DOI: https://doi.org/10.4038/seajme.v19i1.651 **Original Research Paper**

Improving High-Stakes Obstetrics & Gynaecology Exams: A 3-year Analysis of SBAs

Phyo, Z., Soe, T.T., Aung, Y.P., Naing, T.T., Hein, T., Aung, D.

Abstract

Introduction: This study evaluates the quality of Single Best Answers (SBAs) in the summative assessments of Year 5 medical students at Defense Service Medical Academy (DSMA) in Yangon, Myanmar. It examines the Difficulty and Discrimination Indices of SBAs in the Obstetrics and Gynecology curriculum across three cohorts (2022–2024), aiming to improve assessment quality through insights into refining the SBAs and informing broader quality assurance strategies.

Methods: A quantitative analysis was conducted, focusing on 120 SBAs drawn from the Obstetrics and Gynecology curriculum across three cohorts. Statistical methods, including item analysis, Cronbach's Alpha for internal consistency, and cohort performance comparison, were used to assess the psychometric properties of the SBAs.

Results: The Difficulty Index showed that most questions were classified as "easy," while there were notable variations in the Discrimination Index across cohorts, suggesting that some questions did not effectively differentiate between high and low-performing students. Although the exams maintained high internal reliability (Cronbach's Alpha 0.78–0.82), the weak correlation between the difficulty and discrimination indices highlighted the need for improvements in question formulation.

Conclusion: The findings suggest that the exit exam for the medical program should include a higher proportion of average and difficult-level questions. Continuous refinement of SBAs, informed by item analysis and integrated within a comprehensive quality assurance framework that includes robust exam blueprinting, item writing and review protocols, and systematic item banking, is needed to improve difficulty levels and balance discriminatory power. By addressing these issues, DSMA can enhance its assessments, ensuring they more accurately reflect student performance, align with outcome-based medical education, and better prepare students for their future roles in healthcare.

Keywords: Single Best Answer, Medical Education, Difficulty Index, Discrimination Index, Item Analysis

Introduction

Assessment is a critical element in healthcare education, as it evaluates both theoretical knowledge and practical abilities, decision-making, and professional conduct.

Medical Education Department, Defence Service Medical Academy (DSMA), Yangon, Myanmar

Corresponding author: Dr. Zaw Phyo
Email: dr.zawphyoo@gmail.com

In medical education, assessments are essential for determining competency in key areas such as clinical knowledge, patient communication, problem-solving, and ethical decision-making (Boursicot *et al*., 2011). As medical programs evolve, particularly with the shift to integrated, outcome-based curricula, the importance of effective assessment methods has increased.

Single Best Answers (SBAs) are commonly used in medical education due to their efficiency in assessing large volumes of content in a structured, standardized manner (Babo *et al*., 2020). However, while they offer advantages like ease of administration and objectivity, their design must be methodologically sound to accurately measure student competencies. Poorly designed SBAs can lead to biased results and misinterpretation of student abilities, undermining the integrity of the examination system (McTighe & Wiggins, 2013). Ensuring the quality of SBAs through item analysis is crucial for improving assessment reliability and supporting fair evaluations.

This study, carried out at the Defence Services Medical Academy (DSMA) in Yangon, Myanmar, focuses on item analysis as a critical component of assessment evaluation. It is acknowledged that a comprehensive quality assurance framework in medical assessment also involves robust upstream processes. These include exam blueprinting to ensure alignment with curriculum objectives, adherence to standardized item writing guidelines, rigorous peer review of questions before deployment, and the systematic use of an item bank for managing and tracking assessment materials (Downing, 2006; Case & Swanson, 2002). The findings from item analysis, as presented in this paper, are intended to feed into and strengthen these broader institutional quality improvement efforts at DSMA.

SBAs typically consist of a stem, followed by multiple options, one of which is correct. The distractors are designed to challenge students who may not have mastered the material. Well-constructed SBAs can assess a broad range of cognitive skills, from recalling facts to applying knowledge in complex clinical situations (Kumar *et al*., 2021). This approach aligns with Bloom's Taxonomy, which categorizes cognitive skills into knowledge, comprehension, application, analysis, synthesis, and evaluation (Verenna *et al*., 2018). Therefore, well-designed SBAs are essential for assessing both basic knowledge and higher-order cognitive skills needed in clinical practice.

The quality of SBAs is vital for ensuring valid and reliable assessments. Poorly constructed questions—whether too easy, too difficult, or unclear—can produce invalid results that do not reflect students' true knowledge and abilities (Dennick *et al*., 2009). Thus, a systematic evaluation of each question through item analysis is important (Hartati & Yogi, 2019; Siri & Freddano, 2011). This analysis examines key parameters such as the Difficulty Index (P) and Discrimination Index (D), which provide insights into the effectiveness of individual items.

The Difficulty Index (P) measures how challenging a question is for a group of students. A high P indicates that most students answer the question correctly, while a low P suggests that the question is more difficult for the majority (Hingorjo & Jaleel, 2012). The Discrimination Index (D) assesses a question's ability to distinguish between higher- and lower-performing students. A high D ensures the test accurately reflects student proficiency (Olipas & Luciano, 2024).

At the Defence Services Medical Academy (DSMA) in Yangon, Myanmar, the shift from a traditional to an outcome-based integrated curriculum emphasizes the need for high-quality assessment tools like SBAs. The new curriculum integrates multiple disciplines, and assessments at the end of each academic year are critical for measuring students' competencies across various medical domains (Brauer & Ferguson, 2015). Regular item analysis helps ensure that SBAs align with learning objectives and effectively measure desired competencies. Detailed item analysis can also uncover biases in question design, such as poorly constructed distractors, which can compromise the reliability and validity of SBAs (Bone & Prosser, 2020). Addressing these issues is essential to ensuring accurate and fair assessments.

The study aims to evaluate the quality of SBAs used in the summative assessments for Year 5 medical students at DSMA, focusing on the Difficulty Index (P) and Discrimination Index (D). By conducting a detailed item analysis of 120 MCQs used in written examination papers of three cohorts, the study seeks to identify questions that require revision for improved accuracy and reliability.

Methods

Study Design

This research employed a retrospective analytical approach to assess the psychometric properties of SBAs utilized in high-stakes examinations for Year 5 medical students at the Defence Services Medical Academy (DSMA), Yangon, Myanmar. The study recruited three cohorts of students: 2022 (n=39), 2023 (n=34), and 2024 (n=33), each of whom participated in a standardized examination consisting of 120 SBAs from the Obstetrics and Gynecology (OG) curriculum. The primary objective was to evaluate the quality of these SBAs using statistical methods, specifically analyzing their difficulty and discriminatory power.

During this 3 year analysis, the same set of 120 SBAs was administered to the 2022, 2023, and 2024 cohorts without revisions, enabling direct comparison of item psychometrics and cohort performance. Data from this analysis, including flagged items for revision, will guide future exam cycles. While the reuse of items raises concerns about memorization, the broad curriculum and high-stakes nature of the exam were considered to mitigate this issue. However, this is acknowledged as a limitation, and future strategies will include item banking and rotation.

Item Development and Quality Assurance Context

This post-hoc SBA analysis considers established item development best practices like blueprinting, writing guidelines, peer review, and item banking. Detailing their systematic implementation for the analyzed items is beyond the scope of this paper; however, the analysis provides quantitative feedback to strengthen DSMA's broader quality assurance procedures.

Data Collection

The data for were extracted from both manual examination sheets and electronic records, depending on the year of exam administration. Upon completion of the exams, responses were either entered manually or recorded electronically. Data entry was carried out using Microsoft Excel, followed by analysis using the Statistical Package for the Social Sciences (SPSS), version 25.

Statistical Methods and Data Analysis

To assess the psychometric properties of the SBAs, item analysis was conducted, focusing on three key indices: the Difficulty Index (P) and the Discrimination Index (D). These indices provided insights into the items' ability to differentiate between high and low-performing students, their level of difficulty, and the effectiveness of the distractors. The formula are as follows (Mahjabeen *et al*., 2017).

a) Difficulty Index (P)

The Difficulty Index (P) for each SBA was computed to determine the proportion of correct responses from both high and low-achieving students. The formula used for calculation was:

$$P = (HA + LA) / N * 100$$

Where:

- HA represents the number of students answering correctly in high achievers' group.
- LA represents the number of students answering correctly in the low achiever's group.
- N is the total number of students in both groups including non-responders.

Interpretation of P:

- A P value 0.76-1.00 indicates an easy item,
- A P value between 0.25-0.75 suggests average difficulty,

- A P value 0.00-0.24 indicates a difficult item.

b) Discrimination Index (D)

The Discrimination Index (D) was calculated to assess the ability of each item to distinguish between high and low performers.

D = (HA - LA) / N12

Interpretation of D:

- A D value ≤ 0.19 suggests poor discrimination.
- A D value 0.20-0.29 suggests acceptable discrimination,
- A D value between 0.30-0.39 indicates good discrimination,
- A D value ≥ 0.40 indicates excellent discrimination.

Statistical Procedures

After data entry and cleaning, the following analyses were conducted:

a) Descriptive Statistics: Calculated mean, standard deviation, and median for each cohort's responses.

b) Item Analysis: Evaluated the Difficulty (P) and Discrimination (D) indices for each SBAs.

c) Reliability Analysis: Applied Cronbach's alpha to assess the internal consistency of SBAs.

d) Item Revision: Flagged SBAs with poor psychometric properties for future revision or removal.

Items were categorized as follows:

a) **Retain**: Items with a Discrimination Index (D) ≥ 0.40 were generally retained.

b) **Revise**: Items with a Discrimination Index (D) between 0.20–0.39 were flagged for revision.

c) **Reject**: Items with a Discrimination Index (D) ≤ 0.19 were typically rejected.

Ethical Approval

The Ethical Review Committee of the Defence Services Medical Academy, Yangon, Myanmar, granted ethical approval (7/Ethics/ 2024).

Results

This study assessed three cohorts, each administered a test comprising 120 SBAs questions. The distribution of students across these cohorts is presented in table 1. Table shows, each cohort took the same test, ensuring consistency across the study. Despite slight variations in cohort size, all groups were exposed to the identical set of 120 SBAs questions, which enabled a reliable comparison. The maximum possible score was a 120, with a minimum score of zero. No negative marks were awarded.

Table 1: Distribution of students across cohorts

Cohort	No. of Students	No. of Questions
2022	39	120
2023	34	120
2024	33	120

Item Difficulty and Discrimination Indices

The key performance metrics—Standard Deviation (SD), Mean Difficulty Index (P) Mean Discrimination Index (D), Cronbach's Alpha, and the correlation between P and D are used to evaluate item quality across the three cohorts. SD reflects the variation in difficulty and discrimination, while P and D represent the average difficulty and discrimination power of the items, respectively. Cronbach's Alpha measures internal consistency, and the correlation between P and D shows how well difficulty correlates with the ability to distinguish between high and low performers.

The distribution of Item Difficulty and Discrimination Indices varied across the cohorts, and these metrics and distributions are summarized in Table 2.

Table 2: Item Difficulty and Discrimination Indices

Cohort	Mean Difficulty Index (P)	Mean Discrimination Index (D)	Cronbach's Alpha	SD	Correlation between P and D
2022	0.75	0.16	0.80	0.3	r = 0.16 (p> 0.05)
2023	0.74	0.16	0.78	0.7	r = 0.16 (p> 0.05)
2024	0.75	0.18	0.82	0.6	r = 0.18 (p< 0.05)

In Cohort 1 (2022), the SD was 0.3, with a mean difficulty index (P) of 0.75, a discrimination index (D) of 0.16, and a Cronbach's Alpha of 0.80, indicating acceptable internal consistency. The P-D correlation was 0.16 (not significant, p>0.05). In Cohort 2 (2023), the SD rose to 0.7, P slightly decreased to 0.74, D stayed at 0.16, and Cronbach's Alpha dropped to 0.78, suggesting moderate consistency. The P-D correlation remained 0.16 (not significant, p>0.05). In Cohort 3 (2024), the SD was 0.6, P stayed at 0.75, D increased to 0.18, and Cronbach's Alpha increased to 0.82, indicating strong consistency. The P-D correlation was 0.18 (significant, p<0.05).

Overall, internal consistency was high across all cohorts, with only Cohort 3 showing a significant P-D correlation.

Overall Test Performance

The overall performance across cohorts, including mean score, standard deviation (SD), and median score, is summarized in table 3.

Table 3: Overall Test Performance

Cohort	Mean	Standard Deviation (SD)	Median
2022	90.0	8.7	90.0
2023	88.4	8.7	88.0
2024	88.7	9.2	86.5

Table 3 presents the data for three cohorts. In Cohort 1 (2022), the mean was 90.0 out of 120.0, with a standard deviation (SD) of 8.7 and a median of 90.0, indicating a symmetrical distribution. In Cohort 2 (2023), the mean was 88.4, slightly lower than Cohort 1, with the same SD of 8.7. The median was 88.0, suggesting a slight shift to the lower side of the scale. In Cohort 3 (2024), the mean was 88.7, with an SD of 9.2, indicating a slightly wider spread. The median dropped to 86.5, showing a slight skew to lower values.

In summary, the means across the cohorts were similar, but Cohort 3 had a higher SD and a lower median, suggesting a wider spread and a slight shift in central tendency.

Distribution of Category Based on Difficulty and Discrimination Indices

The distribution of 120 questions for each cohort based on the Difficulty Index, Discrimination Index, and final decision (Retain, Revise, Reject) for each cohort is presented in table 4.

Table 4: Distribution of Category Based on Difficulty and Discrimination Indices

Index/Category	Cohort 1 (2022)	Cohort 2 (2023)	Cohort 3 (2024)
Difficulty Index			
Easy (0.76-1.00)	71 (59%)	66 (55%)	70 (58%)
Average (0.25-0.75)	44 (37%)	49 (41%)	44 (37%)
Difficult (0.00-0.24)	5(4%)	5 (4%)	6 (5%)
Discrimination Index			
Excellent (≥0.40)	18 (15%)	24 (20%)	28 (23%)
Good (0.30-0.39)	12 (10%)	9 (8%)	9 (8%)
Acceptable (0.20-0.29)	10 (8%)	14 (12%)	21 (18%)
Poor (≤0.19)	80 (67%)	73 (61%)	62 (52%)
Decision			
Retain (≥0.40)	40 (33%)	39 (33%)	58 (48%)
Revise (0.20-0.39)	76 (63%)	76 (63%)	58 (48%)
Reject (≤0.19)	4 (4%)	5 (4%)	4 (4%)

The analysis of Single Best Answers (SBAs) for the selected cohorts (2022–2024) shows key trends in difficulty, discrimination, and decision-making. Most questions were classified as "Easy" (55%–59%), with 37%–41% being "Average" in difficulty. Only 4%–5% were considered "Difficult."

In terms of discrimination, the percentage of "Excellent" questions (≥0.40) increased from 15% in 2022 to 23% in 2024, indicating progress in differentiating student performance. However, 52%–67% of questions were still deemed "Poor" (≤0.19), suggesting many failed to distinguish between high and low performers effectively.

Regarding decisions, the percentage of "Retain" (≥0.40) questions rose to 48% in 2024, reflecting improved question quality. Yet, 48%–63% were still categorized as "Revise," showing a need for further refinement. Only 4%–5% were "Rejected" (≤0.19), indicating that few questions were unsuitable.

Discussion

This study examines the performance and test characteristics of a 120-item SBA test in Obstetrics and Gynaecology, across three student cohorts (2022–2024), focusing on difficulty, discrimination, and internal consistency to evaluate test quality. While this item analysis provides crucial data for test refinement, it is one component of a larger quality improvement cycle in assessment, which ideally includes proactive measures such as blueprinting, item writer training, and pre-assessment peer review, as mentioned previously. Analysis revealed minimal performance variations across cohorts, with the test maintaining consistent internal reliability, indicating its stability over time.

Difficulty Index showed that the majority of questions were classified as "Easy," with Cohort 1 (2022) having 59% of questions labeled as such, Cohort 2 (2023) 55%, and Cohort 3 (2024) 58%. This trend aligns with the goal of making assessments accessible, ensuring most students answer questions correctly (Bjork *et al*., 2014). However, for high-stakes exams such as medical program exit tests, a greater proportion of "Average" or "Difficult" questions would be more appropriate to ensure adequate differentiation of student knowledge levels.

The Discrimination Index revealed varying effectiveness in differentiating performance. In Cohort 1 (2022), 67% of questions were categorized as "Poor" (≤0.19), while only 15% were rated "Excellent" (≥0.40). By Cohort 3 (2024), the percentage of "Excellent" questions improved to 23%, with 52% still classified as "Poor," suggesting ongoing efforts to enhance item quality. Cohort 2 (2023) mirrored Cohort 1, with 61% of questions deemed "Poor" and 20% rated "Excellent."

In the Decision category, the percentage of questions retained (≥0.40) increased to 48% in Cohort 3 (2024) from 33% in earlier cohorts, indicating better item quality. Despite this improvement, a significant portion (48%–63%) of questions still required revision, highlighting the need for continued refinement. The low percentage of "Rejected" questions (4%) suggests that most items were suitable for further use or revision.

Cronbach's Alpha values ranged from 0.78 to 0.82 across cohorts, demonstrating strong internal consistency (Kane, 2013; Cheung *et al*., 2024). These values meet the threshold for good reliability, ensuring the test consistently measures the intended construct. Despite variations in difficulty and discrimination indices, high Cronbach's Alpha values indicate that the test remained reliable over time.

The weak correlation between difficulty and discrimination indices across all cohorts aligns with previous research (Kline, 2015), suggesting that difficulty alone does not ensure effective discrimination between high and low performers. However, the significant correlation in Cohort 3 (2024) indicates that improvements in item writing enhanced the balance between difficulty and discrimination, reflecting positive changes in test design.

The mean scores were consistent across cohorts, with Cohort 1 (2022) achieving the highest mean score (90.0), followed by Cohort 2 (2023) (88.4) and Cohort 3 (2024) (88.5). Moderate standard deviations across cohorts suggest sufficient variation to differentiate performance, while the alignment of median and mean scores indicates a normally distributed dataset, ensuring that scores accurately represent the full range of student abilities (Von Hippel, 2005).

Decisions on test items, categorized as "Retain," "Revise," and "Reject," varied across cohorts. The number of questions retained increased in Cohort 3 (2024) to 37 items, compared to 30 in Cohort 1 (2022), suggesting improvements in item quality and the test refinement process. This iterative revision process is crucial for enhancing the validity and reliability of educational assessments (Bonner, 2012). The higher retention rate in Cohort 3 indicates that the test is becoming more efficient in evaluating student performance, benefiting from revisions made after each cohort's test administration.

This study has several limitations. First, it uses basic item analysis with classical test theory (P and D indices), lacking advanced models like Item Response Theory (IRT) for deeper insights. Second, the study's title suggests a broader scope than the item analysis alone addresses. It does not mention quality assurance processes in item development, such as exam blueprinting, item writing guidelines, peer review, or item banking, which may affect the accuracy of item performance interpretations. Third, while feedback loops like distractor analysis are intended, they are not yet fully implemented. Fourth, using the same 120 SBAs across three cohorts raises concerns of item exposure and memorization, suggesting the need for item banking and rotation in future assessments. Fifth, the cut-off values for P and D indices were based on literature but should be consistently reviewed for justification. Lastly, the study's focus on a single medical school limits the generalizability of the findings, though the methods are widely applicable.

Conclusion

This study analyzes the performance and quality of a 120-item Single Best Answer (SBA) test in Obstetrics and Gynaecology across three medical student cohorts (2022, 2023, and 2024), focusing on difficulty, discrimination, and internal reliability. The findings show that while slight variations were observed, the test maintained high reliability and validity. Most questions were classified as "Easy," ensuring accessibility for a broad student range. However, the discrimination index varied, indicating the need for ongoing refinement to better differentiate between high and low performers. Internal consistency was strong, with Cronbach's Alpha values between 0.78 and 0.82.

The study emphasizes the importance of continuous item review to improve test quality. The weak correlation between difficulty and discrimination highlights that difficulty alone doesn't ensure effective discrimination. The modest sample size and focus on SBA questions suggest a need for broader research into other test formats and contexts.

Future studies should refine test items for better discrimination and explore advanced statistical methods like Item Response Theory (IRT) to enhance the test's applicability across diverse educational settings. This research contributes valuable insights into educational assessments, underscoring the need for continuous refinement in test development.

References

Babo, R., Babo, L. V., Suhonen, J. T. and Tukiainen, M., 2020. E-Assessment with Multiple-Choice Questions: A 5 Year Study of Students' Opinions and Experience. *International Journal of Learning and Development*, 10(1), pp. 15-29. Available at: https://doi.org/10.28945/4491 [Accessed 19 March 2025].

Bjork, E. L., Little, J. L., Storm, B. C. and Cognition, 2014. Multiple-choice testing as a desirable difficulty in the classroom. *Journal of Applied Research in Memory and Cognition*, 3(3), pp. 165-170. Available at: https://doi.org/10.1016/j.jarmac.2014.03.002 [Accessed 19 March 2025].

Bone, E. and Prosser, M., 2020. Multiple Choice Questions: An Introductory Guide. *University of Melbourne*, [online] Available at: https://melbourne-cshe.unimelb.edu.au/__data/assets/pdf_file/0010/3430648/multiple-choice-questions_final.pdf [Accessed 16 September 2022].

Bonner, S. M., 2012. Validity in classroom assessment: Purposes, properties, and principles. In: *SAGE handbook of research on classroom assessment*. SAGE Publications, pp. 87-107.

Boursicot, K., Etheridge, L., Setna, Z., Sturrock, A., Ker, J., Smee, S. and Sambandam, E., 2011. Performance in assessment: Consensus statement and recommendations from the

Ottawa conference. *Medical Teacher*, 33(5), pp. 370-383. Available at: https://doi.org/10.3109/0142159X.2011.565831 [Accessed 19 March 2025].

Brauer, D. G. and Ferguson, K. J., 2015. The integrated curriculum in medical education: AMEE Guide No. 96. *Medical Teacher*, 37(4), pp. 312-322. Available at: https://doi.org/10.3109/0142159X.2014.970998 [Accessed 19 March 2025].

Case, S.M. and Swanson, D.B., 1998. *Constructing written test questions for the basic and clinical sciences* (2nd ed., pp. 22-25). Philadelphia: National Board of Medical Examiners.

Cheung, G. W., Cooper-Thomas, H. D., Lau, R. S. and Wang, L. C., 2024. Reporting reliability, convergent and discriminant validity with structural equation modeling: A review and best-practice recommendations. *Asia Pacific Journal of Management*, 41(2), pp. 745-783. Available at: https://doi.org/10.1007/s10490-023-09871-y [Accessed 19 March 2025].

Dennick, R., Wilkinson, S. and Purcell, N., 2009. Online eAssessment: AMEE guide no. 39. *Medical Teacher*, 31(3), pp. 192-206. Available at: https://doi.org/10.1080/01421590902792406 [Accessed 19 March 2025].

Downing, S.M., 2006. Twelve steps for effective test development. *Handbook of test development*, *3*, p.25.

Hartati, N. and Yogi, H. P. S., 2019. Item analysis for a better quality test. *English Language in Focus*, 2(1), pp. 59-70. Available at: https://doi.org/10.24853/elif.2.1.59-70 [Accessed 19 March 2025].

Kane, M. T., 2013. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), pp. 1-73. Available at: https://doi.org/10.1111/jedm.12000 [Accessed 19 March 2025].

Kumar, D., Jaipurkar, R., Shekhar, A., Sikri, G. and Srinivas, V., 2021. Item analysis of multiple choice questions: A quality assurance test for an assessment tool. *Medical Journal Armed Forces India*, 77, pp. S85-S89. Available at: https://doi.org/10.1016/j.mjafi.2020.11.007 [Accessed 19 March 2025].

Mahjabeen, W., Alam, S., Hassan, U., Zafar, T., Butt, R., Konain, S. and Rizvi, M., 2017. Difficulty index, discrimination index and distractor efficiency in multiple choice questions. *Annals of PIMS-Shaheed Zulfiqar Ali Bhutto Medical University*, 13(4), pp. 310-315. Available at: https://doi.org/10.48036/apims.v13i4.9 [Accessed 19 March 2025].

McTighe, J. and Wiggins, G., 2013. *Essential questions: Opening doors to student understanding*. ASCD.Olipas, C. N. P. and Luciano, R. G., 2024. Analyzing test performance of BSIT students and question quality: A study on item difficulty index and item discrimination index for test question improvement. *International Journal of Information Technology and Computer Science*, 16(3), pp. 1-12. Available at: https://doi.org/10.5815/ijitcs.2024.03.01 [Accessed 19 March 2025].

Siri, A. and Freddano, M., 2011. The use of item analysis for the improvement of objective examinations. *Procedia-Social and Behavioral Sciences*, 29, pp. 188-197. Available at: https://doi.org/10.1016/j.sbspro.2011.11.224 [Accessed 19 March 2025].

Verenna, A. M. A., Noble, K. A., Pearson, H. E. and Miller, S. M., 2018. Role of comprehension on performance at higher levels of Bloom's taxonomy: Findings from assessments of healthcare professional students. *Anatomical Sciences Education*, 11(5), pp. 433-444. Available at: https://doi.org/10.1002/ase.1768 [Accessed 19 March 2025].

Von Hippel, P. T., 2005. Mean, median, and skew: Correcting a textbook rule. *Journal of Statistics Education*, 13(2). Available at: https://doi.org/10.1080/10691898.2005.11910556 [Accessed 19 March 2025].

Wrigley, W., Van Der Vleuten, C. P., Freeman, A. and Muijtjens, A., 2012. A systemic framework for the progress test: Strengths, constraints and issues: AMEE Guide No. 71. *Medical Teacher*, 34(9), pp. 683-697. Available at: https://doi.org/10.3109/0142159X.2012.704437 [Accessed 19 March 2025].