Evaluation of assessment marks in the clinical years of an undergraduate medical training programme: Where are we and how can we improve?

H Brits,[1] MB ChB, MFamMed, MHPE, FCFP; **G Joubert,**[2] BA, MSc;
J Bezuidenhout,[3] BA, HDE, BA Hons (Psychol), PG Dip (HPE), MEd (Psychol Ed), DTechEd;
L van der Merwe,[4] MB ChB, MMedSc, DA (SA), PhD (HPE)

[1] *Department of Family Medicine, School of Clinical Medicine, University of the Free State, Bloemfontein, South Africa*
[2] *Department of Biostatistics, Faculty of Health Sciences, University of the Free State, Bloemfontein, South Africa*
[3] *Department of Health Sciences Education, Faculty of Health Sciences, University of the Free State, Bloemfontein, South Africa*
[4] *Undergraduate Programme Management, School of Clinical Medicine, University of the Free State, Bloemfontein, South Africa*

Corresponding author: *H Brits (britsh@ufs.ac.za)*

Background. In high-stakes assessments, the accuracy and consistency of the decision to pass or fail a student is as important as the reliability of the assessment.
Objective. To evaluate the reliability of results of high-stakes assessments in the clinical phase of the undergraduate medical programme at the University of the Free State, as a step to make recommendations for improving quality assessment.
Methods. A cohort analytical study design was used. The final, end-of-block marks and the end-of-year assessment marks of both fourth-year and final-year medical students over 3 years were compared for decision reliability, test-retest reliability, stability and reproducibility.
Results. 1 380 marks in 26 assessments were evaluated. The G-index of agreement for decision reliability ranged from 0.86 to 0.98. In 88.9% of assessments, the test-retest correlation coefficient was <0.7. Mean marks for end-of-block and end-of-year assessments were similar. However, the standard deviations of differences between end-of-block and end-of-year assessment marks were high. Multiple-choice questions (MCQs) and objective structured clinical examinations (OSCEs) yielded good reliability results.
Conclusion. The reliability of pass/fail outcome decisions was good. The test reliability, as well as stability and reproducibility of individual student marks, could not be accurately replicated. The use of MCQs and OSCEs are practical examples of where the number of assessments can be increased to improve reliability. In order to increase the number of assessments and to reduce the stress of high-stake assessments, more workplace-based assessment with observed clinical cases is recommended.

Afr J Health Professions Educ 2021;13(4):222-228. https://doi.org/10.7196/AJHPE.2021.v13i4.1379

The pass/fail decision in summative assessment for medicine and other professional qualifications holds many consequences for the various stakeholders.[1-3] Failure, and having to repeat modules, has financial and emotional implications for students, while they may lose trust in the training institution.[4] Student failure may affect throughput rates, as well as the reputation of the faculty or university.[4] However, passing an incompetent student may affect both patients and the healthcare system, e.g. through loss of life and avoidable expenses. It could also lead to misconduct claims against individuals or institutions.[5,6] Miller[7] emphasises this important responsibility relating to assessment:

> 'If we are to be faithful to the charge placed upon us by society to certify the adequacy of clinical performance … then we can no longer evade the responsibility for finding a method that will allow us to do so.'

If we are to be able to defend the outcome of high-stakes examinations, where the outcome has major consequences,[8,9] the assessment must meet the basic requirements of validity, reliability and fairness.[10,11] From a theoretical perspective, it is possible to improve the quality of assessment by addressing criteria such as validity, reliability and fairness.[12]

An assessment is considered valid when it measures what it is supposed to measure.[13,14] In the case of clinical medicine, competence must be measured. Validity in clinical assessment is usually evaluated using Miller's assessment framework.[15] According to this model, a valid assessment for competence must be on the 'show how' and 'does' levels. However, when validity is increased by assessing in real-life situations, the reliability of the assessment may decline, owing to subjective judgements and the lack of standardisation.[16] Before the validity of an assessment can be evaluated, its reliability must be established.[1,4]

The reliability of a clinical assessment is defined as the degree to which a test measures the same concept in different assessments and obtains stable or reproducible results.[17,18] Reproducibility, a synonym for reliability, is described as the closeness of or variation in results of successive measurements of the same assessment carried out under the same or nearly the same conditions.[19] With any assessment, some form of 'measurement error' will occur. This error should be as low as possible to ensure accurate assessment. The calculation of this error determines the reliability of an assessment.[18] Reliability can be evaluated using various measures, depending on the data that are available and what one wants to establish.[20]

From a theoretical viewpoint, an assessment can be considered fair if everybody is subjected to the same assessment, under the same conditions, and all are marked by the same assessors using the same mark sheets.[21] In practice, an assessment is fair when the interpretation of the results is transparent and just, and when nobody is disadvantaged in the process.[22]

One of the aims of assessment evaluation should always be to improve the quality of assessment for all stakeholders.[2] A fine balance should exist between traditional and innovative assessment methods, by selecting judiciously sound assessment methods above tradition or convenience.[9] The decision to change or improve assessment practices or to move towards more innovative assessments should be based on facts rather than preferences.[9]

Pass/fail decisions are made based on predetermined criteria. In high-stakes assessments, the accuracy and consistency of the decision to pass or fail a student are as important as the reliability of the test or assessment.[3,7,23,24] Decision reliability is a term used to measure the consistency with which pass/fail decisions are made.[3]

The best way to evaluate the reliability of a clinical assessment is to assess the same participants under similar circumstances on more than one occasion,[25] which is almost impossible in real-life situations. The reliability of an assessment can be improved by using standardised questions and mark sheets, and multiple and trained markers, and by increasing the number of questions.[20] A high correlation between the different test scores (r>0.7) is indicative of test-retest reliability.[26]

The undergraduate medical programme at the University of the Free State (UFS) is a 5-year, outcomes-based programme that runs over 10 semesters. The clinical phase is presented from semesters 6 - 10. In the clinical phase of the programme, students are assessed in different disciplines. Some disciplines are grouped together to form a module. For example, in the fourth year, the surgery module consists of general surgery, orthopaedics, ophthalmology and otorhinolaryngology. Modules are presented in blocks. Students rotate between different blocks to cover all modules presented in the specific year. At the end of each rotation (block), students are assessed by the end-of-block assessment. In the fourth year, students must pass all disciplines to progress to the fifth year. If students meet minimum requirements in the fifth year, but fail certain disciplines, they are required to repeat only the failed disciplines. Admission to the final end-of-year assessment in the fourth and final year requires that students meet end-of-block academic as well as attendance requirements. Students in the fourth and final years must pass all disciplines in all the modules, including each of the clinical and theoretical components individually (if applicable), to pass the final end-of-year assessment.[27] Regarding clinical cases, students must also pass more than 50% of the cases, irrespective of the overall clinical mark obtained. If a student fails the end-of-year assessment (in either fourth or final year), but meets minimum requirements for reassessment, the student is allowed to do a reassessment within 1 week of the end-of-year assessment.[27] The pass mark for assessments is predetermined at 50%, as per university regulations. No formal standard-setting process exits. Assessments are blueprinted, and assessment rubrics or memoranda are moderated before assessments. Fig. 1 shows a flow diagram of the assessment process.

The end-of-block assessment and the end-of year assessment cover the same content, and are generally conducted by the same assessors (academic staff in clinical departments). Both these assessments consist of theoretical as well as clinical assessments. Different disciplines structure their assessments differently, which makes comparison between disciplines not feasible. No regulation or specific reason was found for conducting an end-of-year assessment after the end-of-block assessments, and it is possibly more traditional than evidence based.

Despite the implications of high-stakes assessment results – such as in the undergraduate medical programme – there are no guidelines for educational institutions to measure the quality of their assessments. Therefore, educational institutions should institute quality assurance measures to ensure quality assessment, and be able to defend these results.

The aim of this study was to evaluate the assessment results of high-stakes assessments in the clinical phase of the undergraduate medical programme. As a first step to improve the quality of assessment in the clinical years of undergraduate medical training, the reliability of current assessments was established. This will assist to make recommendations for improving the quality of current assessments in the undergraduate medical programme, with validity, reliability and fairness in mind.

The objectives were as follows:

(*i*) to determine the decision reliability of the current summative assessments, and whether pass/fail decisions can be defended
(*ii*) to determine the test-retest correlation between different assessments
(*iii*) to compare the reliability results of different assessment methods.

Methods

A cohort analytical study design was used. The study population consisted of all the fourth-year and fifth (final)-year undergraduate medical students at UFS who participated in the last end-of-block and end-of-year assessments of 2016, 2017 and 2018. The last end-of-block marks (obtained during the last rotation of the year) and the end-of-year assessment marks obtained during the final assessment at the end of the academic year were used for data analysis. Data were collected retrospectively. Between the last end-of-block assessment and the end-of-year assessment, no formal training and very little learning takes place, which makes these assessments comparable, but not identical.

The authors used an aggregated approach to look at the reliability of assessments, as an individual approach was impossible owing to the variability in the way each discipline designs multiple choice questions (MCQs), clinical cases and objective structured clinical evaluations (OSCEs) and/or objective

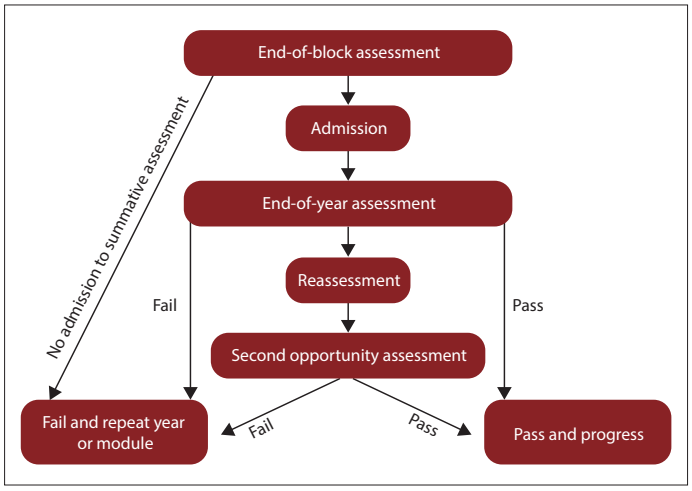

Fig. 1. Flow diagram of assessment process and outcome in fourth and final years of the undergraduate medical programme, University of the Free State.

structured practical evaluations (OSPEs). The reliability of the theoretical and clinical assessments was determined separately. Theoretical assessments consisted of papers with MCQs only, and papers with a combination of MCQs and mostly short written questions. Clinical assessments included clinical cases, OSCEs and OSPEs. In clinical cases, the student assesses a patient unobserved and then reports on findings while the assessors clarify findings and ask predetermined questions. The term OSCE is used for assessments in the form of clinical stations with patients or simulated patients. Students were directly observed at these clinical stations. The term OSPE was used for assessment involving unmanned stations, where students had to interpret diagnostic investigations, e.g. X-rays or laboratory results. Different disciplines use different combinations of assessments; however disciplines use the same combinations during end-of-block and end-of-year assessments.

Table 1 categorises the disciplines as either surgical or medical, indicates the study year(s) in which a discipline is presented and lists the different assessment methods used for each discipline. General surgery, orthopaedics, urology, otorhinolaryngology, ophthalmology, anaesthetics and obstetrics and gynaecology are classified as surgical disciplines (n=7). Internal medicine, paediatrics, family medicine, oncology and psychiatry are classified as medical disciplines (n=5).

Data collection

Student marks, corresponding with respective student numbers, were obtained from the official marks database used by the Faculty of Health Sciences. This is an extensive database with numerous datasets, available in Excel (Microsoft, USA) spreadsheets for each student per discipline and per assessment. It is a secure database with password protection – only authorised access is permitted. All marks, including of reassessments, were used to compare final pass/fail outcome decisions.

Data management and analysis

The Department of Biostatistics performed data analysis using SAS version 9.4 (SAS, USA). Calculations were done per discipline for fourth-year and fifth-year students separately.

The decision reliability between the final end-of-block and end-of-year assessment was calculated using 2 × 2 tables. Due to the skewed data, kappa values could not be calculated[28] and a value ≥0.7 on Holley and Guilford's[29] G-index of agreement (as an alternative for categorical judgement) was considered as reliable. Holley and Gilford's G-index of agreement allows for correlation in the presence of skewed data. As a final step to evaluate the pass/fail outcome decisions, the reassessment outcome decision was compared with the final end-of-block and end-of-year assessment outcome decisions.

To determine test-retest reliability between final end-of-block and end-of-year assessment marks, Pearson correlation coefficients were calculated. A correlation coefficient ≥0.7 was considered as reliable.[26]

The mean and standard deviation (SD) of differences between end-of-block and end-of-year assessment marks were calculated. This mean is used as an indication of assessment stability. The percentage of students whose marks for the end-of-block and end-of-year assessments differed by <10% for the two assessments was calculated to assess reproducibility. The assessment was considered reliable if the reproducibility was ≥80%.

For clinical cases, the individual student marks obtained in consecutive assessments performed on the same day were also compared. The means of the different cases were compared to determine test consistency, and the variance in marks (SD) obtained by individual students was calculated to determine reproducibility.

Table 1. Classification of disciplines, study years of presentation and types of assessment per discipline

Discipline	Classification	Study year	Assessment types
A	Surgical	4 and 5	Theory, clinical
B	Surgical	4 and 5	Theory, clinical
C	Surgical	4 and 5	Theory, clinical
D	Surgical	4	Theory
E	Surgical	4	Theory
F	Surgical	5	Theory, clinical
G	Surgical	5	Theory
H	Medical	4 and 5	Theory, clinical
I	Medical	4 and 5	Theory, clinical
J	Medical	4 and 5	Combined
K	Medical	4	Theory
L	Medical	5	Combined

Ethical considerations, quality and rigour of data management

Ethical approval to conduct the study was obtained from the Health Sciences Research Ethics Committee of UFS (ref. no. UFS-HSD 2019/0001/2304), and permission to use student data was granted by the relevant university authorities. All data were managed confidentially and only student numbers were used. No student or discipline is identified in the published results.

Results

A total of 1 380 marks in a total of 26 administered assessments were evaluated. In Table 2, the numbers of students included in the study per discipline are indicated for the different years. Some disciplines are presented in only one of the study years (Table 1). The study used the marks of 12 disciplines within the medical programme.

Decision reliability of pass/fail decisions

In 2 of the 12 fourth-year assessments, and 7 of the 14 fifth-year assessments, the pass/fail decisions in the final end-of-block concurred with the end-of-year assessments, and all students passed. In the remaining disciplines, there were between 92.5% and 98.9% agreement of the same pass/fail decision outcome between the end-of-block and end-of-year assessments. The G-index of agreement values ranged from 0.86 to 0.98.

Three fourth-year students obtained marks <50% in the final end-of-block assessment. They subsequently failed the end-of-year assessment too, as well as the reassessment, and therefore had to repeat the year. No fifth-year students obtained marks <50% in the final end-of-block assessment, or failed the year. Three fourth-year students and two fifth-year students passed the final end-of-block assessments, and then failed a subcomponent of a discipline/module in the end-of-year assessment. All these students qualified for reassessment, according to the rules, and all passed the reassessment and, therefore, passed the year.

Test-retest correlation of end-of-block and end-of-year results

In the fourth and fifth years, respectively, 12 and 15 assessments were compared for test-retest correlation. Three assessments in the fourth year had correlation coefficients ≥0.70. None of the assessments in the fifth year had correlation coefficients ≥0.70. These results are displayed in Table 3.

Stability of assessment marks per discipline

Table 4 summarises the differences between the final end-of-block and the end-of-year assessment marks per discipline and per study year. The mean differences between marks obtained in the final end-of-block and end-of-year assessments varied between –11.4% (discipline K, fourth-year group) and 7.5% (discipline F, fifth-year group), with discipline K emerging as a clear outlier.

Table 2. Students per discipline for different study years, *n*

	4th year		
Discipline	**2016**	**2017**	**2018**
A	30	37	26
B	30	37	26
C	30	37	26
D	30	37	26
E	28	37	27
F	21	22	17
G	21	22	17
H	21	19	19
I	22	22	16
	5th year		
A	23	29	35
B	23	29	35
C	14	20	18
F	23	29	35
G	31	28	37
H	19	19	19
I	31	28	37
J	19	18	21
L	18	18	21

Reproducibility and assessment methods

The percentage of students whose final end-of-block and end-of-year assessment marks were within a 10% range varied between 33.3% (discipline K fourth year) and 98.9% (discipline I fifth year). The individual marks of students varied considerably, as indicated by the high SD, particularly for the fourth-year group. In Table 5 these percentages are given for the different assessment methods.

Differences between marks for consecutive clinical cases

In three disciplines, students were assessed on two or three clinical cases on the same day. The mean marks obtained per discipline were within 4.5% of each other. The marks that individual students obtained varied by between 0 and 45% for different cases in the same discipline. In Table 6, the mean, SD, minimum and maximum of differences in student marks obtained for consecutive cases are indicated per discipline.

Discussion

The results presented here may be considered representative of the selected study population, as all the student marks were available, in a usable format, in the database.

The aim when evaluating the quality of an assessment should be to identify areas that can be improved in the assessment.[30] Data for this study were obtained with this aim in mind rather than to pronounce judgement on the reliability of current assessment methods and practices.

Calculating the reliability of pass/fail outcome decisions using a kappa coefficient is described in the literature.[9,31] In this study, very few students failed, and the small numbers made the kappa statistic inappropriate for this measurement.[32] A G-index of agreement was, therefore, calculated.[28] In almost half (45.2%) the disciplines investigated, the agreement between the outcomes obtained in assessments was 100%. For the remaining disciplines, the G-index of agreement was >0.85. The decision reliability on pass/fail outcome decisions for clinical assessments in the undergraduate medical programme at UFS can, therefore, be considered excellent. The comprehensive end-of-block assessments, the strict admission requirements to the end-of-year assessment and the reassessment opportunity may be reasons for this finding. Each individual student result, as well as discipline-specific results, are discussed at the examination admission and final

Table 3. Correlation between final end-of-block and end-of-year assessment marks, per discipline, study year and type of assessment

	4th year			5th year		
Discipline	**Theory***	**Clinical***	**Combined***	**Theory***	**Clinical***	**Combined***
A	0.39* (p<0.01)	0.47 (p<0.01)	-	0.61 (p<0.01)	0.34 (p<0.01)	-
B	0.32 (p<0.01)	0.48 (p<0.01)	-	0.23 (p<0.01)	0.24 (p=0.03)	-
C	-	-	0.60 (p<0.01)	0.34 (p=0.01)	0.67 (p<0.01)	-
D	0.80 (p<0.01)	-	-	-	-	-
E	0.50 (p<0.01)	-	-	-	-	-
F	-	-	-	0.57 (p<0.01)	0.40 (p<0.01)	-
G	-	-	-	0.25 (p=0.01)		-
H	0.61 (p<0.01)	-	-	0.62 (p<0.01)	0.35 (p<0.01)	-
I	0.78 (p<0.01)	0.93 (p<0.01)	-	0.64 (p<0.01)	0.32 (p<0.01)	-
J	-		0.23 (p=0.7)	-	-	0.66 (p<0.01)
K	0.43 (p<0.01)	-	-	-	-	-
L	-	-	-	-	-	0.46 (p<0.01)

*Correlation coefficient.

Table 4. Differences between the end-of-block and end-of-year assessment marks, per discipline and study year

	4th year	5th year
Discipline	Mean (SD)*	Mean (SD)*
A Theoretical	3.29 (11.04)	−0.87 (7.25)
A Clinical	−3.44 (9.21)	−4.51 (7.63)
B Theoretical	1.90 (10.67)	−0.06 (9.82)
B Clinical	−1.22 (10.89)	−0.64 (7.63)
C Theoretical	2.05 (6.81)	2.06 (8.62)
C Clinical	-	0.79 (6.12)
D Theoretical	−1.91 (4.94)	-
E Theoretical	−1.51 (6.27)	-
F Theoretical	-	−2.29 (9.10)
F Clinical	-	7.45 (8.59)
G Theoretical	-	−4.16 (10.63)
H Combined	−0.78 (5.69)	1.11 (4.68)
I Theoretical	4.43 (4.75)	4.08 (6.01)
I Clinical	−1.36 (2.10)	−7.27 (7.36)
J Combined	2.65 (12.38)	0.97 (5.55)
K Theoretical	−11.42 (12.84)	-
L Combined	-	-1.44 (5.30)

SD = standard deviation.
*A positive mean indicates that end-of-year marks were higher than end-of-block marks, while a negative mean indicates that end-of-year marks were lower than end-of-block marks.

Table 5. Percentage of students whose final end-of-block assessment and end-of-year marks were within a 10% range, by discipline, year group and assessment method

	Theory		Clinical		
	MCQ	Combined paper	Clinical case	OSCE	OSPE
	4th year, discipline				
A	67.7*	-	-	-	74.2
B	65.6†	-	-	-	64.5
C	88.3	-	-	-	-
D		95.7	-	-	-
E		91.4	-	-	-
H	91.5	-	-	-	-
I	90.2	-	-	98.9	-
J	-	-	-	-	56.7
K	-	33.3	-	-	-
	5th year, discipline				
A	88.5	-	71.3	-	-
B	72.4†	-	-	-	81.6
C	-	75.0	-	90.4	-
F	69.0†	-	-	-	62.1
G	-	64.6	-	-	-
H	98.3	-	60.3	-	-
I	87.5	-	63.5	-	-
J	-	-	-	94.8	-
L	-	-	-	96.5	-

MCQ = multiple choice question; OSCE = objective structured clinical examination; OSPE = objective structured practical evaluation.
*This assessment originally consisted of 30 questions, but the number of questions increased in 2017.
†These assessments consisted of ≤30 or fewer questions per assessment.

Table 6. Differences in marks obtained for consecutive cases per discipline

Discipline	Case	Mean difference	SD	Minimum	Maximum
A	1 and 2	1.99	12.21	−32	45
	1 and 3	1.85	12.78	−35	37
	2 and 3	−0.15	12.98	−37	30
H	1 and 2	−2.32	14.99	−45	38
	1 and 3	−4.36	14.55	−44	35
	2 and 3	−2.04	13.20	−36	43
I	1 and 2	1.67	11.44	−25	30

SD = standard deviation.

examination meeting to ensure defensible outcomes. With the current measures in place, and the addition of standard-setting to ensure accurate pass/fail decisions during end-of-block assessments, the necessity of an end-of-year assessment may be reconsidered.

The test-retest correlations were low, and did not reach a value ≥0.7 for any of the fifth-year students' assessments. This indicates poor reliability for individual assessments. The reliability of an assessment can be affected by the students, the test and the markers.[17,20] Student factors that could contribute to the low test-retest correlations include the fact that students who had passed the recent final end-of-block assessment might be confident about passing the end-of-year assessment, and then opt to study for disciplines/modules in which they had passed the end-of-block assessment some time ago. The added stress of high-stakes assessment, together with uncertainty about future work and placement, could also influence students' performance. Performance stress during high-stakes assessments is well described.[33,34] The effect of additional stress is unpredictable – it can have a positive or negative effect on academic performance.[35] More and regular low-stakes assessments may address the student factors described above. Test factors that could have played a role in this study include the fact that even though the same content was assessed in both the final end-of-block and end-of-year assessments, the questions differed, and no formal standard-setting was performed. Furthermore, not all competencies can be tested in all assessments, and very few assessments performing summative assessments. Competency in one case has poor reproducibility for another case.[1] Finally, the markers stayed the same during both assessments, with the exception of a few additional external assessors. By increasing the number of assessments during rotations, the reliability of overall assessment can also be improved.

The mean marks obtained in the end-of-block compared to end-of-year assessments did not differ much. The exception was the theoretical assessment in one discipline in the fourth year, where the end-of-year mark obtained was 11.4% lower than the final end-of-block module mark. The reason for this difference is not clear, though moderation reports of these assessments could provide some insight. The small variation in the mean marks (end-of-block v. end-of-year) per discipline may be an indication that the assessments were of the same standard. However, the SD was high for all assessments, indicating large differences in the marks obtained by individual students in the two assessments. These differences occurred in theoretical as well as clinical assessments. Poor validity of the assessments, or the student factors discussed above, may be reasons for these differences.

Assessment methods varied across disciplines, and therefore, direct comparisons could not be made between different assessment methods. For theoretical assessments, MCQ papers with >30 questions produced student marks within a 10% range, indicating reproducibility. Reproducibility could not be proved for assessments with <30 questions. The reproducibility of assessments can be improved by increasing the number of questions.[20] Clinical OSCEs yielded good reproducibility results, while OSPEs and clinical cases did not. Patrício *et al.*[36] analysed the results of 366 articles on OSCEs performed in undergraduate medical education, and concluded that OSCEs produce reliable results and are feasible for assessing competence. An OSCE in itself is not reliable, but can produce reliable results if adequate sampling, good-quality questions and mark sheets, time allocation per station and trained assessors are used.[19,20,37] OSPEs lack clinical interaction and demonstration of competence, making OSPEs almost equivalent to written questions.[38]

Clinical cases or long cases are renowned for their poor validity and reliability.[39] Evaluation of the marks obtained for consecutive clinical cases revealed a high SD, despite a stable mean mark. A difference of up to 45% was observed in marks obtained for different clinical cases performed by the same student. A possible reason may be patient selection and reuse of patients for the assessment. It is difficult to find enough suitable, similar and stable patients to use in clinical cases, making long cases less practical and reliable for summative assessment.[40] Assessors also need to make subjective judgements of competence, which may influence reliability. Nevertheless, clinical cases have a definite role to play in low-stakes and formative assessment in which the aim is learning.[1] An advantage of using long cases is that a student can be assessed holistically on an actual case.[40] This advantage is lost when the student's examination of the patient is unobserved, and is followed by the student reporting his/her findings.[41] It has been calculated that 10 clinical cases are necessary to achieve acceptable reliability with clinical cases.[42] These numbers are only possible when workplace-based assessments are used.[9] Based on the above, it is recommended that clinical cases only be used for formative assessment.

Although these results are setting-specific, the recommendations and conclusion can be applied to other settings as they are supported by the latest literature. Reliability is only one aspect of quality assessment to ensure clinical competence. To achieve quality assessment of clinical competence, students should be assessed in real life, or in near-real-life situations. Assessing clinical competence is a complex procedure, with many dimensions requiring different assessment methods.[1,30,43] The highest level of competence, according to Miller's[7] framework for assessment, is 'does'. To ensure the competence of future medical professionals, we should assess them frequently, and in the workplace, and move away from overemphasis on high-stakes assessments.[44] Miller[7] states that:

> 'No single assessment method can provide all the data required for judgment of anything so complex as the delivery of professional services by a successful physician.'

However, real-life situations are not stable and reproducible. This poses challenges in ensuring the reliability of assessments.[45] It is important to take the quality of the assessment process as a whole into account, and to avoid merely focusing on validity, reliability or fairness as individual components to improve the assessment.

Study limitations

The quality of pass/fail decisions for the individual assessments (end-of-block and end-of-year) were not formally established before these assessments were compared with each other. However, the outcome of each assessment per student is discussed during the examination admission meeting and the examination results meeting to ensure accurate decisions.

The validity and fairness of the assessments were not assessed in the present article. This article is only a step in the process to assess the quality of assessment.

The end-of-block and end-of-year assessments that were compared are not identical, but comparable. It is almost impossible to get identical assessments in clinical medicine, as it is performed in real-life situations.

Results of students were grouped together per discipline, and not displayed per individual student per discipline. The aim of this article was not, however, to look at individual students or assessments, but at a collective.

Conclusion

The reliability of pass/fail outcome decisions in clinical assessments in the undergraduate medical programme involved in this study was found to be good. The necessity of end-of-year assessment after comprehensive end-of-block assessments may be questioned. The test reliability, as well as stability and reproducibility of individual student marks, were less acceptable. The use of MCQs and OSCEs are practical examples where the number of assessments can be increased to improve reliability. In order to increase the number of assessments and to reduce the stress of high-stakes assessment, more workplace-based assessment with observed clinical cases can be recommended.

Declaration. This article is based on a PhD study.

Acknowledgements. Me Hettie Human for the language editing.

Author contributions. HB – conceptualisation of study, protocol development, data collection and writing of article; GJ – assisted with concept and methodology, performed data analysis and assisted with interpretation and writeup; JB and LJvdM – promotors who assisted with conceptualisation and planning of the study, as well as critical evaluation and final approval of the manuscript.

Funding. None.

Conflicts of interest. None.

1. Amin Z, Seng CY, Eng KH (editors). Practical Guide to Medical Student Assessment. Singapore: World Scientific Publishing, 2006.
2. Hays RB, Hamlin G, Crane L. Twelve tips for increasing the defensibility of assessment decisions. Med Teach 2015;37(5):433-436. https://doi.org/10.3109/0142159x.2014.943711
3. Möltner A, Tımbıl S, Jünger J. The reliability of the pass/fail decision for assessments comprised of multiple components. Soc Behav Sci 2015;32(4):42. https://doi.org/10.3205/zma000984
4. Najimi A, Sharifirad G, Amini MM, Meftagh SD. Academic failure and students' viewpoint: The influence of individual, internal and external organisational factors. JEHP 2013;2:22. https://doi.org/10.4103/2277-9531.112698
5. Bates DW, Slight SP. Medication errors: What is their impact? Mayo Clin Proc 2014;89(8):1027-1029. https://doi.org/10.1016/j.mayocp.2014.06.014
6. Swaminath G, Raguram R. Medical errors I: The problem. Ind J Psychiatr 2010;52(2):110-112. https://doi.org/10.4103/0019-5545.64580
7. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990;65:S63-S67. https://doi.org/10.1097/00001888-199009000-00045
8. South African Qualification Authority. National policy and criteria for designing and implementing assessment for NQF qualifications and part-qualifications and professional designations in South Africa. Pretoria: SAQA, 2015. http://www.saqa.org.za/docs/pol/2015/National%20Policy%20for%20Assessment.pdf (accessed 12 September 2019).
9. Yudkowski R, Park YS, Downing SM. Introduction to assessment in health professions education. In: Yudkowski R, Park YS, Downing SM (editors). Assessment in Health Professions Education. 2nd ed. New York: Routledge, 2019.
10. Räisänen M, Tuononen T, Postareff L, Hailikari T, Virtanen V. Students' and teachers' experiences of the validity and reliability of assessment in a bioscience course. High Educ Stud 2016;6(4):181-189. https://doi.org/10.5539/hes.v6n4p181

11. Van der Vleuten CPM. The assessment of professional competence: Developments, research and practical implications. Adv Health Sci Educ 1996;1:41-67. https://doi.org/10.1007/BF00596229
12. Peck C. Principles of sound assessment practice in health professions education. ECPP 2017;5(5):150-157.
13. Van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: From methods to programmes. Med Educ 2005;39(3):309-317. https://doi.org/10.1111/j.1365-2929.2005.02094.x
14. Katzenellenbogen J, Joubert G. Data collection and measurement. In: Ehrlich R, Joubert G (editors). Epidemiology – a Research Manual for South Africa. 3rd ed. Cape Town: Oxford University Press, 2014.
15. Pangaro L, Ten Cate O. Frameworks for learner assessment in medicine: AMEE Guide No. 78. Med Teach 2013;35(6):e1197-e1210. https://doi.org/10.3109/0142159X.2013.788789
16. Williams R, Klamen D, Mcgaghie W. Cognitive, social and environmental sources of bias in clinical performance ratings. Teach Learn Med 2003;15:270-292. https://doi.org/10.1207/S15328015TLM1504_11
17. Manterola C, Grande L, Otzen T, García N, Salazar P, Quiroz G. Reliability, precision or reproducibility of the measurements. Methods of assessment, utility and applications in clinical practice. Rev Chilena Infectol 2018;35(6):680-688. https://doi.org/10.4067/S0716-10182018000600680
18. Pietersen J, Maree K. Standardisation of a questionnaire. In: Maree K (editor). First Steps in Research. 7th impression. Pretoria: Van Schaik Publishers, 2016.
19. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: Analysis of measurement errors in continuous variables. UOG 2008;31:466-475. https://doi.org/10.1002/uog.5256
20. Tisi J, Whitehouse G, Maughan S, Burdett N. A review of literature on marking reliability research. Report for Ofqual. Slough: National Foundation for Educational Research, 2013. https://www.nfer.ac.uk/publications/mark01/mark01.pdf (accessed 12 September 2019).
21. Kane M. Validity and fairness. Lang Test 2010;27(2):177-182. https://doi.org/10.1177/0265532209349467
22. Gipps C. Fairness in assessment. In: Wyatt-Smith C, Cumming JJ (editors). Educational Assessment in the 21st Century. Dordrecht: Springer, 2009.
23. Gugiu C, Gugiu M. Determining the minimum reliability standard based on a decision criterion. J Exp Educ 2018;86(3):458-472. https://doi.org/10.1080/00220973.2017.1315712
24. Stoker HW, Impara JC. 7 Basic psychometric issues in licensure testing. In: Impara JC (editor). Licensure Testing: Purposes, Procedures, and Practices.Lincoln, NE: Buros, 1995:167-186. http://digitalcommons.unl.edu/buroslicensure/12 (accessed 12 September 2019).
25. Heale R, Twycross A. Validity and reliability in quantitative studies. Evid Based Nurs 2015;18(3):66-67. https://doi.org/10.1136/eb-2015-102129
26. Sauro J. 2015. How to measure the reliability of your methods and metrics. https://measuringu.com/measure-reliability/ (accessed 12 September 2019).
27. University of the Free State. Faculty of Health Sciences rule book. School of Medicine. Undergraduate qualifications 2019. https://apps.ufs.ac.za/dl/yearbooks/335_yearbook_eng.pdf (accessed 12 September 2019).
28. Xu S, Lorber MF. Interrater agreement statistics with skewed data: Evaluation of alternatives to Cohen's kappa. J Consult Clin Psychol 2014;82(6):1219.
29. Holley JW, Guilford JP. A note on the G index of agreement. EPM 1964;24:749-753. https://doi.org/10.1177/001316446402400402
30. Opposs D, He O. The reliability programme. Final report. London: Ofqual, 2011. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/578899/2011-03-16-the-reliability-programme-final-report.pdf (accessed 12 September 2019).
31. McHugh ML. Interrater reliability: The kappa statistic. Biochem Medica 2012;22(3):276-282.
32. Sim J, Wright CC. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. Phys Ther 2005;85(3):257-268. https://doi.org/10.1093/ptj/85.3.257
33. Attali Y. Effort in low-stakes assessments: What does it take to perform as well as in a high-stakes setting? Educ Psychol Meas 2016;76(6):1045-1058.
34. Beilock SL, Carr TH. On the fragility of skilled performance: What governs choking under pressure? J Exp Psychol Gen 2001;130:701-725.
35. Sandi C. Stress and cognition. WIREs Cogn Sci 2013;4(3):245-261. https://doi.org/10.1002/wcs.1222
36. Patrício MF, Julião M, Fareleira F, Carneiro AV. Is the OSCE a feasible tool to assess competencies in undergraduate medical education? Med Teach 2013;35(6):503-514. https://doi.org/10.3109/0142159X.2013.774330
37. Brits H, Bezuidenhout J, van der Merwe LJ. A framework to benchmark the quality of clinical assessment in a South African undergraduate medical programme. S Afr Fam Pract 2020;62(1):a5030. https://doi.org/10.4102/safp.v62i1.5030
38. Khan KZ, Gaunt K, Ramachandran S, Pushkar P. The objective structured clinical examination (OSCE): AMEE Guide No. 81. Part II: Organisation and administration. Med Teach 2013;35:e1447-1463. https://doi.org/10.3109/0142159X.2013.818635
39. Ponnamperuma GG, Karunathilake IM, McAleer S, Davis MH. The long case and its modifications: A literature review. Med Educ 2009;43:936-941. https://doi.org/10.1111/j.1365-2923.2009.03448.x
40. Kamarudin MA, Mohamad N, Awang MN, Siraj BHH, Yaman MN. The relationship between modified long case and objective structured clinical examination (OSCE) in final professional examination 2011 held in UKM Medical Centre. Procedia Soc Behav Sci 2012;60:241-248. https://doi.org/10.1016/j.sbspro.2012.09.374
41. Wass V, Jolly B. Does observation add to the validity of the long case? Med Educ 2001;35(8):729-734. https://doi.org/10.1046/j.1365-2923.2001.01012.x
42. Wass V, Jones R, Van der Vleuten C. Standardised or real patients to test clinical competence? The long case revisited. Med Educ 2001;35(4):321-325. https://doi.org/10.1046/j.1365-2923.2001.00928.x
43. Norman G. Postgraduate assessment – reliability and validity. Trans Coll Med S Afr 2003;47:71-75.
44. Liu C. An introduction to workplace-based assessments. Gastroenterol Hepatol Bed Bench 2012;5(1):24-28.
45. Clauser BE, Margolis MJ, Swanson DB. Issues of validity and reliability for assessments in medical education. In: Holmboe ES, Durning SJ, Hawkins RE (editors). Practical Guide to the Evaluation of Clinical Competence. 2nd ed. Philadelphia: Elsevier, 2018.

Accepted 3 December 2020.