Journal of University Teaching & Learning Practice

Volume 19
Issue 3 *Quarterly Issue 2*

Article 11

2022

Assessment of multiple choice question exams quality using graphical methods

Mustafa S. Yousuf
Hashemite University, Jordan, msysalman@yahoo.com

Katherine Miles
Hashemite University, Jordan, katemiles@doctors.org.uk

Heather Harvey
Limestone University, United States of America, drhlharvey@gmail.com

Mohammad Al-Tamimi
Hashemite University, Jordan, mohammad.altamimi@hu.edu.jo

Darwish Badran
University of Jordan, Jordan, msysalman@yahoo.com

Recommended Citation

Yousuf, M. S., Miles, K., Harvey, H., Al-Tamimi, M., & Badran, D. (2022). Assessment of multiple choice question exams quality using graphical methods. *Journal of University Teaching & Learning Practice, 19*(3). https://doi.org/#N/A!

Assessment of multiple choice question exams quality using graphical methods

Abstract

Exams should be valid, reliable, and discriminative. Multiple informative methods are used for exam analysis. Displaying analysis results numerically, however, may not be easily comprehended. Using graphical analysis tools could be better for the perception of analysis results. Two such methods were employed: standardized x-bar control charts with standard error of measurement as control limits and receiver operator characteristic curves. Exams of two medical classes were analyzed. For each exam, the mean, standard deviation, reliability, and standard error of measurement were calculated. The means were standardized and plotted against the reference lines of the control chart. The means were chosen as cut-off points to calculate sensitivity and specificity. The receiver operator characteristic curve was plotted and area under the curve determined. Standardized control charts allowed clear, simultaneous comparison of multiple exams. Calculating the control limits from the standard error of measurement created acceptable limits of variability in which the standard deviation and reliability were incorporated. The receiver operator characteristic curve graphically showed the discriminative power of the exam. Observations made with the graphical and classical methods were consistent. Using graphical methods to analyse exams could make their interpretation more accessible and the identification of exams that required further investigation easier.

Practitioner Notes

1. Exams should be valid, reliable, and discriminative
2. Classical methods to analyze exam quality represent data numerically
3. Numerical representation of data may not be readily understood by department staff
4. Graphical methods to analyze exam represent data in easy-to-understand charts
5. Control charts and receiver operator characteristic curves can be employed for such purposes

Keywords

control charts, exam analysis, medical exams, receiver operator characteristic curve

Introduction

In many educational settings, including medical education, learning is often impacted by formative and summative assessments (Knight & LTSN Generic Centre, 2001). Formative assessments occur during the learning process to enhance learning. Summative assessments occur at the end of the learning process and, therefore, reflect the students' final level of achievement and performance (Al-Kadri, 2012). Many types of assessments are used in medical education: essay questions, multiple-choice questions (MCQs), objective structured clinical examinations (OSCEs), long cases, short cases, and others (Tabish, 2008).

Multiple-choice question examinations are widely used to assess medical student learning (Zaidi et al., 2018). At a basic level, multiple-choice questions can assess students' recall of knowledge. However, MCQs can also be written to assess higher levels of cognitive reasoning (McCoubrie, 2004; Palmer and Devitt, 2007; Schuwirth & van der Vleuten, 2004). Consideration of Bloom's taxonomy of six cognitive domains may be helpful when developing MCQ examinations. These six domains are used by students to learn, retain, and apply new information and consist of: knowledge, comprehension, application, analysis, synthesis, and evaluation (Bloom et al., 1956). The first four domains can be assessed through well written multiple-choice questions (Masters et al., 2001).

There are two types of multiple-choice question, true/false questions and single best option questions. True/false questions usually consist of a stem with information and a statement which, then, the student has to indicate if it is true or false. Some true/false questions can consist of a stem question and a list of options from which the student must select all options that are 'true' in response to the stem question (NBME, 2020). In comparison, single best option multiple-choice questions consist of a stem with information and a question followed by three or more options from which the student must choose the single, most accurate answer. The other realistic, less accurate, or incorrect options are referred to as 'distractors'. Often three or four distractors are given for each question item; however, there is no consensus as to the optimal number of distractors required (Gierl et al., 2017).

There are several advantages of multiple-choice question examinations. A broad range of topics in the curriculum can be assessed and the questions linked to specific educational objectives (Brady, 2005). It is a standardized, objective type of assessment that can overcome the subjectivity of essay and oral assessment formats (Hammond et al, 1998). Multiple-choice question examinations are efficient as they enable assessment of a large number of students in a short space of time (Pamplett and Farnill, 1995) and the scores can be quickly generated through machine marking (Hammond et al., 1998).

On the other hand, limitations of multiple-choice question examinations relate to the type of assessment, utility for students, and the challenge of question writing. A major criticism of multiple-choice question examinations is that they encourage superficial learning and memorization of facts (Pamplett and Farnill, 1995) and cannot assess the higher cognitive domains of synthesis and evaluation. Also, for healthcare students, multiple-choice questions do not accurately reflect the complexity of clinical situations (Brady, 2005). For students, multiple-choice question examinations provide limited personalized feedback (Nicol, 2007) which reduces the potential for students to learn from their mistakes. A further drawback to multiple-choice question examinations is that it can be difficult to write good questions with plausible distractors and without construction flaws (Gierl et al., 2017; Holsgrove, 1992). Therefore, it is recommended that a team of experts write multiple-choice questions and that questions are reviewed by colleagues for identification of technical item flaws (Brady, 2005).

For an examination to be an efficient tool for assessment, it needs to have high validity and reliability (Miller et al., 2009). Validity is the degree to which the examination measures what it aims to measure (Schuwirth & van der Vleuten, 2004). Reliability is the degree of consistency of the examination and measure of confidence that the same results would be obtained if the exam was re-administered to the same students with all other factors being equal (Miller et al., 2009).

At a basic level, the mean, percentages, and standard deviation of an examination can be used for simple analysis of exam results. There are several approaches to statistical analysis of multiple-choice question examinations, including: classical test theory, factor analysis, cluster analysis, item response theory, and model analysis (Ding and Beichner, 2009). Each approach has a slightly different purpose and algorithm, with the ultimate goal of making sense of the raw data. The statistical analysis approach that is feasible and provides the best interpretation of the data, is the one to use (Ding and Beichner, 2009).

Various methods can be employed to review an exam to enhance its quality (Pugh et al., 2016; Zaidi et al., 2017) and involving the staff in this process is important (Zaidi et al., 2016). However, for faculty members who are not trained in using different statistical techniques, it may be difficult to understand data presented numerically or in tabular format. Indeed, academic staff may not easily engage with these numerical analyses (Crisp & Palmer, 2007). To enhance the accessibility of the analysis, a method must be used that can be easily created and interpreted. Illustrating information in a suitable graph may achieve this goal (In & Lee, 2017; Tait et al., 2010). An important point in successful data presentation is not to clutter the data in a table, but to take the key points and display them in a suitable graph (Lowe & Borkan, 2021). Graphical presentation of data can communicate quantitative information in a meaningful way (Cleveland and McGill, 1985). When a graph is created, quantitative data is encoded using position, size, shape, symbols, and color. A person visually decodes this information when they look at a graph or pictorial representation of data. Meyer et al. (1999) conducted experiments to compare the interpretation of tabular and graphical presentation of data for two types of tasks. For the first data extraction task, trends were easier to read when data was presented graphically; whereas, point comparisons were easier to make when data was presented in tabular format (Meyer et al., 1999). For the second prediction task, graphical presentation of the data had clear advantages and participants could use their prior knowledge together with the decoded information from the graph more efficiently than if data was presented in a table (Meyer et al., 1999). Graphical formats are particularly powerful when the information presented is a task relevant to the person reading the graph, as the person can make use of the visual patterns presented (Meyer et al., 1999). Graphical representation of data may also last in a person's memory longer (Bavdekar, 2015).

In the faculty where this research was conducted, single-best-option MCQ exams are used to test the students. The classical analysis of mean, standard deviation, and percentages is used to analyze the exams and the results are presented numerically to the staff. From the results of such analysis, the staff must decide if the exam was optimal or not and if further analysis is required. We propose the use of two graphical methods to provide simple and comprehensive visual analysis of MCQ type exam results, namely: standardized x-bar control charts and receiver operator characteristic (ROC) curves.

X-bar control charts

In control charts (Figure 1), a series of data points are plotted against three main reference lines: a central line (CL), an upper control limit (UCL), and a lower control limit (LCL). The data plotted represent statistics of a certain process measured at various times. Control charts were first created by Walter Shewhart as a means to obtain statistical control on the products of industrial processes (Shewhart, 1931). In later years, control charts were used in various fields like banking (Yasin et al., 1991), human performance (Burney & Al-Darrab, 1998), and education (Besha, 2012; Hrynkevych, 2017; Patil et al., 2020; Schafer et al., 2011; Tomak et al., 2016).

The central line of a control chart represents the mean of the population. Control limits are calculated from the standard deviation of the population. If these parameters are unknown, estimates are used instead. The control limits represent boundaries within which variability between measurements is considered acceptable. Measurements falling within the control limits are considered in-control and require no further investigation; points falling outside are considered out-of-control and require additional inspection. Suitable estimates of the control limits must be chosen, however. If the control limits form a wide band around the central line, some out-of-control points might be missed (Figure 1, a). If, on the other hand, the control limits form a narrow band, too many points will be out-of-control (Montgomery, 2013) as can be seen in Figure 1, b. Creating a not-too-wide and not-too-narrow band will probably reflect a truer picture of the plotted data (Figure 1, c).

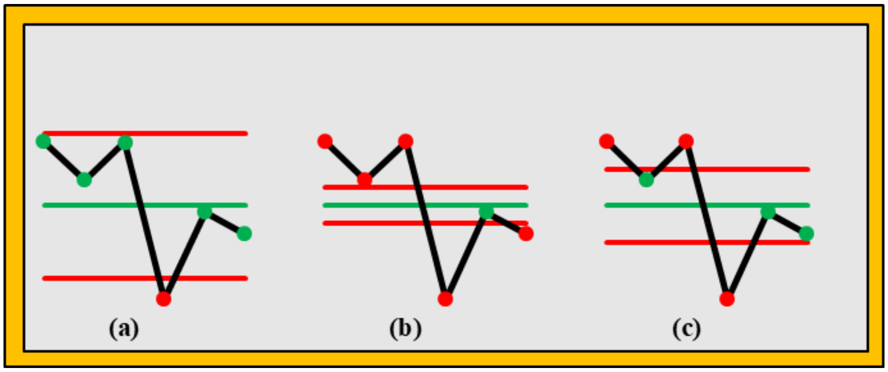

Figure 1:
Control chart with various width of the control limits. The upper and lower (red) lines represent the upper and lower control limits, respectively. The central (green) line is the central line (CL). (a) Control limits were set so high that only one value was out of control. (b) Only one point was in control due to low control limits. (c) Choosing suitable control limits will give a better picture of the data. (No actual data was used to plot these charts).

According to classical test theory, a student’s exam score can be considered to be the sum of a true score and an error score. The true score comes from the questions answered correctly by the student due to knowledge of the material and skill in taking the exam. The error score comes from questions answered correctly by some other way like chance or cheating. The true score is what reflects the actual competency of the student. The standard error of measurement (SEM), calculated from the standard deviation and reliability of the test, can be used to create a range within which the true score is found (refer to Harvill, 1991, and Musselwhite & Wesolowski, 2018, for an in-depth discussion of the SEM).

This concept can be expanded to a cohort of students taking several exams. Exams are usually compared by their means. Variability in the means of different exams are expected. To determine if this variability was because an exam was too difficult (or too easy) for the students, other factors must be considered like standard deviation, item difficulty, and item discrimination. Exam scores represent a statistic of the educational process that is measured repeatedly over time. Therefore, we propose to use control chart to analyze exam results and employing the SEM to create the control limits of the control charts, thus forming a band within which variability between test means can be considered acceptable.

Receiver operator characteristic curve

The receiver operator characteristic (ROC) curve (Figure 2) is a graph that plots the true positive rate (sensitivity) of a diagnostic test against the false positive rate (1-specificity) of the test for different values of the cut-off threshold (King & Eckersley, 2019; Krzanowski & Hand, 2019). This curve was first used by Egan et al. (1961) in signal detection to determine the ability of a receiver to differentiate between signals (true detection) and noise (false reports). Since then, it has been used in various disciplines. In medicine, it has been used in radiology (Lusted, 1971) and medical diagnostic tests (Baduashvili et al., 2019; Hajian-Tilaki, 2013; Obuchowski & Bullen, 2018). ROC curves have also been used in education (Bowers and Zhou 2019) and in exam analysis (Dhakal et al., 2018; Taib & Yusoff, 2014). Determining the area under the ROC curve (AUC) gives an estimate of the probability that the test can discriminate between normal and abnormal (Hanley & McNeil, 1982). To facilitate the interpretation of the curve, a straight line, representing the 50% probability that the test result is positive, is plotted. The further the curve is from this line, the better the test is considered. Exams can be considered as a diagnostic test for the level of performance of the students and ROC curves can be used to analyze them.

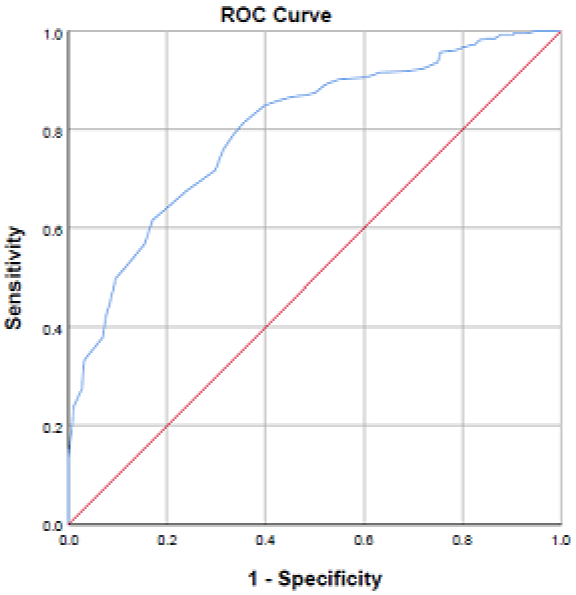

Figure 2:
Receiver operator characteristic (ROC) curve. The red line represents the 50% probability that the test result is positive. (No actual data was used to graph this curve).

Method

The study was conducted at the Faculty of Medicine, The Hashemite University in Jordan during the first term of the academic year 2019-2020. To keep the classes and subjects confidential, codes were used. To illustrate the two suggested methods, analysis was performed on the exams of two different cohorts of students in that term: (1) Cohort A, studied 3 subjects (each had 3 exams) and a fourth subject with only one exam included in the analysis, the exams were designated A1-A10; (2) Cohort B, studied 3 subjects (each had 3 exams) and a fourth subject with only one exam included in the analysis, the exams were designated B1-B10. The two cohorts were at different educational levels and studied different courses. In the faculty where the research was conducted, the performance of the students in the various subjects is assessed by three exams: two exams during the term (example, exams A1 and A2) and one final exam at the end of the term (example, exam A3). All exams are of the multiple-choice-question type. Clinical courses, however, are assessed by two exams: a practical, OSCE type, exam (not analyzed in this study) and a final, MCQ type, exam (exams A10 and B10, included in this study).

The exams analyzed consisted of several dichotomously scored multiple-choice questions with five options each. For each exam, a Microsoft Excel file tabulating the answers of all the students to all the questions was generated. Each correct answer was coded as 1, and each incorrect answer was coded as 0. The results were imported to IBM SPSS version 25 to calculate the mean, standard deviation, and reliability (Cronbach's Alpha). For dichotomous exams, reliability should be calculated by the Kuder-Richardson formula 20 (Kuder & Richardson, 1937). However, for dichotomous questions, Cronbach's Alpha yields the same result (Ritter, 2010). In addition, the average difficulty and discrimination indices (sum of the index for all the questions divided by the number of questions) were calculated by Microsoft Excel. For the evaluation of the discrimination index, the results of the upper and lower 27% of the examinees were compared (Kelley, 1939).

X-bar control charts

For each exam, the SEM was determined by the equation:

(1) … $SEM = SD * \sqrt{1 - Reliability}$ SD, standard deviation of the exam

The number of questions in the exams were not, necessarily, the same. Before plotting, the mean and the SEM were rescaled to percentages: (statistic / number of questions) x 100. These rescaled values were averaged and the means of the exams were standardized according to the average SEM as follows:

(2) … $Standardized\ mean = \frac{(mean\ of\ exam - average\ of\ means)}{average\ SEM}$

The central line (the population mean) was given the value of the average of the means. After standardization, the CL was plotted at the 0 average SEM point. The UCL was chosen as +2 SEM and the LCL was chosen as -2 SEM. An additional upper warning limit (UWL = +1 SEM) and a lower warning limit (LWL = -1 SEM) were also plotted. The standardized means of the exams were graphed against these reference lines. Microsoft Excel was used for these calculations and to plot the control charts.

Receiver operator characteristic curve

The performance of the students was determined using the grade point average (GPA) of the preceding term. Accordingly, the students were classified into: (1) above average student*:* the students whose GPA was more than or equal to the average GPA of the entire cohort; (2) below average students*:* the students whose GPA was less than the average GPA of the entire cohort. After scoring the exam, it was used as a diagnostic test according to the following rules: (1) the test was (+ve) if the student's score in the exam was more than or equal to the mean of the exam and (2) the test was (-ve) if the student's score was less than the mean of the exam.

Only students with known GPA and exam scores were included in this analysis. The mean of these valid scores was used as the cut-off point to determine the sensitivity and specificity for each exam. The ROC curve was plotted and the AUC was calculated. In addition, for courses formed of several exams, the total mark was calculated and ROC curve analysis was performed based on these marks. MedCalc version 19 was used for these calculations and to plot the ROC curve.

Results

Classical analysis

The total number of Cohort A students registered in the different subjects ranged between 475-490. For Cohort B students, the total number ranged between 235-240. The students who were absent from the exams took an essay-type makeup exam and, therefore, their results were not included in the analysis. Classical analysis of the exams is shown in Table 1.

Table 1:
Classical analysis for the exams of Cohort A and Cohort B.

Cohort A							
Exam	Examinees	Questions	Mean	SD	Average DI	Average Disc. I	Reliability
A1	465	30	23.08	3.97	0.769	0.315	0.744
A2	471	30	22.21	4.13	0.740	0.331	0.731
A3	475	40	30.39	5.85	0.760	0.350	0.826
A4	467	30	18.12	3.76	0.604	0.306	0.637
A5	472	30	19.16	4.55	0.639	0.373	0.775
A6	467	40	25.82	5.80	0.645	0.351	0.781
A7	482	30	23.35	4.16	0.778	0.335	0.770
A8	482	30	21.73	4.56	0.724	0.363	0.782
A9	484	40	30.19	5.85	0.755	0.345	0.843
A10	476	40	26.54	5.00	0.663	0.301	0.754

Cohort B

					Average	Average	
Exam	Examinees	Questions	Mean	SD	DI	Disc. I	Reliability
B1	229	60	37.06	9.66	0.618	0.386	0.879
B2	231	20	16.81	2.26	0.841	0.251	0.607
B3	230	80	55.50	10.94	0.694	0.333	0.891
B4	239	60	39.71	7.74	0.662	0.318	0.830
B5	238	20	16.64	2.09	0.832	0.244	0.551
B6	238	80	50.25	11.02	0.628	0.336	0.881
B7	233	60	45.72	7.23	0.762	0.298	0.841
B8	230	20	19.32	0.94	0.966	0.092	0.314
B9	206	80	57.33	10.19	0.717	0.313	0.887
B10	225	40	28.00	4.56	0.700	0.278	0.723

DI = difficulty index; Disc. I = discrimination index; SD = standard deviation.

X-bar control charts

The control chart for Cohort A is shown in Figure 3. To explain the methodology, consider exam A3. The number of questions in this exam was 40 and the mean was 30.39 (Table 1). The standard deviation and reliability were 5.85 and 0.826, respectively. According to formula (1), the SEM was 2.44. For exam A3, the rescaled mean and SEM were, thus, 75.98 and 6.10, respectively. Similar calculations were made for all the exams of Cohort A yielding an average (rescaled) mean of 70.79 and an average (rescaled) SEM of 6.72. Hence, according to formula (2), the standardized mean of exam A3 was 0.77 which was plotted on the control chart. For this cohort, exams A4 and A5 were relatively difficult as their means fell below the LWL but above the LCL. Exam A7 was relatively easy as the mean fell above the UWL but still below the UCL. No points were out-of-control as none of the means fell outside the control limits.

For Cohort B, Figure 4 shows the control chart. Three exams were out-of-control: two (B1 and B6) were difficult as their means fell below the LCL, and one exam (B8) was easy as its mean fell above the UCL. B2 and B5 were relatively easy and B4 was relatively difficult.

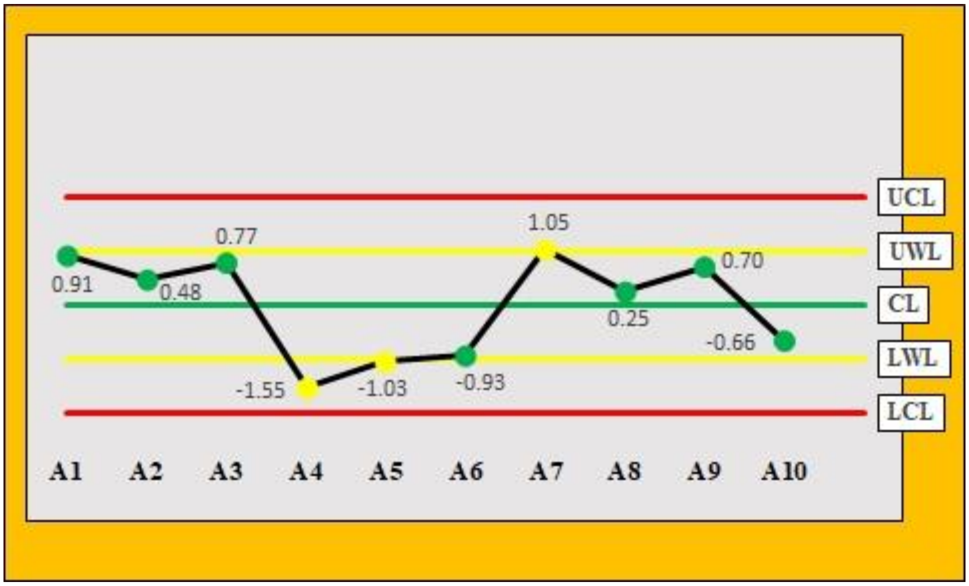

Figure 3:
Control chart of the standardized means of exams A1-A10 of Cohort A. The numbers indicate the standardized mean of exam. UCL = upper control limit = +2; UWL = upper warning limit = +1; CL = central line = 0; LWL = lower warning limit = -1; LCL = lower control limit = -2.

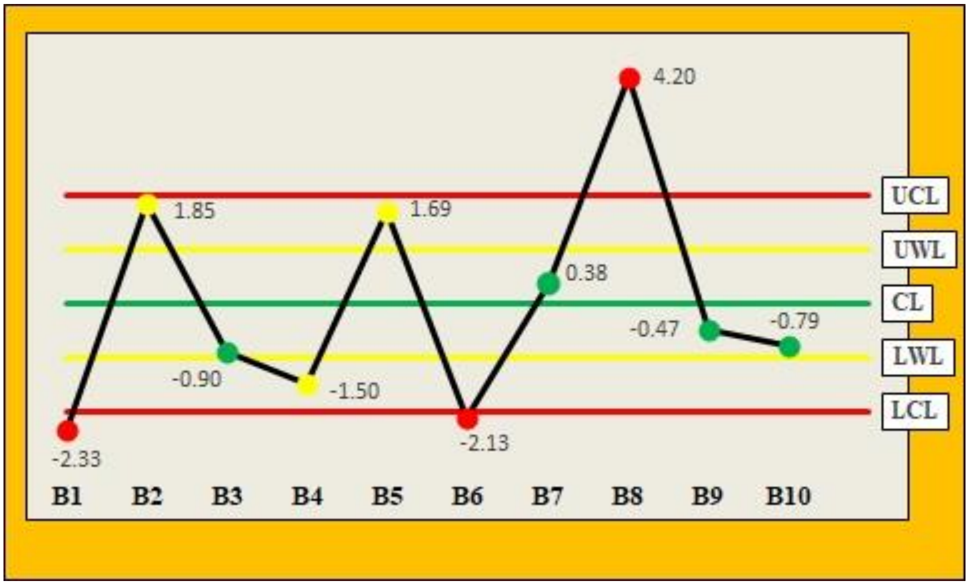

Figure 4:
Control chart of the standardized means of exams B1-B10 of Cohort B. The numbers indicate the standardized mean of exam. UCL = upper control limit = +2; UWL = upper warning limit = +1; CL = central line = 0; LWL = lower warning limit = -1; LCL = lower control limit = -2.

Receiver operator characteristic curve

The sensitivity, specificity, and the AUC for Cohort A exams are shown in Table 2. The highest sensitivity was recorded for the total marks of exams A7-9 (66.40%). The highest specificity was recorded for A4 (73.40%). The largest AUC (0.736), however, was found for the total marks of exams A1-3 with a sensitivity of 64.23% and a specificity of 67.65%.

Table 2:
ROC curve analysis for the exams of Cohort A.

Exam	Cut-off point [a]	Sensitivity	Specificity	AUC (95% CI)
A1	23.10	57.99	72.53	0.729 [0.686 to 0.769]
A2	22.15	57.18	61.76	0.670 [0.626 to 0.713]
A3	30.51	63.56	68.63	0.698 [0.654 to 0.739]
Total	74.95	64.23	67.65	0.736 [0.694 to 0.775]
A4	19.13	53.80	73.40	0.685 [0.641 to 0.727]
A5	20.34	58.15	61.70	0.647 [0.602 to 0.691]
A6	25.98	58.63	69.89	0.687 [0.642 to 0.729]
Total	65.22	57.07	68.09	0.695 [0.651 to 0.737]
A7	23.50	63.11	59.60	0.678 [0.634 to 0.721]
A8	21.89	64.48	63.64	0.674 [0.629 to 0.716]
A9	30.45	65.85	63.37	0.706 [0.663 to 0.747]
Total	75.01	66.40	60.78	0.709 [0.665 to 0.749]
A10	26.59	59.73	60.78	0.639 [0.594 to 0.683]

AUC = area under the curve; CI = confidence interval; ROC = receiver operator characteristic.

[a] The mean of the exam, as calculated for the ROC curve, is chosen as the cut-off point.

For Cohort B exams, sensitivity, specificity, and the AUC are shown in Table 3. The highest sensitivity was recorded for B4 (81.51%). The highest specificity was recorded for the total marks of B4-6 (80.51%). The largest AUC (0.861) was, however, found for B3 (sensitivity 80.17% and specificity of 76.07%) and the total marks of B4-6 (sensitivity 80.67% and specificity of 80.51%). The smallest AUC (0.609) was found for B8 with sensitivity of 64.04% and specificity of 53.98%. The ROC curve for B3 and B8 are shown in Figure 5.

Table 3:
ROC curve analysis for the exams of Cohort B.

Exam	Cut-off point [a]	Sensitivity	Specificity	AUC (95% CI)
B1	24.44	75.86	77.59	0.822 [0.767 to 0.869]
B2	16.86	75.86	49.57	0.709 [0.646 to 0.767]
B3	27.73	80.17	76.07	0.861 [0.810 to 0.903]
Total	68.88	79.31	77.78	0.848 [0.796 to 0.892]
B4	26.49	81.51	71.19	0.847 [0.794 to 0.890]
B5	16.65	71.43	52.99	0.675 [0.611 to 0.735]
B6	25.11	77.31	76.92	0.835 [0.782 to 0.880]
Total	68.16	80.67	80.51	0.861 [0.811 to 0.903]
B7	30.52	74.56	65.52	0.807 [0.750 to 0.856]
B8	19.32	64.04	53.98	0.609 [0.542 to 0.673]
B9	28.67	76.42	74.23	0.852 [0.796 to 0.898]
Total	74.57	80.87	62.93	0.809 [0.753 to 0.858]
B10	28.05	65.63	63.56	0.725 [0.660 to 0.784]

AUC = area under the curve; CI = confidence interval; ROC = receiver operator characteristic.
[a] The mean of the exam, as calculated for the ROC curve, is chosen as the cut-off point.

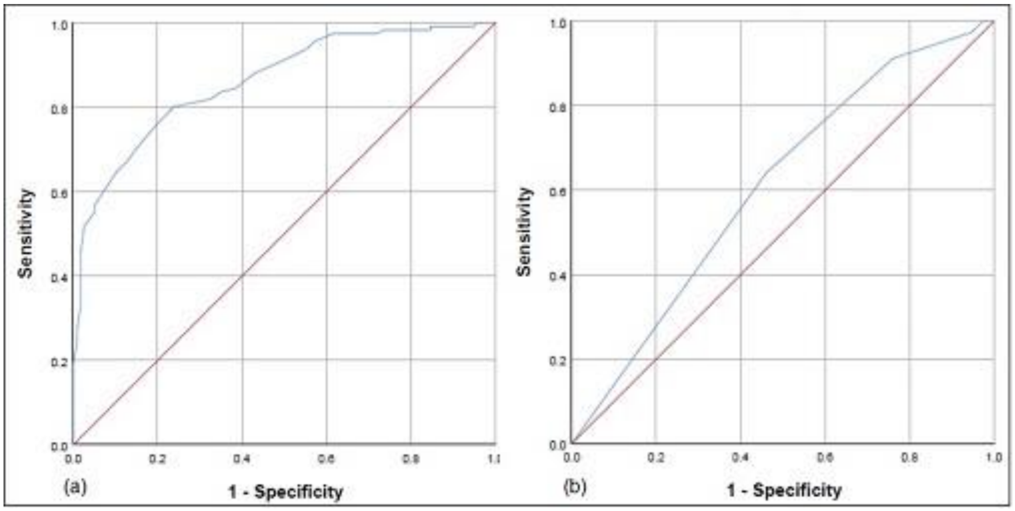

Figure 5:
ROC curves for the B3 (a) and B8 (b) exams of Cohort B.

Tabular vs graphical formats

The results of the exams where shown to members of the department in both the tabular (Tables 1, 2, and 3) and graphical (Figures 3, 4, and 5) formats. The staff were simply asked to identify the exams that were not optimal. The numbers shown in the tables had to be explained and discussed before determining which exams were difficult and which were easy. When looking at the graphs, however, the staff found it much more feasible to identify the out-of-control exams.

Discussion

To assess if an exam has achieved the goals for which it was made, the results of the exam has to be analyzed. Classically, this usually includes calculating the mean, standard deviation, and reliability of the exam and the difficulty and discrimination indices of the exam questions. These statistics are, usually, presented as numbers or a table (like Table 1). It is undeniable that these methods are quite informative. But classical exam analysis tables cannot easily show if the variability between exams could be considered acceptable or not, which is critical to determine which exams require enhancement to improve their quality.

By using control charts, the exams taken by a cohort of students in one term were compared with each other. This controlled for the differences that could exist between different cohorts. The exams were adjusted so that they could be easily compared on the same scale. Moreover, the means of these exams were compared to the mean of the means, which can be considered as the true mean of our population of students. Control charts allow for variability between exams by creating a band (bounded by control limits) within which the value of the mean can fluctuate. Different methods are used to calculate the control limits; this depends on the data and the purpose for using the control charts. Besha (2012) used the range of the data to calculate the control limits to establish a new grading scheme, while Hrynkevych (2017) used the range for the assessment of education quality. Tomak et al. (2016) used range and standard deviation to create the limits for the comparison of difficulty indices of the exam. Alabi-Labaika and Ahani (2015) compared the results of examinations in two departments by x-bar control charts using the standard deviation to evaluate the control limits. On the other hand, Patil et al. (2020) chose a predetermined value as their desired upper limit.

In this research, the aim for using control charts was to analyze and compare exam means. Using the SEM to establish the control limits led to the incorporation of both the standard deviation and reliability of the exam into the calculation and this created a not-too-wide-not-too-narrow band within which the variation of exam means can be considered acceptable. Moreover, since the exams analyzed were dichotomous and one mark was given for each correct answer, and since the means were rescaled, the plotted points also represented the average difficulty index of that exam (Kuder & Richardson, 1937, formula 22). So, by this method, four of the main statistics shown in Table 1 were used to create the control charts. Rescaling and standardizing the means according to the SEM made the comparison even easier as it made the control limits straight lines instead of zig-zag lines that occur when there are different sample sizes (as in this study).

It is not easy to compare exams by looking at their statistics unless they have the same number of questions. Comparing B1 and B4 (Table 1), a difference of about 2.5/60 points between the means was found with the other statistics being similar. By looking at this number, one might think that the difference is not important. The control chart (Figure 4), however, clearly showed that B1 was an out-of-control difficult exam (its mean fell below the LCL) and B4 was relatively difficult (its mean fell between the LWL and the LCL). Looking at the statistics of B3 and B6, a difference between the means of about 5/80 points was sufficient to make these two exams separate from each other in the control chart (Figure 4), where B6 was an out-of-control difficult exam and B3 was an in-control acceptable exam (its mean was between the CL and LWL). Without looking at the chart, two exam means that differ by only 5/80 points might be considered at the same level.

Other features of the chart of Cohort B (Figure 4) may also require exploration. Three practical exams were analyzed: one was easy (B8) and two were relatively easy (B2 and B5). This should prompt the instructors involved to improve the quality of these exams. Other interesting features seen in the chart were the great variability between the exams and that more exams were found below the CL than above. After examining the positions and trend of the various points plotted on the chart, the department could make a better decision as to which exams require more thorough investigation allowing the instructors to have some ideas of how to redesign their exams to enhance their quality for future use.

All points in the control chart for Cohort A (Figure 3) were within the control limits and only three exams were relatively out-of-control points. Looking at such a chart, the department might decide that the variability seen might be regarded as justifiable and the exams were well designed and require no further enhancement. However, exams A4-6 belong to a subject taught by an instructor different from the instructors of the other subjects taught. The subject is generally considered not more difficult than the others. Examining the chart, however, indicated that these exams (A4-6) were more difficult. This might indicate a problem with the methods used by that instructor to teach the subject or design the exams. After looking at the chart, the department might discuss this with the instructor to try to enhance the exams for the future.

Exam B8 had a very high mean (Table 1). This could indicate that either the exam was too easy because of poor design or the students of that cohort were very good that they answered the exam well. By looking at this exam’s other statistics, the standard deviation was very low, so most of the marks were clustered around the mean; the average difficulty index was very high, which indicated that the exam was easy; and the average discrimination index was very low, which indicated that this exam did not discriminate very well between high and low performing students. Without going through all this analysis, the exam can be directly compared with the other exams of Cohort B by simply looking at the control chart (Figure 4) from which we could see that the mean of exam B8 was far above the UCL making it an out-of-control easy exam. In addition, the mean can be used as a cut-off point to calculate sensitivity and specificity. For B8, these were approximately 64% and 54%, respectively. This meant that about half of the low performing students did well in this exam. The area under the ROC curve indicated the ability of the exam to discriminate between the high and low performers. The AUC of this exam was the smallest (Figure 5, b), and its value (0.609) meant that the discrimination was poor (Hosmer et al., 2013). This showed that the control chart, the ROC curve, and the classical analysis were highly consistent.

For Class B, the total of B4-6 had the highest specificity and AUC, but not the highest sensitivity (that was found for B4). It is not enough to look at either the sensitivity or the specificity alone. Considering both of these values together gives a better understanding of the exam. The graph of the ROC curve and the value of the AUC, however, directly showed the discriminative ability of the exams.

Although ROC curve analysis is widely used in medical diagnostic tests, its use in exam analysis has not been as frequent. Dhakal et al. (2018) used ROC curve analysis to compare multiple-choice and short-answer exams for medical students. Taib and Yusoff (2014) used this method to compare multiple-choice and long-case exams for medical students with the passing grade as the cut-off point. In the current study, the mean was chosen as the cut-off point because, ideally, the goal of any medical school is to produce good doctors that are at least ‘above average’ in their performance.

Limitations

The suggested graphical methods were used to assess the quality of multiple choice question exams. The application of such methods for other types of exams should be studied to determine their benefits. For the control chart to have any meaningful interpretation, several exams must be included in the analysis and, therefore, it must be performed at the end of a term or a year when the students have completed several courses. This means that this method can only be used for quality improvement of exams in the future. Plotting the graphs was easily done using statistical software. However, incorporating these methods directly into the faculty exam analysis software would make it more feasible. Although the staff expressed their preference for the graphical methods, formal statistical tests should be carried out in the future to determine if the graphical methods were easier to understand than the numerical methods. With further research, other graphical methods may be employed to analyze exam results.

Conclusion

Control charts enabled the comparison of several exams using graphs rather than tables. The use of the SEM to create the control limits led to the incorporation of the standard deviation and reliability into the calculation of these limits. This led to the creation of acceptable limits within which variability can occur which allowed the easy determination of out-of-control exams that require further attention. Using ROC curve analysis gave a straightforward graphical method to determine the discriminative power of an exam. Accordingly, control charts and ROC curve analysis of MCQ exams provided new, simple, and comprehensive methods for exam analysis that “forces us to notice what we never expected to see” (Tukey, 1977, p.vi). These methods may be of great benefit anywhere MCQ type exams are used (high school, undergraduate study, postgraduate study, or others) to facilitate the presentation and comprehension of exam results; thus, assisting in the process of exam quality improvement and standardization.

References

Alabi-Labaika, A. B., & Ahani, E. (2015). Comparing two examination results using means of sample means and control charts. *Journal of Education and Practice*, *6*(4), 106-113.

Al-Kadri, H. M., Al-Moamary, M. S., Roberts, C., & van der Vleuten, C. P. M. (2012). Exploring assessment factors contributing to students' study strategies: Literature review. *Medical Teacher, 34*(sup1), S42-S50. https://doi.org/10.3109/0142159X.2012.656756

Baduashvili, A., Guyatt, G., & Evans, A. T. (2019). ROC anatomy - Getting the most out of your diagnostic test. *Journal of General Internal Medicine, 34*(9), 1892-1898. https://doi.org/10.1007/s11606-019-05125-0

Bavdekar, S. B. (2015). Using tables and graphs for reporting data. *The Journal of the Association of Physicians of India*, 63(10), 59-63.

Besha, B. (2012). Students' performance evaluation using statistical quality control. *International Journal of Science and Advanced Technology, 2*(12), 75-79.

Bloom, B., Englehart, M., Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. Longmans, Green.

Bowers, A. J., & Zhou, X. (2019). Receiver operating characteristic (ROC) area under the curve (AUC): A diagnostic measure for evaluating the accuracy of predictors of education outcomes. *Journal of Education for Students Placed at Risk (JESPAR), 24*(1), 20-46. https://doi.org/10.1080/10824669.2018.1523734

Brady, A.-M. (2005). Assessment of learning with multiple-choice questions. *Nurse Education in Practice*, 5(4), 238-242. https://doi.org/https://doi.org/10.1016/j.nepr.2004.12.005

Burney, F. A., & Al-Darrab, I. (1998). Performance evaluation using statistical quality control techniques. *Work Study, 47*(6), 204-212. https://doi.org/10.1108/00438029810238606

Cleveland W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, 229(4716), 828-833. https://doi.org/10.1126/science.229.4716.828

Crisp, G. T., & Palmer, E. J. (2007). Engaging academics with a simplified analysis of their multiple-choice question (MCQ) assessment results. *Journal of University Teaching & Learning Practice, 4*(2), 88-106.

Dhakal, A., Yadav, S. K., & Dhungana, G. P. (2018). Assessing multiple-choice questions (MCQs) and structured short-answer questions (SSAQs) in human anatomy to predict students' examination performance. *Journal of Research in Medical Education & Ethics, 8*(2), 127-131. https://doi.org/10.5958/2231-6728.2018.00024.0

Ding, L., & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics - Physics Education Research, 5*(2), 020103. https://doi.org/10.1103/PhysRevSTPER.5.020103

Egan, J. P., Greenberg, G. Z., & Schulman, A. I. (1961). Operating characteristics, signal detectability, and the method of free response. *The Journal of the Acoustical Society of America, 33*(8), 993-1007. https://doi.org/10.1121/1.1908935

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87(6), 1082-1116. https://doi.org/10.3102/0034654317726529

Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian Journal of Internal Medicine, 4*(2), 627-635.

Hammond, E. J., McIndoe, A. K., Sansome, A. J., & Spargo, P. M. (1998). Multiple-choice examinations: Adopting an evidence-based approach to exam technique. *Anaesthesia*, 53(11), 1105-1108. https://doi.org/10.1046/j.1365-2044.1998.00583.x

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36. https://doi.org/10.1148/radiology.143.1.7063747

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practice, 10*(2), 33-41. https://doi.org/10.1111/j.1745-3992.1991.tb00195.x

Holsgrove, G. J. (1992). Guide to postgraduate exams: Multiple-choice questions. *British Journal of Hospital Medicine*, 48(11), 757-761.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). John Wiley & Sons, Hoboken.

Hrynkevych, O. S. (2017). Statistical analysis of higher education quality with use of control charts. *Advanced Science Letters*, *23*(10), 10070-10072. https://doi.org/10.1166/asl.2017.10390

In, J., & Lee, S. (2017). Statistical data presentation. *Korean Journal of Anesthesiology, 70*(3), 267-276. https://doi.org/10.4097/kjae.2017.70.3.267

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30*(1), 17-24. https://doi.org/10.1037/h0057123

King, A. P., & Eckersley, R. J. (2019). Descriptive statistics III: ROC analysis. In *Statistics for Biomedical Engineers and Scientists* (pp. 57-69). Academic Press.

Knight, P., & LTSN Generic Centre. (2001). *A briefing on key concepts: Formative and summative, criterion and norm-referenced assessment*. Learning and Teaching Support Network, York.

Krzanowski, W. J., & Hand, D. J. (2009). *ROC curves for continuous data*. CRC Press, Boca Raton. https://doi.org/10.1201/9781439800225

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151-160. https://doi.org/10.1007/BF02288391

Lowe, R. C., & Borkan, S. C. (2021). Effective medical lecturing: Practice becomes theory. *Medical Science Educator*, 31, 935-943. https://doi.org/10.1007/s40670-020-01172-z

Lusted, L. B. (1971). Signal detectability and medical decision-making. *Science, 171*(3977), 1217-1219. https://doi.org/10.1126/science.171.3977.1217

Masters, J. C., Hulsmeyer, B. S., Pike, M. E., Leichty, K., Miller, M. T., & Verst, A. L. (2001). Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. *Journal of Nursing Education* 40(1), 25-32. https://doi.org/10.3928/0148-4834-20010101-07

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: A literature review. *Medical teacher*, *26*(8), 709-712. https://doi.org/10.1080/01421590400013495

Meyer, J., Shamo, M. K., & Gopher, D. (1999). Information structure and the relative efficacy of tables and graphs. *Human Factors*, 41(4), 570-587. https://doi.org/10.1518/001872099779656707

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching* (10th ed.). Pearson Education, Upper Saddle River.

Montgomery, D. C. (2013). *Introduction to statistical quality control* (7 ed.). John Wiley & Sons, Hoboken.

Musselwhite, D. J., & Wesolowski, B. C. (2018). Standard error of measurement. In B. B. Frey (Ed.), *The SAGE encyclopedia of educational research, measurement, and evaluation* (Vol. 1, pp. 1588-1590). SAGE Publications, Thousand Oaks. https://doi.org/10.4135/9781506326139.n658

The National Board of Medical Examiners (NBME) (2020). NBME Item-writing Guide Retrieved 05-April-2022 from https://www.nbme.org/item-writing-guide

Nicol, D. (2007). E-assessment by design: Using multiple-choice tests to good effect. *Journal of Further and Higher Education*, 31(1), 53-64. https://doi.org/10.1080/03098770601167922

Obuchowski, N. A., & Bullen, J. A. (2018). Receiver operating characteristic (ROC) curves: Review of methods with applications in diagnostic medicine. *Physics in Medicine & Biology, 63*(7), 07TR01.

Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education: Modified essay or multiple choice questions? Research paper. *BMC Medical Education*, 7(1), 49. https://doi.org/10.1186/1472-6920-7-49

Pamphlett, R., & Farnill, D. (1995). Effect of anxiety on performance in multiple choice examination. *Medical Education*, 29(4), 297-302. https://doi.org/10.1111/j.1365-2923.1995.tb02852.x

Patil, T. S., Belitskaya-Levy, I., & Allaudeen, N. (2020). Increasing the frequency of night float teaching with a daily management system: Where medical education meets quality improvement. *Medical Science Educator, 30*(4), 1399-1403. https://doi.org/10.1007/s40670-020-01106-9

Pugh, D., Champlain, A. D., Gierl, M., Lai, H., & Touchie, C. (2016). Using cognitive models to develop quality multiple-choice questions. *Medical teacher*, *38*(8), 838-843. https://doi.org/10.3109/0142159X.2016.1150989

Ritter, N. L. (2010, February). *Understanding a widely misunderstood statistic: Cronbach's alpha* [Paper]. Annual meeting of the Southwest Educational Research Association, New Orleans, USA.

Schafer, W. D., Coverdale, B. J., Luxenberg, H., & Ying, J. (2011). Quality control charts in large-scale assessment programs. *Practical Assessment, Research, and Evaluation, 16*. https://doi.org/10.7275/5t2n-f843

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2004). Different written assessment methods: what can be said about their strengths and weaknesses? *Medical Education*, *38*(9), 974-979. https://doi.org/10.1111/j.1365-2929.2004.01916.x

Shewhart, W. A. (1931). *Economic control of quality of manufactured product* Van Nostrand Reinhold Company, Princeton.

Tabish, S. A. (2008). Assessment methods in medical education. *International journal of health sciences*, *2*(2), 3-7.

Taib, F., & Yusoff, M. S. B. (2014). Difficulty index, discrimination index, sensitivity and specificity of long case and multiple choice questions to predict medical students' examination performance. *Journal of Taibah University Medical Sciences*, *9*(2), 110-114. https://doi.org/10.1016/j.jtumed.2013.12.002

Tait, A. R., Voepel-Lewis, T., Zikmund-Fisher, B. J., & Fagerlin, A. (2010). The effect of format on parents' understanding of the risks and benefits of clinical research: A comparison between text, tables, and graphics. *Journal of Health Communication*, *15*(5), 487-501. https://doi.org/10.1080/10810730.2010.492560

Tomak, L., Bek, Y., & Cengiz, M. A. (2016). Graphical modeling for item difficulty in medical faculty exams. *Nigerian Journal of Clinical Practice*, *19*(1), 58-65. https://doi.org/10.4103/1119-3077.173701

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley Publishing Company, Reading.

Yasin, M. M., Green, R. F., & Wafa, M. (1991). Statistical quality control in retail banking. *International Journal of Bank Marketing*, *9*(2), 12-16. https://doi.org/10.1108/02652329110137729

Zaidi, N. L. B., Grob, K. L., Monrad, S. M., Kurtz, J. B., Tai, A., Ahmed, A. Z., Gruppen, L. D., & Santen, S. A. (2018). Pushing critical thinking skills with multiple-choice questions: Does Bloom's taxonomy work? *Academic Medicine*, 93(6), 856-859. https://doi.org/10.1097/acm.0000000000002087

Zaidi, N. L. B., Grob, K. L., Yang, J., Santen, S. A., Monrad, S. U., Miller, J. M., & Purkiss, J. A. (2016). Theory, process, and validation evidence for a staff-driven medical education exam quality improvement process. *Medical Science Educator*, *26*(3), 331-336. https://doi.org/10.1007/s40670-

016-0275-2

Zaidi, N. L. B., Monrad, S. U., Grob, K. L., Gruppen, L. D., Cherry-Bukowiec, J. R., & Santen, S. A. (2017). Building an exam through rigorous exam quality improvement. *Medical Science Educator, 27*(4), 793-798. https://doi.org/10.1007/s40670-017-0469-2