JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Measures of Difficulty Index of Various Sets of Multiple-Choice Questionnaires

Ishan Malik
University School of Automation and Robotics (USAR),
Guru Gobind Singh Indraprastha University, East Delhi Campus, Delhi-110092

Kamal Nain
Department of Statistics, Hindu College, University of Delhi, Delhi-110007

This paper discusses the effect of question difficulty on students' performance in multiple choice questions (MCQs). It also explains the essence of the impact to support balanced assessments in learning environments that are fair and productive. The relevance of this is how grading strategies can be changed to expose the real abilities of the learner and thereby benefit educational assessment with improved learning. This study aims at bridging this gap through an analysis to find out the areas of question difficulty that might affect students' performance and, therefore, suggest changes to the grading system that would take care of those areas. The data in this paper is analyzed using statistical methods in order to evaluate different grading strategies. It uses a simulated dataset of student responses to MCQs of differing levels of difficulty but without necessarily detailing each of its steps. The results concluded that question difficulty does hold a level of significance with performance. Grading adjustments, which include weighted scoring, normalization, penalization, and bell curve application, do have varied effects on reflecting true student capability in assessments. This research develops an understanding of how question difficulty impacts student performance and provides insights into fair practices for assessment. It is to be noted that an educator will have to develop a balanced examination design along with balanced grading, which is a true reflection of student knowledge and skills.

Keywords— Question Difficulty, Student Performance, Grading Strategies, Statistical Methods, Educational Assessment

1 Introduction

Understanding of such influences that question difficulty can produce on the students' performances in MCQs becomes an important aspect, ensuring quality educational assessment, and consequently improved learning outcomes. The very nature of MCQs is such that there can be a problem of over-difficulty in the assessment of the exams if these are not properly counterbalanced by the educator. In doing so, it moves to explore the complexity of the relationship between student performance and question difficulty and how the variation of question difficulty can either challenge or retard student achievement. Such would be invaluable insights for educators, curriculum developers, and policymakers, among others who may be interested in optimizing the strategies of teaching and designs of examination that would foster a fair and effective learning environment.

In the literature, no research addresses the general correlation between student performance and the difficulty of questions being asked, nor the differences in grading methodologies when applied in an attempt to balance this interrelationship.

Works of studies such as [Karelia et al., 2013; Towns, 2014; Kar et al., 2015] have noted the requirement for the balance of their analysis between question difficulty and the effect it has on student performance. There is, however, still room for properly understanding how different grading strategies can be designed in such a manner that they take into account the difficulty spectrum of questions on examinations. In other words, the exam should be both tough and fair [Sevenair and Burkett, 1988].

This is the exact gap the present research hopes to fill by providing detailed analysis of how the leveling of question difficulty affects student performance and how, in the face of these differences, grading strategies can be modified. Using a simulated dataset of student response patterns to MCQ items with different levels of difficulty, this paper describes an innovative methodological approach to grading, reflective of genuine student capabilities. This study would thus add value to the extent that it has great potential to inform much-needed improvements in educational practice, specifically around the crafting of assessments to ensure that they are rigorous and fair. This study intends to add to existing literature that provides an underlying foundation for other future research seeking optimal educational outcomes. This study further refines the subtleties of question difficulty and the implications that arise from its grading, thereby adding a layer over an already established knowledge and helping to pave the way for future research in the quest for ideal educational outcome.

2 Related Works

The examination standards domain and the diversity of approaches and methodologies have been addressed in this assessment. Earlier landmark work in this area was of Bandaranayake (2008), Zainudin et al. (2012), Karelia et al. (2013), Towns (2014), Kar et al.(2015), Baldwin (1984) and Hambleton et al. (1991). Andrew and Hecht (1976) found differences in the standards set by different judges, showing that for the most part, there were no differences in group standards when compared to individual judgments.

Berk (1986) reviewed a practical guide offering 38 standard-setting and adjusting methods recommended to educational and testing authorities. Beuk (1984) proposed to balance absolute versus relative standards systematically in examination settings by suggesting a linear function for the compromise between the pass score and pass rate.

Brennan (1992) considered the use of generalizability theory for the purposes of including more information of the different sources of error in measurement than is possible using classical test theory and ANOVA. In one of the works, Chambers, Boulet, and Gary (2000) reported on data related to time allocation in SP assessments and found that the mean time was adequate.

De Gruijter (1985) has described the compromising models as those that continue to compromise the standards, with the help of the observed score distribution, for effective balance between absolute and relative standards. The Medical Council of Canada used an innovative approach to set the pass marks for completing a national OSCE which was described by Dauphinee et al. (1997). This method is effective in identifying borderline performances with the use of global judgments of physician examiners.

Floreck and De Champlain (2001) saw the point that standardization was important in SP examinations for maintaining reliability in a large-scale implementation. Hodges et al. (2002) reported that there may be a very subtle difference in the style of interviewing between the expert clinician and novice that may give clear indication of the need of OSCE measures to tap into the small nuances of behavior for the full evaluation of the clinical skill.

Kane (1994) went on to reiterate that it should be within the validity of the passing score that the reflection of the specified performance standards and reasonableness in intended decisions are encapsulated. Chinn and Horitz (2002) reviewed Angoff's standard setting techniques and concluded that there was a reasonable stability in percentage-based judgments, but there existed variability in dichotomous ratings post introduction of item difficulty data and group discussions.

As revealed by Boulet, De Champlain, and McKinley (2003), it is difficult to come up with standards, especially on performance assessments that involve SPs in the medical field. Ben-David (2000) has emphasized the fact there is a growing need for innovative new approaches to setting standards in assessing professional competence, and he also emphasized that there is no universally recommended method in this regard because the field is new.

De Champlain et al. (2001) assessed automated scoring and developed cost-effective models to evaluate clinical skills that sustained the important aspects of expert judgment. At last, Clauser et al. (1997) compared the adequacy of different methods for grading performance assessment with automated ratings versus those of experts and have put forward its possible applications for different tasks. All of these studies contribute meaningfully to the evolving field of methodologies for assessment, bringing to light complexities and necessities of fair and accurate evaluation within worlds of education and profession.

3 Methodology

This research examines student performance on multiple-choice questions (MCQs) of varying difficulty levels. The aim is to understand how different question difficulties impact student performance, guiding improvements in teaching methods and exam design. Figure 1 shows the proposed methodology for this research.

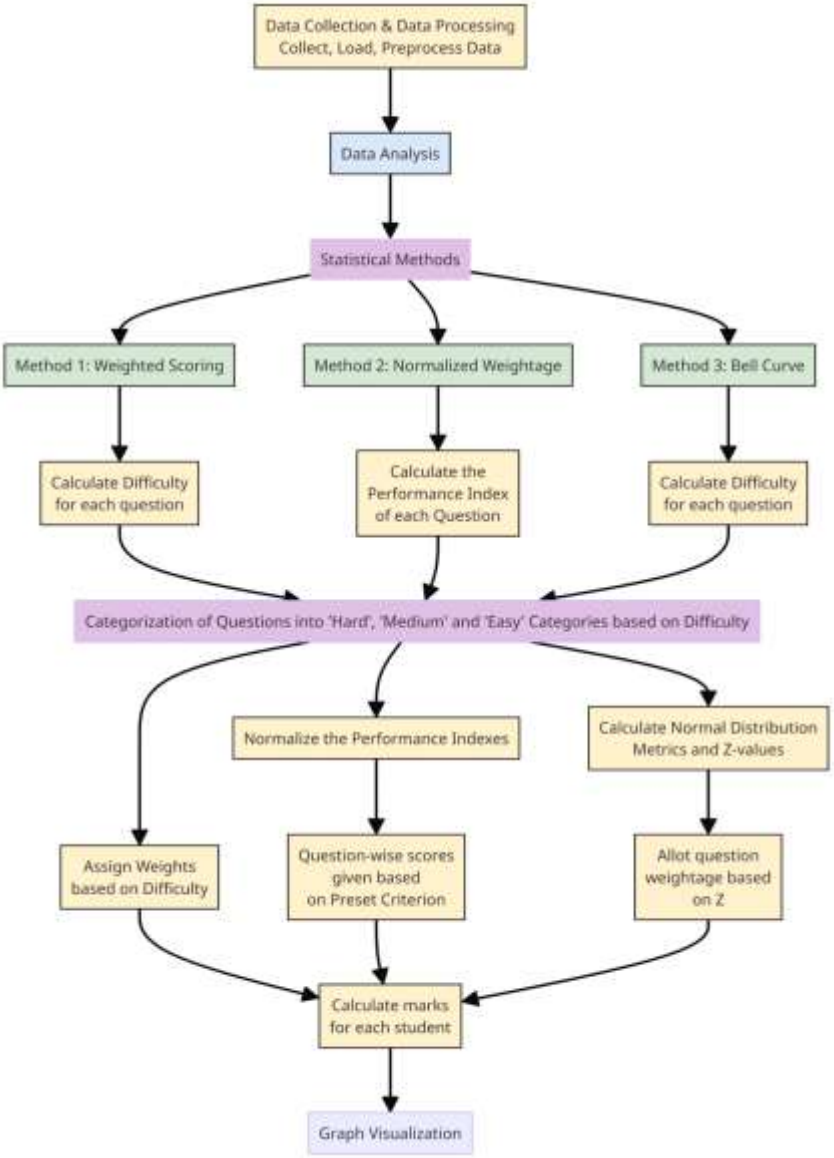

Figure 1: Proposed Methodology

3.1 Methodological Approach

A Research Problem and Data Requirement

This research aimed to find out the student performance on multiple-choice questions (MCQs) of various difficulty levels. The study required data on students' answers to a set of 100 MCQs. The simulated data presented the responses of students to questions of varying difficulty, forming a basis for assessment and refinement of grading strategies.

B Aim of the Research

To determine a clear understanding of how different levels of question difficulties affect student performance. Data were collected to represent different levels of question difficulty and student accuracy. This allowed identifying trends in performance, indicating where teaching methods or exam designs need adjustments.

C Type of Data Collected

Quantitative data was necessary to fulfill the research aim. The generated dataset included numerical representations of students' answers and question difficulties. Quantitative analysis enabled accurate measurement of performance grading and its effectiveness.

D Data Collection Methods

Primary data was used, generated through a Python script that simulates realistic student responses. The script created a dataset representing answers from 20 students to 100 MCQs. Simulated data allowed testing and comparing different grading methods in a controlled environment, free from ethical and practical issues.

E Objectives and Research Questions

The methodology was tailored to effectively assess the research questions. Each method addressed grading nuances based on question difficulty and student performance, validating the research's aims and questions.

3.2 Methods of Analysis

A Data Processing and Analysis

Data were processed and analyzed using statistical methods to evaluate grading strategies. Methods included difficulty-based weightage for questions, normalization, and statistical distributions application. These methods offered detailed assessments of student performance, considering each question's difficulty.

B Tools and Statistical Methods

Python libraries like pandas, numpy, seaborn, and matplotlib were used for data processing and analysis. These tools facilitated efficient handling of large datasets and complex calculations, providing clear insights into grading strategies' effectiveness.

3.3 Evaluation and Justification of Methodological Choices

The chosen methods provided a balanced approach to analyzing MCQ performance. Techniques were selected to accurately reflect the difficulty level of questions, allowing for a fair grading system. The approach relied on simulated data, which might not fully capture real student behavior nuances, but allowed for controlled, in-depth analysis.

3.4 Grading Strategies Analysis

3.4.1 Method 1: Weighted Scoring Based on Question Difficulty

This method assigns weights to questions based on their difficulty to calculate student scores. For i-th question, the difficulty is calculated using the formula:

$$\text{difficulty}_i = \frac{\text{number of correct answers}}{\text{total number of students that attempted (i)-th question}} \quad (1)$$

Based on this difficulty level, weights are assigned to the questions. The weight (w_i) for the (i)-th question, determined by its difficulty level, is defined as follows:

$$w_i = \begin{cases} 2 & \text{if difficulty}_i \leq 0.125, \\ 1.75 & \text{if } 0.125 < \text{difficulty}_i \leq 0.375, \\ 1.5 & \text{if } 0.375 < \text{difficulty}_i \leq 0.5, \\ 1.25 & \text{if } 0.5 < \text{difficulty}_i \leq 0.625, \\ 1 & \text{if difficulty}_i > 0.625 \end{cases} \quad (2)$$

3.4.2 Method 2: Normalized Weightage with Penalization

In this method, scores are adjusted based on the normalized weightage of questions, with penalties for incorrect answers [Sachdev, (2019)]. The normalization process commences with the calculation of the performance index (p_i) for each (i)-th question as the ratio of the number of correct answers to the total number of students:

$$p_i = \frac{\text{number of correct answers for question } i}{\text{total number of students that attempted (i)-th question}} \quad (3)$$

When no student attempts (i)-th question, (p_i) becomes 0, leading to $(\frac{1}{p_i})$ to trend towards infinity. To address this issue, the range of (p_i) is clipped to lie between 0.2 and 0.8. This ensures stability and prevents computational extremes. The aggregate of the reciprocals of the performance indices determines $(\frac{1}{\lambda})$, the sum of all $(\frac{1}{p_i})$'s:

$$\frac{1}{\lambda} = \sum_{i=1}^{n} \frac{1}{p_i}, \quad (4)$$

where (n) is the total number of questions. The normalized weightage (w_i) for each question is calculated as follows:

$$w_i = \frac{\frac{1}{p_i}}{\frac{1}{\lambda}} = \frac{\lambda}{p_i}, \quad (5)$$

which simplifies to multiplying the reciprocal of the performance index (p_i) by (λ). Marks are then allotted to a student (j) on question (i) based on this normalized weightage. The scoring rule is defined as:

$$\text{score}_{ij} = \begin{cases} 100 \times w_i & \text{if the answer is correct,} \\ -\frac{100 \times w_i}{3} & \text{if the answer is incorrect,} \\ 0 & \text{if unanswered.} \end{cases} \quad (6)$$

This approach ensures that the weightage of each question is normalized relative to the overall difficulty of the set of questions, facilitating a fair and balanced scoring mechanism.

3.4.3 Method 3: Bell Curve Application

This method applies a bell curve grading system, standardizing scores across a normal distribution for each (i)-th question. The difficulty of each question (i) is quantified before applying the Z-score formula. Please refer to Equation (1) for calculating the difficulty (difficulty_i) of each question (i). Then, the Z-scores for each question (i) are calculated using the formula:

$$z_i = \frac{\text{difficulty}_i - \mu}{\sigma} \tag{7}$$

where (μ) is the mean difficulty of all questions, and (σ) is the standard deviation of these difficulty levels. Z-scores adjust the grading scale, placing the performance of the (j)-th student in the context of the overall group, leading to a distribution of scores that follow a bell curve. Marks for the (j)-th student on the (i)-th question are then assigned based on these Z-scores, with score ranges mapped to a predefined grading scale, ensuring that grades are distributed evenly across the performance spectrum. The predefined grading scale for the (j)-th student on the (i)-th question, based on (z_i) ranges, is defined as follows:

$$w_i = \begin{cases} 4 & \text{if } -3 < z_i \leq -2, \\ 3 & \text{if } -2 < z_i \leq -1, \\ 2 & \text{if } -1 < z_i \leq 0, \\ 1 & \text{if } z_i > 0 \end{cases} \tag{8}$$

3.4.4 Difficulty Categorization and Visualization

Questions are categorized based on their difficulty levels, aiding in the analysis of question distribution. Questions are segmented into 'Easy', 'Medium', and 'Hard' Categories based on their difficulty (difficulty_i) calculated in Equation (1). The categorization is derived using the following formula:

$$\text{Difficulty Level}_i = \begin{cases} \text{Hard} & \text{if difficulty}_i \leq 0.125, \\ \text{Medium} & \text{if } 0.125 < \text{difficulty}_i \leq 0.625, \\ \text{Easy} & \text{if difficulty}_i > 0.625 \end{cases} \tag{9}$$

This categorization helps in visualizing the distribution of questions by difficulty levels, enabling a better understanding of the assessment's balance and fairness. The process involves using numerical thresholds based on the calculated difficulties to categorize questions, which is crucial for a structured analysis of question difficulty levels and subsequent visualization.

4 Results and Discussion

4.1 Overview of Analysis

The study was a performance-validation exercise of students' Multiple Choice Questions (MCQ) examinations and utilized various methods based on the MCQ dataset. Subsequently, the combination of answered questions and unanswered questions (coded as 100) was used to analyze data from 20 students answering 100 questions, employing three different methodologies for the evaluation and adjustment of the scoring system and the categorization of question difficulty. Table 1 shows the responses from students, which were taken at an initial point. In other words, it shows a subset of the dataset with answers from students to a number of initial sets of questions. The table represents an unprocessed raw snap of data forming a basis for further methodological adjustments and analysis.

Table 1: Student Responses to Questions

Question No.	Student Responses									
Q1	2	2	2	2	2	2	2	2	2	3
Q2	2	4	2	2	2	2	1	100	2	2
Q3	4	4	4	4	4	1	4	4	1	1
Q4	2	4	3	4	4	4	1	4	4	4
Q5	4	4	4	4	4	4	4	3	4	4

Table 2: Correct Answers for Questions (1-10 shown out of 100)

Question No.	Correct Answer
Q1	2
Q2	2
Q3	4
Q4	4
Q5	4
Q6	4
Q7	1
Q8	2
Q9	4
Q10	3
... 90 more questions ...	

Correct answers for problems contained in the attached sample are given in Table 2. Such an answer key serves very well in evaluating students' performance and making comparative studies of various scoring methods.

Table 3: Truncated Results for Students (20 students, 100 questions)

Question	Student_1	Student_2	Student_3	…	Student_19	Student_20
Q1	1	1	1	...	1	1
Q2	1	0	1	...	100	0
. . .	. . .	. . .	. . .	. . .	. . .	. . .
Q10	0	1	0	...	0	0
Data continues for each student across 100 questions						

Table 3 presents students' result that were derived after matching the student responses in Table 1 with the answer key in Table 2. It is a snapshot of performance data for 20 students on a 100-question test. This shows how there is variance in student response and lays the basis of how different grading strategies impact student performance.

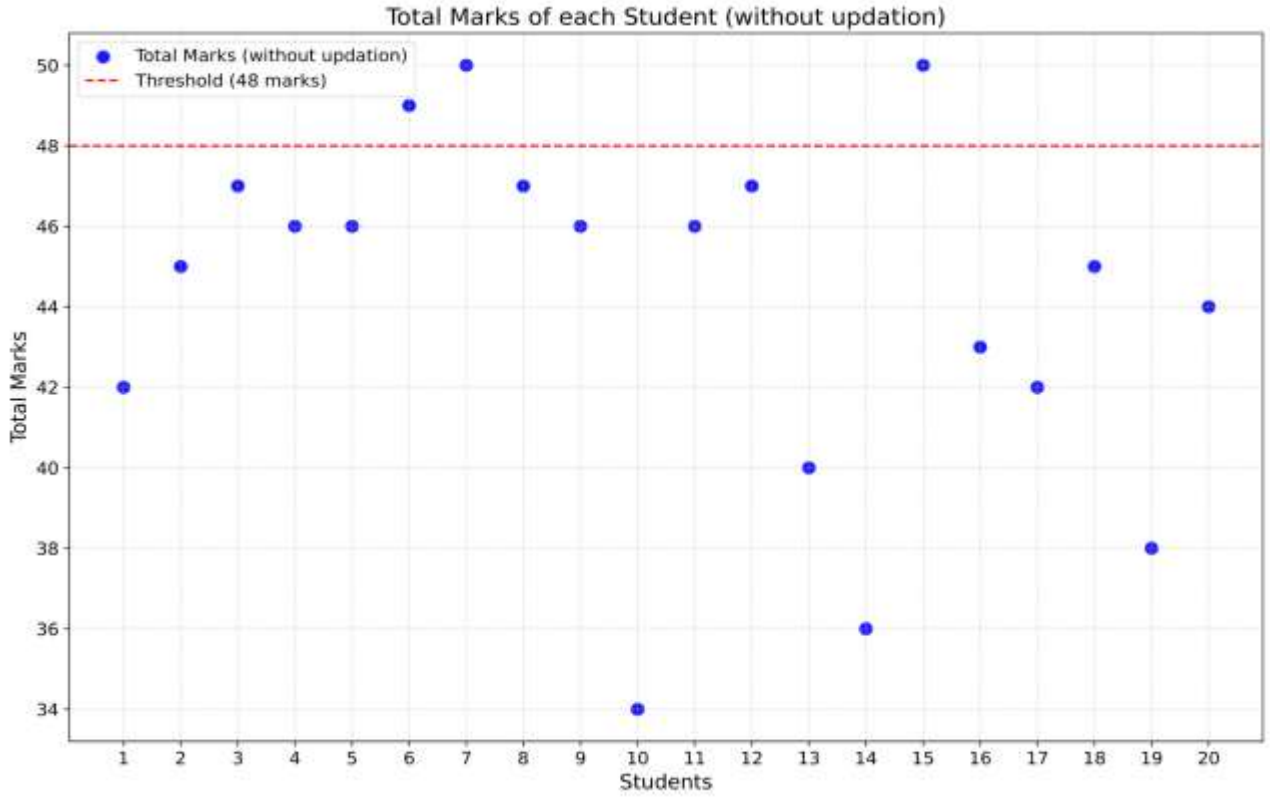

Figure 2: Original Distribution of Student Marks Before Weightage Adjustments

The scatter plot presented above reflects the raw scores of the students before the evaluation or adjustment methodology was taken up that would change some of these scores based on the weightages. In fact, it sets the first line for further comparisons of methods. Each of these tables and figures plays a crucial role in visualizing the data and supporting the analysis of the assessment's balance and fairness, as well as the effectiveness of the applied scoring methodologies.

4.2 Method 1: Basic Weightage Calculation

First, the number of correct attempts of students were counted and then divided with the total number of students to calculate the difficulty of each question. Later, in the next stage, according to the predefined criteria, the weight of every question was decided. Total marks per student were computed only after establishing these weightages. Replacement of any instance of 100, indicative of non-answered questions, with 0 was done, respectively, after which these scores were aggregated to come up with each student's total marks. This is shown in Table 4 and, it made it easy to come up with a scatter plot graph of the updated total marks for the students. The graph clearly shows the discrepancies that were previously present in the performance of the students. Moreover, this graph delineates the pass mark threshold at 48 marks (Figure 3).

Table 4: Summary of Original and Adjusted Weights for Questions

Question No.	Original Weight	Adjusted Weight
1	0.90	1.00
2	0.65	1.00
3	0.65	1.00
…	…	…
96	0.15	1.75
97	0.20	1.75
98	0.05	2.00
99	0.25	1.75
100	0.15	1.75

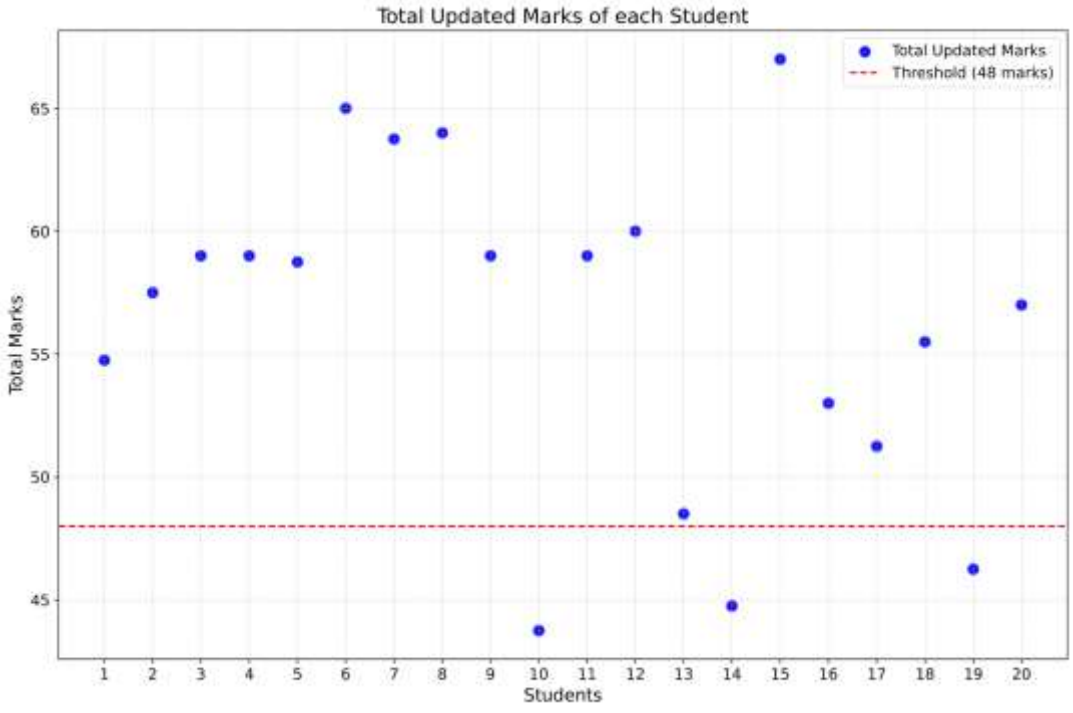

Figure 3: Weightage-adjusted distribution of student performance

The scatter plot presented above shows a distribution of marks scored by students for the preset weightage criterion. It describes the effect of such criteria on student scores and reflects moderate variations in performances. This shows the adjustments of weightage in shaping the overall distribution of marks, with a nuanced attitude, enunciating the impact that these predefined factors derive over the student outcomes. Please note that after applying this method the weightage of Easy questions remained the same but the weightage of Medium and Hard Questions was increased respectively, which led to the Total Marks or the Maximum Achievable Marks of the Paper to change and increase from the predetermined 100 marks. It is important to bring to notice that this change is done only for comparison purposes and the Total Marks or the Maximum Achievable Marks of the Paper still remains at 100 Marks only.

4.3 Method 2: Advanced Weightage Calculation

The second method advanced the analysis in that the clipped range of the reciprocal values of the performance index (p_i) of each question are computed and then normalized to get a final weightage for each question. This approach also took into account updating the student marks based on this final weightage, with conditions applied that would simulate the correct, incorrect, and not answered scenarios. The scatter plot in Figure 4 shows a more uniform spread of marks, with some students whose marks were affected greatly by the adjustment in weightage, indicated through a 10-mark threshold line.

Table 5: Comparative Analysis of Question Weightages: Initial, Inverse, and Normalized Values

Question No.	**Performance Index** (p_i)	**Reciprocal Value** $1/(p_i)$	**Normalized Weightage** (λ/p_i)
1	0.9	1.25	0.00
2	0.65	1.54	0.01
3	0.65	1.54	0.01
…	…	…	…
97	0.25	4.00	0.02
98	0.15	4.00	0.02
99	0.2	4.00	0.02
100	0.15	4.00	0.02

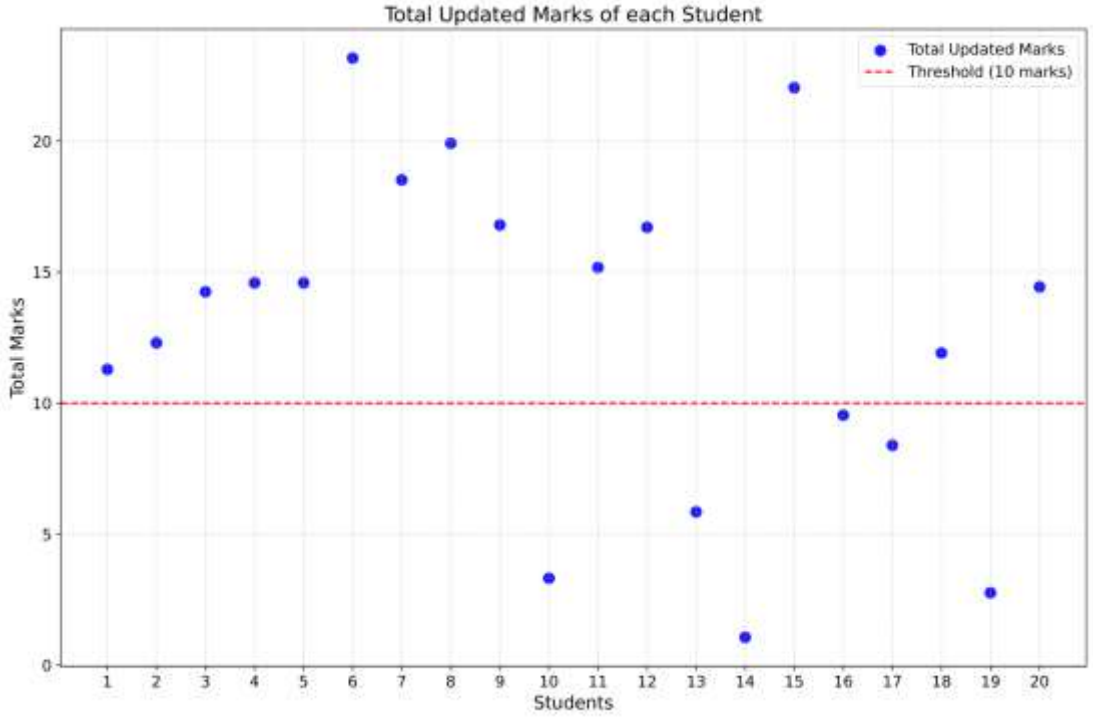

Figure 4: Advanced Weightage Impact on Students' Marks Distribution

The scatter plot in Figure 4 shows the distribution of students' marks calculated post applying an advanced weightage calculation technique. The impact of using normalized scoring along-with negative marking for incorrect answers can be seen clearly across the cohort. The plot shows a varied effect of weightage adjustments on student performances, with a reference line at 10 marks. Again, kindly take note that similar to Method 1, after applying this method the weightage of Easy questions remained the same but the weightage of Medium and Hard Questions was increased respectively, which led to the Total Marks or the Maximum Achievable Marks of the Paper to change and increase from the predetermined 100 marks. Thus, it is again important to bring to notice that this change is done only for comparison purposes and the Total Marks or the Maximum Achievable Marks of the Paper still remains at 100 Marks only.

4.4 Method 3: Bell Curve Application

The third approach consisted of a statistical model that used a normal curve distribution in allocating marks. This model tried to assess the difficulty of each question by first, finding the mean and standard deviation of the (difficulty_i) for all $\{i\}$ questions, after which the respective z-values were calculated to find out the final weightage of each question using a predefined criteria. Then, updated marks for each question were allotted based on these final weightages $\{w_i\}$, and the bell curve method was applied to the students' total updated marks. This is evidently brought out from the scatter plot and normal distribution graph of the marks allocation done by this method. The threshold line is set at 70 marks. (Figures 5 and 6).

Table 6: Summary of Normalized Weightage and Z-values for Questions

Question No.	Normalized Weightage	Z-value	Updated Z-value
1	0.90	0.33	1
2	0.65	–0.23	2
3	0.65	–0.23	2
4	0.75	0.00	2
…	…	…	
96	0.20	–1.25	3
97	0.05	–1.59	3
98	0.25	–1.14	3
99	0.15	–1.36	3
100	0.20	–1.25	3

Statistics	Value
Mean	0.44
Standard Deviation	0.26

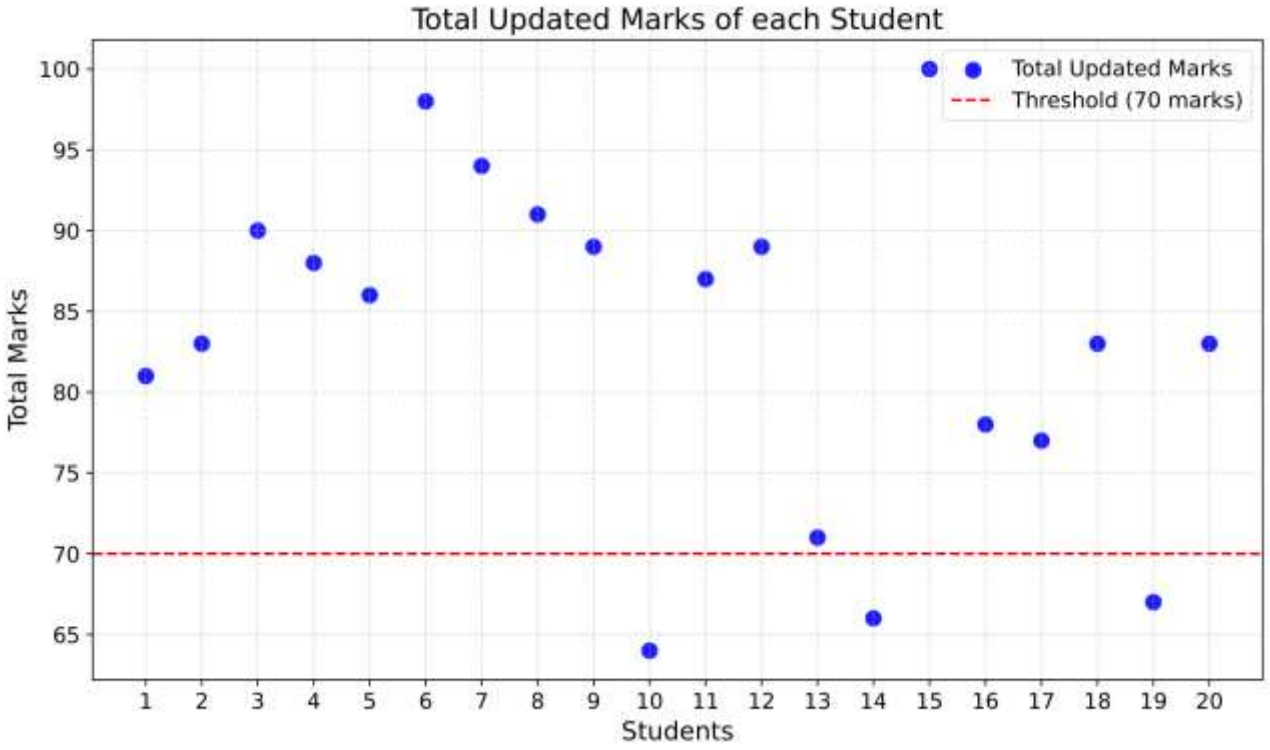

Figure 5: Distribution of the adjusted student marks by Bell curve method

The graph presented in Figure 5 shows the distribution of individual student scores after making the bell curve adjustments.

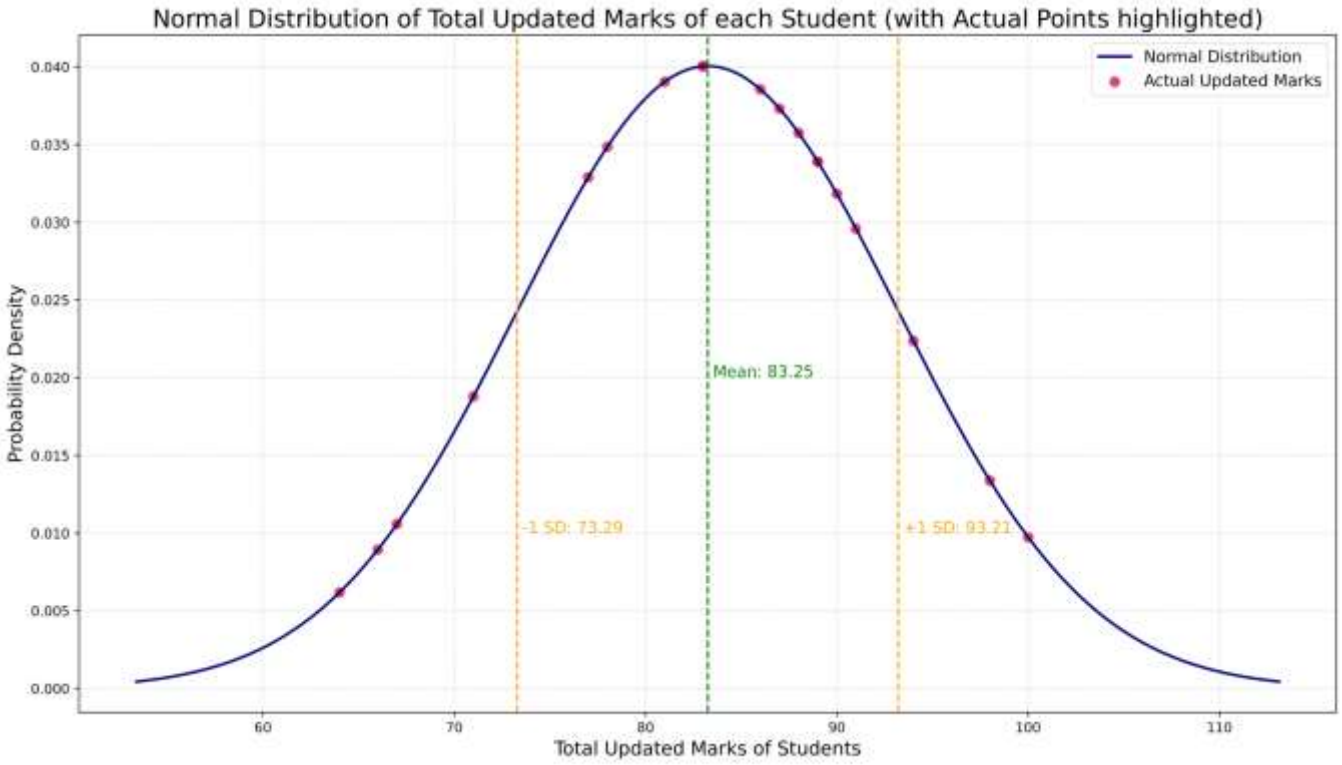

Figure 6: Normal Distribution Curve of Updated Student Marks

In Figure 6, the normal distribution curve for the updated students' marks was drawn with the mean marked at 83.25 and standard deviations at -1 SD (73.29) and +1 SD (93.21) to show how scores are spread across the distribution with respect to the mean. Updated Student Marks are overlaid to show their alignment with the theoretical distribution. Once again, similar to the previous methods used, the weightage of Easy questions remained the same but the weightage of Medium and Hard Questions was increased respectively after applying this method, which led to the Total Marks or the Maximum Achievable Marks of the Paper to change and increase from the predetermined 100 marks. Hence, it is again brought to notice that this change is done only for comparison purposes and the Total Marks or the Maximum Achievable Marks of the Paper still remains at 100 Marks only.

4.5 Question-wise Difficulty Analysis

Separate analysis was done to classify all questions into Difficulty Levels based on difficulty calculations done in Equation (1). Visual distributions using bar and pie charts have been made for showing the distribution of question difficulty into three Categories (Easy, Medium, and Hard), where it was found that statistically, many questions fell under the category of medium difficulty (Figures 7a and 7b).

Table 7: Weights and Difficulty Levels of Questions

Question No.	Weight	Adjusted Weight	Difficulty
1	0.9	1.0	Easy
2	0.65	1.0	Easy
3	0.65	1.0	Easy
4	0.75	1.0	Easy
5	0.9	1.0	Easy
... data continues up to 100 entries ...			

Table 8: Count of Items by Difficulty Level

Difficulty Level	Count
Hard	13
Medium	62
Easy	25

7a) Bar Plot of Question Difficulty Distribution 7b) Pie Chart of Question Difficulty Proportions

Figure 7: Analysis of Distribution and Proportion of Question Difficulty

Figure 7 essentially shows the questions in an exam categorized by levels of difficulty. It emphasizes the balanced nature of the Test Questions in terms of the quantity of questions of Easy, Medium, and Hard Difficulty Levels.

4.6 Key Findings

1. **Investigating the Impact of Question Difficulty:**

The main objectives were to find out how differently the level of difficulty affected student performance in the multiple-choice questions (MCQs). A comprehensive dataset was prepared using a Python script; it was basically composed of simulated responses to 20 students taking 100 MCQs with varied levels of difficulty. These simulated responses are, in turn, analyzed, showing the relationship of difficulty of questions and accuracy of students that may draw some comments. Complications may take place through simulated data, but this way, allowed a controlled testing and comparison of different strategies of evaluation. Such simulated data were created within the methodology, analyzed, and are used as a really good ground for the further investigation of how the approaches to grading vary.

1. **Advancements in Weighted Scoring Analysis:**

The study has used a weighted system of scoring to ensure that the students' scores are more effective, based on the questions' level of difficulty. A weight is assigned to each question, reflecting difficulty based on the percent of students who answered the question correctly. This was, for instance, a way to ensure each question's value in the test gets accurately represented so that it informs more about the student's performance and, eventually, the grading system. Weighted scoring allowed differences in the students' performance to be noticed, where further justification to investigate the use of normalization and penalization techniques was granted.

2. **Refinement through Normalization and Penalization**

The research delved into normalization and penalization as strategies of enhancing assessment of student achievements. It equally introduced a new grading system that would be able to normalize the weights of the questions and apply penalties for answering wrong to balance the grading scale. Designed to fairly represent the student's knowledge, this approach aimed at minimizing the effect of question difficulty on overall results by penalizing mistakes and normalizing scores appropriately. Normalized weighting and penalization, if applied, would at the least bring the result closer to a grading system of better equity and potentially a one that aligns more closely with what a student has understood and achieved as a learning outcome.

3. **Evaluation through Bell Curve Grading**

The research critically examined the application of Normal Distribution using a Bell Curve in the grading system. The grading, therefore, was scaled with a standard bell curve statistical model to normalize the students' scores against the difficulty of the questions. This approach tried to create a fair and balanced grading system in which the score distributions are lined up similar to a bell curve, leading the way for consistent and just evaluation of student performance. It is hence an outstanding example of how the study does contribute to standard performance measurement through a wider comparison of different methods allowed for grading with its impact on assessment of students.

4. **Categorization and Visualization of Question Difficulty**

A significant aspect of the research was the systematic categorization and analysis of question difficulty. Questions per se were classified into 'Easy', 'Medium' and 'Hard' Categories as per the weightage at which they were counted so that the scrutiny at question-difficulty spread could become easier within the given examination. The categorization provided relevant insights into the structure of the exam, signaling the extent to which the test successfully managed to challenge the students, even offering potential recommendations on how this design of the test can be improved. Therefore, it is important that educators prepare an assessment that is balanced in taking a true test of the student's ability and a one that prepares them for the varied academic challenges.

5 Conclusion

The current study presupposes an experimental approach to establish the way question difficulty affects student performance on a Multiple-Choice Questions (MCQ) type of test paper. The question difficulty's significance on student performance is highly established using weighted scoring methods in order to make a fairly valid student performance evaluation. The introduction of both normalization and penalization techniques was put in place to bring equality of difficulty across the examination, along with the bell curve method to standardize scores against a normalized distribution.

The research was done through the use of data simulations and hence gives no room for the real nature of student behaviors and variation in performance. This article could have been affected by some potential limitations such as the sample size being not more than 20 students. However, this helped in ensuring a focused analysis of the proposed methodologies.

Future research will need to take place in a larger, more diverse educational setting and over a longer period of time to observe how the impact of their methodologies affect the learning outcomes and student satisfaction. This also helps to understand how question difficulty relates to student involvement.

Ultimately, the findings in the study did suggest that in order to carefully align examination design and grading to educational policy, there may be an urgent need for balanced assessments representing skills and knowledge of the students appropriately. And the teachers, of course, need to be capable of developing not only the assessments but also the appropriate grading scales for the different levels of difficulty of all those exams and also, develop the teaching strategies to assist students in being successful at all those different levels. This study may also create resources to assist in deciding the examination difficulty and student preparation effectively.

6 References

Baldwin, B. A. (1984). The role of difficulty and discrimination in constructing multiple-choice examinations: With guidelines for practical application. *Journal of Accounting Education, 2*(1), 19-28.

Bandaranayake, R.C. (2008). Setting and maintaining standards in multiple choice examinations: AMEE guide no. 37. *Medical Teacher, 30*(9-10), 836-845, DOI: 10.1080/01421590802204247.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory (Vol. 2)*. Sage.

Kar, S.S., Lakshminarayanan, S., & Mahalakshmy, T. (2015). Basic principles of constructing multiple choice questions. *Indian Journal of Community and Family Medicine, 1*(2), 65-69. Doi:10.4103/2395-2113.251640

Karelia, B.N., Pillai, A., & Vegada, B.N. (2013). The levels of difficulty and discrimination indices and relationship between them in four-response type multiple choice questions of pharmacology summative tests of year II MBBS students. *IeJSME, 7*(2), 41-46.

Sachdev, R. (2019). Measures of difficulty index of multiple sets of multiple choice questionnaire. International Journal of Science, Engineering and Computer Technology, 9(1-4), 21-23.

Sevenair, J.P., & Burkett, A.R. (1988). Difficulty and discrimination of multiple choice questions: A counterintuitive result. *Journal of Chemical Education, 65*(5), 44-45.

Towns, M.H. (2014). Guide to developing high-quality, reliable, and valid multiple-choice assessments. *Journal of Chemical Education, 91*(9), 1426-1431. Doi:10.1021/ed500076x.

Zainudin, S., Ahmad, K., Ali, N. M., & Zainal, N. F. A. (2012). Determining course outcomes achievement through examination difficulty index measurement. *Procedia-Social and Behavioral Sciences, 59*, 270-276.

Angoff, W. H. (1984). Scales, norms, and equivalent scores. *Educational Testing Service*.

Andrew, B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. *Educational Psychology Measurement, 36*, 45-50.

Bandaranayake, R. (2000). Content expertise of Angoff judges. Proceedings of the Ninth International Ottawa Conference on Medical Education., Cape Town, 1-3 March 2000.

Ben-David, M. F. (2000). Standard setting in student assessment. *Medical Teacher, 22*(2), 120-130, (AMEE Guide No. 18).

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56*, 137-172.

Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement, 21*, 147-152.

Boulet, J. R., de Champlain, A. F., & McKinley, D. W. (2003). Setting defensible performance standards on OSCEs and standardized patient examinations. *Medical Teacher, 25*, 245-249.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice, 11*(4), 27-34.

Chambers, K. A., Boulet, J. R., & Gary, N. E. (2000). The management of patient encounter time in a high-stakes assessment using standardized patients. *Medical Education, 34*(10), 813-817.

Chinn, R. N., & Horitz, N. R. (2002). Alternative approaches to standard setting for licensing and certification examinations. *Applied Measurement in Education, 15*(1), 1-14.

Dauphinee, W. D., Blackmore, D., Smee, S., Rothman, A. I., & Reznick, R. (1997). Using the judgments of physician examiners in setting the standards for a national multi-center high stakes OSCE. *Advances in Health Sciences Education, 2*, 201-211.

De Champlain, A. F., Margolis, M. J., MacMillan, M. K., & Kiss, D. J. (2001). Predicting mastery level on a large-scale standardized patient test: a comparison of case and instrument score-based models using discriminant function analysis. *Advances in Health Sciences Education, 6*(2), 151-158.

Floreck, L. M., & De Champlain, A. F. (2001). Assessing sources of score variability in a multisite medical performance assessment: an application of hierarchical linear modeling. *Academic Medicine, 76*(10 Suppl.), S93—S95.

Friedman Ben-David, M. (2000). AMEE Guide No. 18: Standard setting in student assessment. *Medical Teacher, 22*(2), 120-130.

Hodges, B., McNaughton, N., Regehr, G., Tiberius, R., & Hanson, M. (2002). The challenge of creating new OSCE measures to capture the characteristics of expertise. *Medical Education, 36*(8), 742-748.

Case, S. M., & Swansen, D. A. (1998). Constructing Written Test Questions in the Basic Sciences. Philadelphia, PA: National Board of Medical Examiners, 3750 Market Street, PA 19104.

De Gruijter, D. (1985). Compromise models for establishing examination standards. *Journal of Educational Measurement, 22*, 263-269.

Jaeger, R. M., Mullis, I. V. S., Bourque, M. L., & Shakrani, S. (1996). Setting performance standards for performance assessment: some fundamental issues, current practice, and technical dilemmas. In: Technical Issues in Large-Scale Performance Assessment, NCES 96-802 (US Department of Education, Office of Educational Research and Improvement).

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425-461.

Clauser, B. E., Margolis, M. J., Clyman, S. G., & Ross, L. P. (1997). Development of automated scoring algorithms for complex performance assessments: A comparison of two approaches. *Journal of Educational Measurement, 34*(2), 141-161.