in 2022 (Cronbach's alpha= 0.77).

There was still evidence for correlation between sections of the remotely-delivered open-book written paper (during the pandemic in 2020 and 2021) evidencing factorial validity but correlations were less positive and less significant than when using closed-book time-limited format: short notes v essay, Pearson's r=0.37+0.11, p<0.01 mean +sd., n=2; short notes v case study: Pearson's r=0.35+0.14 p<0.05 mean + sd., n=2; case study v essay: Pearson's r=0.42+0.14 p<0.01 mean + sd., n=2. Performance in the mid-year class test still correlated with that in the end of year paper but less strongly than before indicating a reduction in predictive validity: 2020 Pearson's r= 0.25, p=0.06; 2021 Pearson's r=0.33, p<0.05. Correlation improved markedly in 2022 on return to closed-book, time-limited in-person assessment: Pearson's r=0.77, p<0.0001. with use of open-book assessments of extended duration during the pandemic, correlation between the Pharmacology for Dentistry module mark and those for the Physiology for Dentistry (Pearson's r= 0.45, p<0.001) and Disease Mechanisms for Dentistry (Pearson's r=0.57, p<0.0001) modules for the same cohort of 60 second year dental students (2021) was less positive but still significant. Correlation of marks achieved across modules improved markedly in 2022 on return to closed-book, time-limited in-person assessment: correlation between the Pharmacology for Dentistry module mark and Physiology for Dentistry module mark: Pearson's r= 0.73, p<0.0001; correlation between the Pharmacology for Dentistry module mark and Disease Mechanisms for Dentistry module mark: Pearson's r=0.75, p<0.0001.

Medical Curriculum

Reliability of the short notes paper increased with the number of questions included: during 2013-2016, in which candidates completed 5 questions in 60 minutes in closed-book format, the mean performance was 35.39 out of 50 (71%) and Cronbach's alpha was 0.79+0.06, mean + sd., n=4; from 2017-2019 during which candidates completed 3 questions in 36 minutes in closed-book format, the mean performance was 17.92 out of 30 (60%) and Cronbach's alpha was 0.66+0.10, mean + sd., n=3. The mean performance in the closed-book SBA class test introduced from 2017 onwards was 13.5 out of 20 (68%) for the 20 pharmacology questions completed in 30 minutes and KR20 was 0.67+0.03, mean + sd., n=3. Predictive validity was evidenced by correlation between pharmacology questions in the SBA class test and end of year short notes pharmacology questions in 2019 (Pearson's r=0.53, p<0.001). Factorial validity was shown by correlation between the pharmacology question styles within the short notes examination: basic principles of pharmacology v drug comparison question Pearson's r=0.46, n=7, p<0.0001.; basic principles of pharmacology v clinical case study Pearson's r=0.40, n=7, p<0.0001; drug comparison question v clinical case study Pearson's r=0.52, n=7, p<0.0001. Concurrent validity was evidenced by correlation between the pharmacology SBA questions and the pathology SBA questions within the same class test undertaken in 2019: Pearson's r=0.42, p<0.001. There was also correlation between sections of the closed-book, time-limited short notes assessment: pharmacology v pathology (10 questions in total 2013-2016): Pearson's r =0.61+0.06, mean + sd., n=4, p<0.001; pharmacology v pathology (6 questions in total 2017-2019) Pearson's r= 0.54+0.18, mean + sd., n=4, p<0.001; pharmacology v biochemistry and genetics (7 questions in total 2017-2019) Pearson's r= 0.53+0.16, mean + sd., n=3, p<0.001.

Performance in an 2020 open-book extended duration (24 hours) short notes examination delivered remotely during the pandemic was markedly greater than in the pre-pandemic closed-book time-limited examination; 86.94%+2.73, mean + sd., n=267. Reliability was reduced but not significantly different to that of the closed-book time limited examination in previous years; Cronbach's alpha was 0.70 (5 questions). Factorial validity between different question types within the paper was also reduced but remained statistically significant: basic principles of pharmacology v drug comparison question: Pearson's r= 0.33, p<0.0001; basic principles of pharmacology v clinical case study: Pearson's

r= 0.20, p<0.001; drug comparison question v clinical case study: Pearson's r= 0.36, p<0.0001. Concurrent validity was also reduced but remained significant: correlation with the pathology section of the paper: Pearson's r=0.30, p<0.001 .Similarly, predictive (criterion) validity was reduced but remained significant: correlation with the time-limited closed book SBA biochemistry and genetics class test earlier in the year: Pearson's r= 0.213, p<0.001.

Discussion

There is a lack of consensus regarding whether candidates tend to score more highly in SBA /MCQ assessments than in free response SAQ assessments (Sam et al 2016; Sullivan, 2011; Wilkinson and Shaw, 2015) and whether the former enhance reliability at the expense of validity (Hift, 2014; Patil et al., 2015). Preston et al (2020) have proposed that students perform more highly in SBA/MCQ assessments in part because these are most frequently employed and students become more familiar with this format as they progress through the course. SBA were incorporated into the profile for assessment of basic and clinical pharmacology learning outcomes within our undergraduate medical curriculum in 2017; we have found no evidence that students score consistently higher in such assessments than in traditional free response short notes examinations. However, our study was restricted to students in the early years of the 5 year medical and dental programmes during which students may still be becoming familiar with this format.

Furthermore, in our experience, appropriately constructed free response short notes papers can be as reliable as SBA papers for assessing basic and clinical pharmacology learning outcomes in closed-book, time-limited assessments: utilising approximately 30-35 minutes allocated to the pharmacology section of a larger assessment with the medical curriculum permitted the use of either 3 short notes questions or 20 SBA: Cronbach's alpha was similar regardless of which strategy was employed. Reliability of each assessment method would be expected to increase if more time was available enabling an increased number of questions to be used; for example we found Cronbach's alpha was higher when 5 short notes questions were set rather than 3. Student perception in regards to subjectivity in marking of free response answers (Holzinger et al, 2020) can be mitigated by construction of detailed model answers and internal (and external) quality assurance of the paper. In our experience such measures have resulted in an inter-maker variability of less than 5% particularly when internal examiners are subject experts of many years' standing and experienced markers. SBA arguably offer the convenience of automated marking, but short notes questions in our experience afford more opportunity for meaningful feedback at cohort and individual level which students consider to be very important to allow improvements (Preston et al, 2020), whereas protection of the security of the question bank of quality assured discriminating SBA questions is often prioritised over provision of specific and sufficiently detailed feedback to candidates and can encourage dependence by students on external SBA question banks of variable quality. Furthermore, Preston et al (2020) have reported that students perceive that performance in short answer questions more accurately reflects the effort they put into learning and their knowledge of the content material than that afforded by SBA/MCQ-based assessment.

For both forms of assessment, content and construct validity increase when a greater number of questions is used. The argument that SBA papers by virtue of containing a greater number of questions and allocating 1-2 minutes to answer each enhances content validity relative to a much smaller number of free response short notes questions allocated 10-15 minutes each for completion is misleading: appropriately constructed short notes questions (potentially with multiple parts) together with detailed model answers and marking schemes can permit assessment of multiple learning outcomes and aspects of the subject material within the same question. In both our medical and dental short notes assessments, factorial validity was evident when comparing across question types, for example comparing case studies and drug comparison questions, which supports the construction

of short notes papers that include a range of question formats. Positive correlation between SBA and short notes assessments within our medical curriculum also advocates for inclusion of a range of written assessment methods rather than dependence on one assessment type. The performance of our dental students in the written assessments did not however correlate with performance in the group presentation; this can be attributed to the latter assessing a different set of subject specific and generic learning outcomes relative to the written papers and also group size diluting individual performance.

The Covid-19 pandemic led to a dependence on remotely-delivered open-book assessments often of extended duration. A desire to protect the security of SBA question banks so that robust quality-assured SBA questions could be re-used at a future date also encouraged renewed interest in the use of short notes papers as an assessment strategy during the pandemic. Not unexpectedly, open-book, remotely-delivered assessments encouraged grade inflation although this could be mitigated in part by limiting the extent to which the duration of the assessment was extended to the absolute minimum necessary to facilitate downloading and uploading of the paper and typing of answers by candidates. The decision was taken not to alter the standard set cut score (which is based on assessor judgement of the performance of a borderline student by the modified-Angoff method, George et al., 2006) to account for the Covid-19 related change to open-book assessment delivery. Despite grade inflation, there was still opportunity to discriminate between stronger and weaker candidates, as evidenced by the spread of marks. Dependence on open-book extended duration assessments resulted in reduction in assessment reliability; this was unavoidable but the impact was more modest than anticipated and could be mitigated by inclusion of an increased number of questions. Open-book remotely delivered assessments did not reduce content or construct validity since sampling strategy and blueprinting was unchanged. There were modest reductions however in factorial, predictive and concurrent validity of assessments employed; however although less positive, correlations remained statistically significant. Such reductions may be accounted for by the open-book format and extended duration; access to learning resources and increased time available to construct answers often of greater length may have increased variation in a candidate's answering of differing styles and types of question within pharmacology and also more widely, across all subjects assessed concurrently.

A possible limitation of the current study is that analysis was undertaken by the investigators using historical spreadsheets provided by the Progress and Assessment Office Administrative Manager of anonymised data so it was not possible to undertake subgroup analysis of possible influencing variables: only the final ratified overall module mark was available to the investigators in de-anonymised form. The majority of students enrolled in our medical (60-75%) and dental (65-80%) programmes in the last decade are female. Routine quality assurance by our Progress and Assessment Office of the summative assessments undertaken within our medical and dental courses has not however found any statistical difference between male and female candidates in terms of their academic performance in either MCQ/SBA or free response short notes assessments providing reassurance that the assessment strategy employed would not confer a gender-related advantage. Small subgroup size has limited opportunity for analysis of other potential influences such as those with registered disability, or home /international status. Our medical and dental students enter university with a strong foundation in the sciences, acquired during secondary education, and/or a primary degree in a relevant subject. Caution is necessary when extrapolating to other disciplines for whom the SBA and/or SAQ assessment format may be less suitable: it would be interesting to explore the wider application of the study findings by comparing the reliability and validity of SBA and SAQ assessments in measuring performance of students undertaking courses in the arts and humanities which may wish to assess achievement of different generic and subject-specific learning outcomes.

In conclusion, appropriately constructed papers employing short notes free response answers are reliable and valid for assessing basic and clinical pharmacology learning outcomes in closed-book, time-limited written assessments. We would argue for their retention within the assessment strategy, alone or possibly used in combination with other forms of written assessment (such as SBA) to increase assessment variety, discourage cueing and foster deeper learning and critical understanding (Witt et al., 2022) as they afford opportunity for provision of detailed constructive feedback to candidates. Students also acknowledge the benefit afforded by inclusion of a variety of assessment types to accommodate for a range of learning styles and needs (Holzinger et al., 2020; Preston et al, 2020). Dependence on short notes papers for remote delivery of open-book extended duration assessments during the Covid-19 pandemic did not reduce content or construct validity but did cause modest reduction in factorial, predictive and concurrent validity and reliability. With the benefit of experience, we propose that this could potentially be mitigated in future by careful consideration of the optimum number of questions employed and restriction to assessment duration. Thorough item analysis could discard ambiguous or poorly performing questions to enhance reliability. Adaptation of questions to place greater emphasis on critical understanding and problem-solving rather than simple factual recall would also compensate for inability to rigorously enforce closed-book conditions on a remotely-delivered assessment without the significant logistical challenge of online proctoring of a large cohort of candidates.

References

Andreatta, P.B., Gruppen, L.D. (2009) 'Conceptualising and classifying validity evidence for simulation', *Medical Education 2009*;43, pp. 1028-1035.

Downing, S.M. (2003) On the meaningful in'terpretation of assessment data', *Medical Education 2003*; 37, pp. 830-837.

Fallatah, H.I., Tekian, A., Park, Y.S., Al Shawa, L. (2015) 'The validity and reliability of the 6th Year internal medicine examination', *BMC Medical Education 2015*; 15, pp.10.

General Dental Council (2015) Preparing for Practice 2015. https://www.gdc-uk.org/docs/default-source/quality-assurance/preparing-for-practice-%28revised-2015%29.pdf (Accessed: September 2022).

General Medical Council (2018) Outcomes for Graduates 2018. https://www.gmc-uk.org/education/standards-guidance-and-curricula/standards-and-outcomes/outcomes-for-graduates/outcomes-for-graduates (Accessed 21 September 2022).

George, S, Haque, M.S., Oyebode, F. (2006) 'Standard setting: comparison of two methods', *BMC Medical Education* 2006; 6, pp. 46.

Heeneman, S., Schut, S., Donkers, J., van der Vleuten C.P.M., Muijtiens, A.J.J.M. (2016) 'Embedding of the progress test in an assessment program d,esigned according to the principles of programmatic assessment', Medical Teacher; *DOI: 10.1080/0142159X.2016.1230183*

Hift, R.J. (2014) 'Should essays and other open-ended type questions retain a place in written summative assessment I clinical medicine?,' *BMC Medical Education 2014*; 14, pp.249.

Holzinger, A., Lettner, S., Steiner-Hofbauer, V., Melser, M.C. (2020) 'How to assess? Perceptions and preferences of undergraduate medical students concerning traditional assessment methods', *BMC Medical Education 2020*; 20, pp. 312.

Jaenicke, J., Jhet, M. T, Dave D, Makeev, V., Nasim, S.A. (2020) 'Student reflections on clinical prioritisation questions and their assessment', *Med Teach 2020*; 42. pp. 1193-1194.

Patil, S.Y., Gosavi, M., Bannur, H.B., Ratnakar, A. (2015) 'Blueprinting in assessment: a tool to increase validity of undergraduate written examinations in pathology', *Int J App Basic Med Res 2015*; 5 Suppl 1.

Preston, R., Gratani, M., Owens, K., Roche, P., Zimany, M., Malau-Aduli, B.(2020) 'Exploring the impact of assessment on medical students'', *Learning. Assessment & Evaluation in Higher Education 2020*; 1, pp. 109-124.

Sam, A.H., Hameed, S., Harris, J., Meeran, K. (2016) 'Validity of very short answer versus single best answer questions for undergraduate assessment', *BMC Medical Education 2016*; 16, pp.266.

Sullivan, G.M. (2011) A primer on the validity of assessment instruments', *J Grad Med Ed 2011*; 3, pp.120-121.

Wilkinson, T., Shaw, H. (2015) 'Are spot test multiple choice questions easier to answer than short answer questions?', *FASEB J 2015*; 29. Pp.344 (1).

Witt, E.E., Onorato, S.E., Schwartzstein, R.M. (2022) 'Medical students and the drive for a single right answer', *ATS Scholar 2022*; 3, pp. 27-37.