Development of a baseline assessment tool to establish students' foundational knowledge of life sciences at entry to university

L Pienaar,[1] MSc; **R Prince,**[2] MSc; **A Abrahams,**[3] PhD, PGDip (Health Professional Education)

[1] *Department of Health Sciences Education, Faculty of Health Sciences, University of Cape Town, South Africa*
[2] *Centre for Educational Testing for Access and Placement, Centre for Higher Education Development, University of Cape Town, South Africa*
[3] *Department of Human Biology, Faculty of Health Sciences, University of Cape Town, South Africa*

Corresponding author: *A Abrahams (amaal.abrahams@uct.ac.za)*

Background. Universities in South Africa use the Grade 12 school-leaving examinations to measure whether students have the knowledge and skills needed to enter tertiary-level education. However, there is much discussion on the effectiveness of these assessments to measure the preparedness of students for their first year at university. To facilitate the appropriate teaching and learning of anatomy and physiology, there is a need to assess students' baseline knowledge of life sciences at entry to their first year at university.
Objectives. To develop and refine an anatomy and physiology foundational knowledge assessment (A&P foundational knowledge assessment), which looks back to the content of the Grade 12 life sciences curriculum and forward to the first-year anatomy and physiology curricula.
Methods. Three hundred and seventy-one first-year students (occupational therapy, physiotherapy and MB ChB) wrote the A&P foundational knowledge assessment. Classic item and test analysis was done using Iteman 4.3 software (Assessment Systems Corp., USA).
Results. The Kuder-Richardson formula 20 (KR-20) reliability score, which ranges from 0 to 1, was 0.64 for all the students. For MB ChB students, the KR-20 value was lower (0.57) compared with that for occupational therapy and physiotherapy students (0.66). The KR-20 scores for the 21 physiology and 16 anatomy items were 0.48 and 0.57, respectively. A KR-20 score of >0.50 is considered acceptable. The mean difficulty index (range 0 - 1) for physiology was 0.60, and the mean discrimination index was 0.15. For anatomy, the mean item difficulty index was 0.57 and mean discrimination index was 0.21.
Conclusion. Based on the acceptable reliability value, the assessment was shown to be an effective instrument to measure students' foundational knowledge in human anatomy and physiology, which is part of life sciences.

Afr J Health Professions Educ 2021;13(1):77-82. https://doi.org/10.7196/AJHPE.2021.v13i1.1226

Morrow[1] distinguishes between gaining access to an institution of learning and access to knowledge. To facilitate the latter, one needs to ascertain what students know about a subject to inform teaching to scaffold the learning process. The development of valid and reliable instruments is essential to guide and thus enhance teaching and learning activities. Moreover, in the absence of pass or fail decisions associated with summative assessments, a low stakes baseline or formative assessment provides a less intimidating experience and an ideal starting point to establish what students know. Importantly, these baseline assessments offer an opportunity to elucidate cognitive and content knowledge gaps that may exist.[2,3] This is valuable information, as students from diverse educational and socioeconomic backgrounds may be provided with access to an institution, but may not have access to the subject knowledge that is required to succeed. In the South African (SA) context, there are a few tailor-made test instruments that explore students' preparation for university. The Health Sciences Placement Tests developed by the Alternative Admissions Research Project (AARP), currently the Centre for Educational Testing for Access and Placement (CETAP), were used by Wadee and Cliff[4] to predict academic success during the first year of study. The individual component scores of the test were aligned with biology, chemistry, physics, sociology, psychology and fundamentals of medical and clinical sciences to identify specific predictor domains of student success in a medical degree. Allers *et al.*[5] used medical students' National Senior Certificate (NSC) and National Benchmark Test (NBT) results to compile profiles of successful students and those who failed physiology, and to identify predictors for success in physiology. An earlier study at the University of Cape Town (UCT) focused on developing a mathematical literacy questionnaire to identify which students entering the MB ChB programme were in need of extra mathematical literacy interventions. The diagnostic tool showed a moderate improvement post intervention in the mathematical literacy of these students.[6] Potgieter and Davidowitz[7] developed the Chemistry Competence Test to probe the school-university interface in SA for the level of conceptual understanding of chemistry. Their study is similar to ours; however, none of these instruments addresses constructs obtained from Grade 12 life sciences that are aligned to topics in the first-year anatomy and physiology curricula. In SA, the school-leaving (Grade 12) examinations, administered by the Department of Basic Education, measure whether students have acquired the knowledge and skills needed to exit the school system and enter tertiary-level education. Some students admitted to the first-year MB ChB, physiotherapy and occupational therapy programmes may not have studied and completed life sciences as a Grade 12 subject, but rather mathematics and physical science, as these are entrance requirement subjects. Even for students who have studied life sciences, given the variation in curriculum design and pedagogies employed by the various public and private high schools in SA, the Grade 12 marks for life sciences do not give a dependable indication of a student's foundational knowledge for first-year health professional courses. There are often huge variations in students' cognitive, problem-solving and reasoning skills[8] that emerge

from the different schooling experiences in this country. SA studies have also shown that in the physical sciences, students with poor cognitive and problem-solving skills often struggle in their first year at university.[9,10] Poor performance has been linked to the lack of dialogical discourse, rote learning and failure of teachers to ensure that students actively engage with the content.[11] Students in the health sciences are taught anatomy and physiology in their first year, with the assumption that they have acquired the prerequisite foundational knowledge and cognitive skills from their Grade 10 - 12 curricula. Currently, when students enter their first year at university, there is no assessment to measure their level of knowledge and to identify those who may benefit from early academic support in anatomy and physiology. Consequently, students who fail the anatomy and physiology courses must repeat these, or are placed in an extended degree programme.[12] In our study, we describe the development and refinement of a novel baseline assessment, which aligns content of the Grade 12 life sciences curricula with content that is relevant to the first-year anatomy and physiology curricula.

Methods

Study design

The study describes the development of a baseline assessment to determine students' foundational knowledge of human anatomy and physiology on entry to university. Psychometric analysis was performed to measure the reliability of the test instrument.

Characteristics of the study population (sampling) and procedure

A total of 371 students admitted into the first-year MB ChB, physiotherapy and occupational therapy programmes completed the multiple-choice assessment. These undergraduate programmes in the health sciences faculty offer anatomy and physiology courses during the first semester of the first year. Whereas the admission requirements for the first year at university vary, it is expected that all students should have mathematics and physical sciences or life sciences to be eligible for acceptance into the different programmes. At the higher education institution where the study was conducted, the minimum entrance requirements for first-year MB ChB students is at least 60% for mathematics, physical sciences and English and at least 50% for the 3 next best subjects. The entrance requirements for the first year in the Division of Physiotherapy are that students obtain at least 50% for all subjects, which must include mathematics and physical sciences or life sciences. For the Division of Occupational Therapy, students are required to have obtained at least 50% for all subjects, which must include physical sciences or life sciences and mathematics (or 60% for mathematical literacy). All students wrote the same NSC examination, and therefore were expected to have acquired the foundational knowledge and cognitive skills in the subjects. A 37-item 4-option (single best answer) multiple-choice assessment was developed and administered during the first week of lectures on the university's online learning system. The responses of participants (including the questions) were exported into an Excel spreadsheet (Microsoft Corp., USA) for further analysis by a test development co-ordinator. Participants' names and student numbers were anonymised by the primary investigator (AA). Only students who gave consent and were ≥18 years old at the time, were eligible participants.

Development of the data collection instrument

The pool of multiple-choice assessment items was developed by the principal investigator and co-investigators to explore students' cognitive skills and conceptual knowledge in anatomy and physiology. The content for the selected items was based on the content of the 2012 - 2015 Grade 12 final life sciences examination papers and relates to the curriculum covered in the first-year anatomy and physiology courses. The items were evaluated by 2 disciplinary experts to ensure adequate coverage of the domain and field of anatomy and physiology. The researchers carefully selected the distractors for each item, based on content and concepts that past students found challenging. This was done to give researchers a better understanding of students' knowledge, misconceptions and reasoning abilities at the start of the academic year. Each of the items for the anatomy and physiology foundational knowledge assessment (A&P foundational knowledge assessment) was rated according to the cognitive domains of Bloom's taxonomy. This taxonomy is frequently[13-16] used to classify the cognition required in multiple-choice items. The items were designed with specific focus on Bloom's taxonomy categories of knowledge and comprehension, as these were thought to be the most appropriate at an entry level to first-year anatomy and physiology.[17] For each of the items, the knowledge domain and comprehension required to solve the problem were analysed and documented in a specification table (Table 1) before administering the test. This helped to shed light on the students' level of understanding of the concepts and sub-concepts of the content tested. Topics included: homeostasis, anatomical terminology, levels of organisation of the human body, body systems, the endocrine system, development and inheritance. In 2016, the assessment was piloted using data from 134 students in the first-year anatomy and physiology courses to allow for revision and refinement.

Table 1. Classification of each item in each domain assessed into the cognitive levels of Bloom's taxonomy

Bloom's taxonomy categories	Items in each domain, *n*	Medical students, mean (SEM)	Occupational therapy and physiotherapy students, mean (SEM)	*p*-value
Level 1: knowledge		26.59 (4.20)	21.18 (4.66)	<0.0001
Physiology	10			
Anatomy	11			
Level 2: cognitive skills (comprehension)		26.43 (4.29)	21.13 (4.66)	<0.0001
Physiology	11			
Anatomy	5			

SEM = standard error of the measurement.

Analysis

Item difficulty and item discrimination were calculated for each item of the assessment using Iteman version 4.3 software (Assessment Systems Corp., USA).[18] The classic test theory (CTT) analysis[19] allowed insight into the reliability of the overall test and of the anatomy and physiology domains independently.

Individual questions were analysed according to: (*i*) reliability of the test without the item; (*ii*) item difficulty; (*iii*) discrimination indices; and (*iv*) correlation indices. The total scores, standard deviations, standard error of the measurement (SEM) and distribution scores were also calculated using the CTT Iteman software. Analysis was performed on the whole group and by splitting the students into 2 groups – the MB ChB students in one group and the combined physiotherapy and occupational therapy students in another group. The reason for splitting the students into the 2 groups is that, in the health sciences faculty, not only are the admission requirements similar for the occupational therapy and physiotherapy students, but these students register for the same anatomy and physiology course in their first year. Data collected were coded by a research assistant. The coded data were recorded in an Excel program (Microsoft Corp., USA) and later transferred to SPSS version 25 (IBM Corp., USA). In addition to the abovementioned analyses, the mean and distribution scores of the 2 groups were compared with each other.

Ethical approval

The proposal received ethical approval from the Human Research Ethics Committee of the Faculty of Health Sciences at the University of Cape Town (ref. no. HREC 7982016).

Results

Overall reliability, item difficulty and discrimination performance of the baseline assessment

Analyses were performed on the test as a whole and on the 37 items consisting of 2 domains: anatomy and physiology. For the 2 student cohorts, using the Kuder-Richardson formula 20 (KR-20), the test had an overall reliability of 0.64 (alpha), with the SEM at 2.53. As an assessment, the instrument demonstrated an acceptable level of reliability, given that it was a low-stakes assessment with only 37 items. The small SEM shows that the observed scores were closely distributed around a student's 'true' score. We then evaluated the performance for each of the items by analysing the point-biserial correlation discrimination indices (*rpbis*). This was done to ascertain how well the item differentiated between low- and high-scoring students and how easy or challenging each item was, expressed as an item difficulty index (*p*-value). Items were deemed difficult if the index was ≤0.25 and easy if it was ≥0.95. The discrimination index ranges from −1 to +1. Here, negative scoring items are considered to have poor discrimination, given that low-scoring students were more likely to choose the correct option than high-scoring students. A discrimination index of ≥0.2 was considered to offer the best discrimination.[20-22] Overall, the mean score for all the test items was 21.73, with 59% of students answering the items correctly (mean *p*=0.59). Of the 37 items, 6 items were flagged as not discriminating well, with 2 of the 6 items having negative discrimination and 2 of the 6 items having a *p*-value of 0.1 - 0.2.

Overall, item difficulty was reasonable, although the item discrimination index of all 37 items would need to be reviewed, as the mean *rpbis* of 0.18 was lower than expected (Table 2).

Reliability, item difficulty and discrimination performance in the anatomy and physiology domains

The physiology and anatomy domains consisted of 21 and 16 items, respectively. Individually, the domains demonstrated lower reliability than the overall assessment, with physiology KR-20 at 0.48 and anatomy KR-20 at 0.57. Given the small number of items in each domain, it is not surprising that the reliability tended to be on the low end of acceptable, as reliability is a function of the number of items in a test. The mean item difficulty index (*p*-value) for physiology was 0.60, with a lower mean discrimination index of 0.15, whereas anatomy had a mean item difficulty index (*p*-value) of 0.57 and a marginally higher mean discrimination index of 0.21. Overall, some of the items in the physiology domain failed to discriminate the high-scoring students from those who achieved lower scores. The items that failed to discriminate would need to be reviewed; for anatomy the item discrimination was reasonable. The anatomy items were only slightly more challenging, with students achieving a correct score of 56.6% compared with physiology at 60.4% (Table 2). The combined scores for all the items and the distribution of raw scores for physiology and anatomy, respectively, were normally distributed. The two domains had a low correlation of 0.37, indicating that what was being tested was distinct (Fig. 1).

Comparison of test scores for medical, physiotherapy and occupational therapy students

In addition to establishing the overall reliability of the assessment, we were interested in determining whether the occupational therapy, physiotherapy and medical students performed similarly or differently in the baseline assessment. The baseline assessment scores of the medical students were slightly higher than those of the occupational therapy and physiotherapy students. The mean test scores for the medical students for the anatomy and physiology test was 26.03. Of the 227 medical students, an average of 59% answered the items correctly. In comparison, the occupational therapy and physiotherapy students (*n*=144) had a mean score of 20.06 for the anatomy and physiology test, where 53% of the students answered the items correctly. We then compared the students' results on the baseline assessment with those of their final Grade 12 life sciences results for both medical, occupational therapy and physiotherapy students. Of the 371 registered first-year students, the

Table 2. Item difficulty and discrimination for all scored items in each domain

Domain	Items, *n*	Mean (SD)	Minimum score	Maximum score	Mean *p*-value	Mean *rpbis*
Items scored, *N*	37	21.73 (4.25)	10	34	0.59	0.18
Physiology	21	12.68 (2.69)	5	19	0.60	0.15
Anatomy	16	9.05 (2.45)	0	16	0.57	0.21

SD = standard deviation; *rpbis* = point-biserial correlation discrimination indices.

majority had studied life sciences (n=356) (medical, n=217; physiotherapy and occupational therapy, n=139). Fewer than 10% of students reported not having studied life sciences (medical, n=11; physiotherapy and occupational therapy, n=4). Medical students entered university with a slightly higher final school grade for life sciences than physiotherapy and occupational therapy students. The mean life sciences grade for medical students was 88% (n=217) (range 74 - 100%). The physiotherapy and occupational therapy grade was 71.50% (n=139) (range 47 - 94%). The results of the baseline assessment showed a similar trend as for the life sciences grades, as medical students scored slightly higher than occupational therapy and physiotherapy students for anatomy and physiology. When we compared the performance of students based on Bloom's taxonomy in knowledge and comprehension (Table 1), the medical students scored significantly higher in both cognitive categories than occupational therapy and physiotherapy students. This aspect of the research will be explored in greater detail in another study, focused on the understanding of the life sciences conceptual knowledge and problem-solving skills students possess at the start of the academic year.

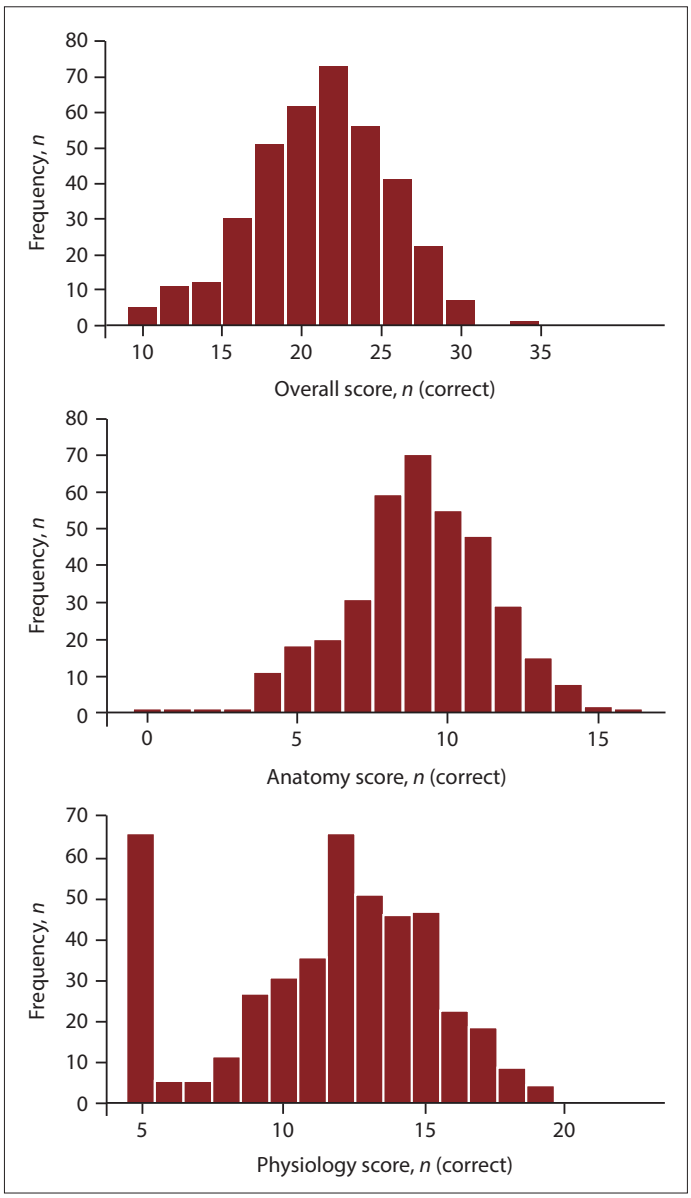

Fig. 1. Distribution of the raw scores across combined and individual domains (anatomy and physiology).

Reliability of the test for the 2 student groups

The test had a lower reliability for the medical students (0.57), with a larger SEM (2.78) than for the occupational therapy and physiotherapy students, where the test reliability score was (0.66), with an SEM of 2.66. When calculating the reliability of the individual domains of anatomy and physiology for each discipline, the test achieved lower reliability. In physiotherapy and occupational therapy, the anatomy and physiology domains achieved a reliability of 0.60 and 0.42, respectively, whereas for medical students a slightly lower reliability was achieved for anatomy (0.52); physiology was higher (0.43).

Discussion

This study describes the development and refinement of an assessment instrument that aims to establish students' foundational knowledge of human anatomy and physiology learnt in Grade 12 life sciences. The benefits of such an assessment can be to inculcate content expertise, alter attitudes, promote student growth and offer an opportunity to receive feedback from peers and lecturers.[23] Moreover, the assessment can also provide an opportunity for lecturers to understand students' prior knowledge[24,25] and to use this information to inform teaching.[26,27] Establishing the knowledge that students bring into learning spaces is less well explained in the literature.[28] The baseline assessment administered to first-year health sciences students shows potential as an efficient and acceptable method to establish students' prior knowledge. Ideally, assessments that inform pass or fail decisions, such as summative examinations, should have a reliability coefficient of >0.8.[29] This is because the consequences of these decisions have an impact on the students' future, whereas the aim of the baseline assessment was to inform teaching and learning. In our study, the assessment achieved a modest reliability of 0.64, which is lower than assessments that examine student preparedness.[4,7] It can be attributed to the fact that we purposely selected key themes that served as indicators of entry-level foundational knowledge. This resulted in a limited number of items generated, and the associated shorter length of test administration most likely played a part in the lower reliability that is usually associated with formative assessments (0.70 - 0.79).[30] Even in light of the reliability score, this instrument provides a way to gather information to benefit teaching and learning, as it looks back to the content of the Grade 12 life sciences curricula and forward to first-year anatomy and physiology curricula. In this way it links prior knowledge with knowledge of the forthcoming subject. Using this test, we were able to determine that ~60% of students had basic knowledge to build on as they entered university. Previous studies have shown that students who performed well in secondary school subjects were more likely to perform well at university.[31,32] Medical students who entered with higher Grade 12 life sciences grades performed better in the baseline assessment. The multiple-choice baseline assessment performed differently across the 2 groups, with lower reliability achieved in the medical students' group. Reliability can be affected by the formulation of the items, such as a mistake on the correct scoring key. Other factors that can affect reliability are poorly prepared students guessing correctly and well-prepared students somehow justifying the wrong answer.[33,34] In the assessment, only 6 items were found to be problematic. The majority (57%) of the items were developed at the knowledge level, which may have

contributed further to the difference in performance between the 2 groups. For the continued use of the baseline assessment, the items must be reviewed and adapted to accommodate the range of student groups being tested, as well as increasing both the number of questions and length of the test.[35] Indirectly, this baseline assessment also provided an opportunity for students to access and retrieve information that was learnt previously, which may be used in the future.[36,37] This retrieval facilitates learning even if the attempt is unsuccessful,[36] as it shows students where their knowledge gaps are. In this way, students can adjust their learning strategies. As can be observed from the test scores, students possess a reasonable level of understanding of the subject. Where it is shown that students do not possess the required content knowledge early in the academic programme, it offers an opportunity to effect changes to teaching, such as academic support interventions.[38] In our study, students were offered additional support alongside the course when their test score was ≤55%. (The results of this study will be reported in a future publication.) Instituting these types of interventions is especially important where students who work with patients need to use their science knowledge to inform clinical decisions. Using a less intimidating low-stakes baseline assessment may clarify the concepts that link prior knowledge to current experiences.

Lessons learnt

A range of criteria exists that supports the development of rigorous assessment instruments that can withstand internal and external scrutiny. We set out to develop a reliable instrument, while giving attention to validity and the educational benefit.[39] During the development process, we were mindful to ensure that the instrument assesses students' knowledge at the appropriate cognitive level, using the relevant content. The results indicate that additional work is required to improve the quality of the items to consistently achieve accurate results. Alongside the criteria mentioned in developing the instrument, we also paid attention to fairness and feasibility.[40] Using a method of assessment that students have some familiarity with from high school is helpful, especially if the assessment is to be taken early in the year. The items were reviewed for content validity and language by lecturers associated with the programmes and educationalists to ensure that English additional language speakers were not negatively affected. A baseline assessment used to inform teaching can be challenging to administer early in the course, where class schedules are set in advance. In our faculty, timetable planning in some instances is done months or a year in advance, which limits the freedom of introducing assessment for learning opportunities. Within a structured curriculum, the emphasis is often on gaining content knowledge and developing appropriate formative and summative assessments, which dictate the timetable. The success of developing baseline assessments to inform teaching and learning not only requires the buy-in from relevant stakeholders, but should be core to the design of a first-year curriculum.

Conclusion

Establishing baseline assessments clarifies assumptions regarding the knowledge of students when starting university study. It enables lecturers to create scaffolded learning opportunities to bridge the gaps in knowledge, and in doing so helps to facilitate access to subject knowledge. While we recognise that to establish statistical levels of reliability, repeated assessment opportunities are needed until an optimal level of reliability is achieved, this may not always be required if a low-stakes assessment directed at learning is developed.

Declaration. None.

Acknowledgements. We thank Ms Stevie Biffen, a research assistant at UCT, for her contribution to data capturing, and Prof. Shirley Pendlebury and Ms Vera Frith, UCT, for their helpful comments on the manuscript.

Author contributions. LP contributed to the design of the research, analysis and writing of the manuscript. RP analysed the data and contributed to writing of the manuscript. AA conceptualised, implemented and contributed to analysis of the results and to writing of the manuscript.

Funding. This work was supported by the UCT teaching grant and the Human Biology Emerging Researcher award.

Conflicts of interest. None.

1. Morrow W. Epistemological access in the university. AD Issues 1993;1(1):3-4.
2. Decristan J, Klieme E, Kunter M, et al. Embedded formative assessment and classroom process quality: How do they interact in promoting science understanding? Am Educ Res J 2015;52(6):1133-1159. https://doi.org/10.3102/0002831215596412
3. Weurlander M, Söderberg M, Scheja M, Hult H, Wernerson A. Exploring formative assessment as a tool for learning: Students' experiences of different methods of formative assessment. Assess Eval Higher Educ 2012;37(6):747-760. https://doi.org/10.1080/02602938.2011.572153
4. Wadee AA, Cliff A. Pre-admission tests of learning potential as predictors of academic success of first-year medical students. S Afr J Higher Educ 2016;30(2):264-278. https://doi.org/10.20853/30-2-619
5. Allers NJ, Hay L, Janse van Rensburg RC. Preliminary study: Predictors for success in an important premedical subject at a South African medical school. Afr J Health Professions Educ 2016;8(1):81-83. https://doi.org/10.7196/ajhpe.2016.v8i1.647
6. Prince R, Frith V, Jaftha J. Mathematical Literacy of Students in First Year of Medical School at a South African University. Proceedings of the 12th Annual Conference of the Southern African Association for Research in Mathematics and Science Education, Cape Town. Cape Town: UCT, 2004. http://hdl.handle.net/11427/27581 (accessed 25 January 2021).
7. Potgieter M, Davidowitz B. Grade 12 achievement rating scales in the new National Senior Certificate as indication of preparedness for tertiary chemistry. S Afr J Chem 2010;12:75-82.
8. Hartman N, Kathard H, Perez G, et al. Health sciences undergraduate education at the University of Cape Town: A story of transformation. S Afr Med J 2012;102(6):477-480.
9. Selvaratnam M. Chemistry students' competence throughout their BSc course in some problem-solving strategies. S Afr J Chem 2011;64(1):44-48.
10. Tigere E. Investigating the problem solving skills proficiency of Grade 12 physical science learners in Highveld Ridge East and West circuits when solving stoichiometry problems. MSc thesis. Pretoria: University of South Africa, 2014.
11. Van Schalkwyk S, Bitzer E, van der Walt C. Acquiring academic literacy: A case of first-year extended degree programme students. South Afr Linguist Applied Language Studies 2009;27(2):189-201. https://doi.org/10.2989/salals.2009.27.2.6.869
12. Alexander R, Badenhorst E, Gibbs T. Intervention programme: A supported learning programme for educationally disadvantaged students. Med Teach 2005;27(1):66-70. https://doi.org/10.1080/01421590400016472
13. Crowe A, Dirks C, Wenderoth MP. Biology in bloom: Implementing Bloom's taxonomy to enhance student learning in biology. CBE-Life Sciences Educ 2008;7(4):368-381. https://doi.org/10.1187/cbe.08-05-0024
14. Kim MK, Patel RA, Uchizono JA, Beck L. Incorporation of Bloom's taxonomy into multiple-choice examination questions for a pharmacotherapeutics course. Am J Pharm Educ 2012:76(6):1-8. https://doi.org/10.5688/ajpe766114
15. Noble T. Integrating the revised Bloom's taxonomy with multiple intelligences: A planning tool for curriculum differentiation. Teach College Rec 2004;106(1):193-211. https://doi.org/10.1111/j.1467-9620.2004.00328.x
16. Palmer EJ, Devitt PG. Assessment of higher order cognitive skills in undergraduate education: Modified essay or multiple choice questions? BMC Med Educ 2007;7(1):49. https://doi.org/10.1186/1472-6920-7-49
17. Krathwohl DR. A revision of Bloom's taxonomy: An overview. Theory Pract 2002;41(4):212-218. https://doi.org/10.1207/s15430421tip4104_2
18. Guyer R, Thompson NA. User's Manual for Iteman 4.3. Woodbury, MN: Assessment Systems Corporation, 2013.
19. Hambleton RK, Jones RW. Comparison of classical test theory and item response theory and their applications to test development. Educ Measurement: Issues Pract 1993;12(3):38-47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x
20. Thompson N. Classical Item and Test Analysis with CITAS. Woodbury, MN: Assessment Systems Corporation, 2009.
21. DiBattista D, Kurzawa L. Examination of the quality of multiple-choice items on classroom test. Can J Scholar Teach Learn 2011;2(2). https://doi.org/10.5206/cjsotl-rcacea.2011.2.4
22. Downing SM, Yudkowsky R. Assessment in Health Professions Education. New York: Routledge, 2009. https://doi.org/10.4324/9780203880135
23. Gibbs G, Simpson C, Gravestock P, Hills M. Conditions under which assessment supports students' learning. Learn Teach Higher Educ 2004;1:3-31.
24. Edens K, Shields C. A Vygotskian approach to promote and formatively assess academic concept learning. Assess Eval Higher Educ 2015;40(7):928-942. https://doi.org/10.1080/02602938.2014.957643
25. Gultice A, Witham A, Kallmeyer R. Are your students ready for anatomy and physiology? Developing tools to identify students at risk for failure. Adv Physiol Educ 2015;39(2):108-115. https://doi.org/10.1152/advan.00112.2014
26. Rushton A. Formative assessment: A key to deep learning? Med Teach 2005;27(6):509-513. https://doi.org/10.1080/01421590500129159
27. Silverthorn DU, Thorn PM, Svinicki MD. It's difficult to change the way we teach: Lessons from the integrative themes in physiology curriculum module project. Adv Phys Educ 2006;30(4):204-214. https://doi.org/10.1152/advan.00064.2006
28. Kulasegaram K, Rangachari PK. Beyond 'formative': Assessments to enrich student learning. Adv Phys Educ 2018;42(1):5-14. https://doi.org/10.1152/advan.00122.2017
29. Kibble JD. Best practices in summative assessment. Adv Phys Educ 2017;41(1):110-119. https://doi.org/10.1152/advan.00116.2016
30. Downing SM. Reliability: On the reproducibility of assessment data. Med Educ 2004;38(9):1006-1012. https://doi.org/10.1111/j.1365-2929.2004.01932.x
31. Green R, Brown E, Ward A. Secondary school science predictors of academic performance in university bioscience subjects. Anat Sci Educ 2009;2(3):113-118. https://doi.org/10.1002/ase.82

32. McKenzie K, Schweitzer R. Who succeeds at university? Factors predicting academic performance in first year Australian university students. Higher Educ Res Dev 2001;20(1):21-33. https://doi.org/10.1080/07924360120043621
33. Collins J. Writing multiple-choice questions for continuing medical education activities and self-assessment modules. Radiographics 2006;26(2):543-551. https://doi.org/10.1080/07924360120043621
34. Williams JB. Assertion reason multiple choice testing as a tool for deep learning: A qualitative analysis. Assess Eval Higher Educ 2006;31(3):287- 301. https://doi.org/10.1080/02602930500352857
35. Wolf R. The Validity and Reliability of Outcome Measure. Monitoring the Standards of Education. Oxford: Pergamon, 1994:121-132.
36. Kornell N, Vaughn KE. How retrieval attempts affect learning: A review and synthesis. In: Ross BH, ed. Psychology of Learning and Motivation. Cambridge, MS: Academic Press, 2016:183-215. https://doi.org/10.1016/bs.plm.2016.03.003
37. Logan JM, Thompson AJ, Marshak DW. Testing to enhance retention in human anatomy. Anatom Sci Educ 2011;4(5):243-248. https://doi.org/10.1002/ase.250
38. Mattheis A, Jensen M. Fostering improved anatomy and physiology instructor pedagogy. Adv Phys Educ 2014;38(4):321-329. https://doi.org/10.1152/advan.00061.2014
39. Schuwirth LWT, van der Vleuten CPM. Changing education, changing assessment, changing research? Med Educ 2004;38(8):805-812. https://doi.org/10.1111/j.1365-2929.2004.01851.x
40. Van der Vleuten CPM. The assessment of professional competence: Developments, research and practical implications. Adv Health Sci Educ 1996;1(1):41-67. https://doi.org/10.1007/BF00596229

Accepted 16 January 2020.