Original Article

Expert Prediction Versus Difficulty Index Measured by Psychometric Analysis; A Mixed Method Study Interpreted through Diagnostic Judgment by Cognitive Modeling Framework

Memoona Mansoor[1], Shazia Imran[1], Rehmah Sarfraz[1], Ali Tayyab[1]*

Islamabad Medical & Dental College, Islamabad, Pakistan.

***Corresponding Author**

Memoona Mansoor
memoona.mansoor@imdcollege.edu.pk

Received: 12th March, 2024
Revised: 15th May, 2024
Accepted: 25th May, 2024

DOI: https://doi.org/10.51846/jucmd.v3i2.3047

Abstract

Objective: The item difficulty is determined in two ways; one relies on expert judgments, and the other on psychometric analysis. This study compared item developers' perceptions of item difficulty with psychometric analysis results and explored their thought processes in categorizing items.
Methodology: This explanatory sequential mixed method study was conducted from October to December in 2022 in three phases (quantitative, qualitative, and mixed method strand). Difficulty ranking of items by 20 subject experts, for all the preclinical years' end-of-module exams was compared with that obtained by psychometric analysis from the OMR (Optical Mark Reader). Cohen's Kappa was used to check the agreement and Pearson's correlation was used to infer the correlation between the two measures (item writers' perception of item difficulty and Rightmark analysis). All the item developers (20) were interviewed through an open-ended two-item questionnaire. Interviews were recorded and transcribed. Themes and subthemes were identified from interview data through manual coding. The anonymity of the participants was maintained.
Results: A total of 1150 items from Anatomy, Physiology, Biochemistry, Pharmacology, Pathology & Forensic Medicine were compared. These items were developed by 20 content experts. There was a weak positive (r=0.11) but significant correlation (p=0.00) between faculty perception and Right mark analysis of the item difficulty. However, there was no agreement between the two measurements (Cohen's Kappa k=0.042, p=0.027). The interviews of item developers identified four major themes: Academic performance, learning habits, the content targeted, and the item's construction.
Conclusion: Experts consider contextual factors which cover content and student background, when ranking items, while psychometric analysis is based on item performance data. Thus, contextual nuances may lead to differences in judgment.

Keywords: Assessment, Expert Prediction, Dia Com Framework, Item Difficulty, Test Psychometrics

Introduction

An ideal MCQ paper requires a balance of easy, moderate, and difficult questions. Item difficulty is a psychometric concept that measures how easy or difficult a test item is to answer correctly. It is determined in two ways. One common approach to predicting item difficulty is to rely on expert judgments, where subject matter experts rate the expected difficulty of each item based on their experience and intuition. The item developers may have different reasoning for ranking the items on a scale of difficulty (easy, moderate, and difficult) and other characteristics (cognitive level). This method has no set criteria for categorizing items on the difficulty scale. Another approach is to use empirical data from test administrations, where item difficulty is estimated using psychometric analysis (statistical analysis).[1] Both methods have advantages and limitations. The literature on the comparison of the two methods reveals variable findings but emphasizes the complexities of assessing item difficulty in educational contexts. It showed a correct estimate by the faculty of less than 50% on the one hand, an underestimation of difficulty by the faculty, and an overestimation by the students on the other hand.[2,3]

In Pakistan, there is a drive to adopt MCQs as the primary assessment tool for theoretical knowledge in undergraduate healthcare education. Departments of Medical Education in almost every institute are training the faculty to develop quality MCQs and interpret the statistical analysis of difficulty and discrimination indices. The pre-testing of newly added items is not only administratively and financially costly, but can also compromise item security which increases the reliance on expert judgment of item difficulty.[4,5] However, the prediction of item difficulty is an important task for test developers and researchers, as it can inform the selection and calibration of items for different purposes and populations for example for the standard setting of pass scores. In addition to that, this skill is a part of the teachers' assessment competence which is differentiable from other teacher competencies.[6]

Loibl and colleagues termed this skill as "Diagnostic judgment" by the teachers; they proposed a framework to understand the cognitive processes causal to such decisions by the faculty, the Diagnostic Judgment by Cognitive Modeling (The DiaCom Framework).[7]

Although post-hoc report generation and interpretation are routine in institutes where MCQs are the desired assessment tool, our local literature is not only lacking in data about the comparison of item developers' judgment and the post-hoc analysis but also in explaining the cognitive processes in this regard. Therefore, comparing the faculty judgment with the values obtained from the post-test scores and understanding the experts' reasoning for this seems

pertinent. In this study, we aim to compare the item developers' perception of item difficulty with the one obtained from psychometric analysis and discover the thought process behind the ranking given by them. We believe this will inform our future decisions while designing question papers. It will be a stimulus to the developers for critical thinking allowing them to revisit their judgments while ranking the items for future implementation of standard setting for determining passing scores.

Methodology

This was an explanatory sequential mixed-method study, which was conducted after the approval of IRB-IMDC (94/IMDC/IRB-2022) at Islamabad Medical and Dental College from October to December 2022.

Phase 1 Quantitative strand

The ranking of items on difficulty and cognitive basis by test developers was done while submitting the items to the MCQ bank; all item writers submitted their items individually to the bank. The two parameters were recorded from there, of all the end-of-module papers of 1st, 2nd, and 3rd year MBBS (20 modules) in the year 2021. These were compared with the difficulty index measured through psychometric analysis (condensed report) obtained from the OMR (Optical Mark Reader) or Rightmark. The item categorization of difficulty index (DI) on psychometric analysis was easy items <0.76 DI, moderate 0.45-0.75 DI, and difficult <0.44 DI.[8]

SPSS Version 20 was used for all statistical analyses. Cohen's Kappa was used to check the agreement between item writers' perception of the item's difficulty and the Rightmark analysis of the difficulty of items (difficulty Index). Pearson's correlation was used to find out the correlation between item writers' perception of item difficulty and assigned cognitive level as well as item writers' perception of item difficulty and Rightmark analysis of the difficulty level of items.

The data from the question bank and the condensed reports were used with the permission of the Head of the Examination Department after approval from IRB.

Phase 2 Qualitative strand

All the faculty members (twenty) of the first three years of MBBS whose items (MCQ) were used in the quantitative strand were interviewed (criterion/purposive sampling) through a two-item open-ended questionnaire.

1.What criterion do you use for ranking the item as easy moderate and hard?

2.What is the basis for applying this criterion (factors influencing this reasoning)?

Cognitive pre-testing of the interview questions was done for comprehensibility and respondent difficulties. It was done through think aloud technique, probing, and debriefing. Items were replaced and modified on that basis. The participants were Assistant professors, Associate Professors, and Professors in their disciplines. All the participants had prior training for item construction and identification of item writing flaws. Written informed consent was obtained and all interviews were face-to-face by the first two authors. A concurrent probing technique was used during the interviews.

All the interviews were audio recorded and transcribed manually by the authors. No repeat interviews were done, and the average time of the interviews was 10 minutes. The transcribed files were reviewed by all the researchers individually and matched with the audio. Modifications were done through consensus. Two strategies were used for generating meaning from transcribed data:

1. Open coding of the words/phrases, counting frequencies of the repeated words/phrases, followed by axial coding by combining the codes with the constant comparative method. Clustering was done by creating categories of words/phrases with similar meanings or connotations. The themes were further divided into subthemes.

2. Plausibility was checked with the help of analytic memos created while going through the transcribed data in the form of short phrases.

The findings were then interpreted based on Diagnostic judgment by Cognitive Modeling Framework by Katharina Loibl and colleagues.[7]

The anonymity of the participants was maintained. There were no potential risks for the participants and no monetary benefit was given to them.

Phase 3 Mixed method strand

The qualitative data was interrogated again by the first two authors, to gain additional insight into the quantitative results.

Results

Quantitative strand

A total of 1150 items of the twenty item writers (from Anatomy, Physiology, Biochemistry, Pathology, Forensic Medicine, and Pharmacology) who taught in a system-based integrated curriculum in first three years of MBBS were used in the study. All items were single best MCQs.

The item writers assigned a difficulty level to the items based on their experience and perception. Rightmark (Avison Scanner AD 240) calculated the statistical difficulty index for each of the items.

The total number of items submitted by each writer and the comparison of item writer perception to the Rightmark statistical analysis is shown in Table 1. The item writers also categorized the questions according to their cognitive level (Recall or Application). Of the 1150 questions, 763 (66.30%) were labeled as "recall", while 387 (33.70%) were "application" by the item writers.

Table 1: Number of items submitted by each item writer and the comparison of item writers' perception & and Rightmark analysis of difficulty

ITEM SUBMITTER	SUBJECT	ITEM WRITER PERCEPTION (LEFT) & RIGHTMARK ANALYSIS OF DIFFICULTY (RIGHT)						
		EASY		MODERATE		DIFFICULT		TOTAL
1	Physiology	50	47	22	23	11	13	83
2	Anatomy	3	20	58	24	2	19	63
3	Biochemistry	4	20	30	11	2	5	36
4	Anatomy	38	46	66	38	6	26	110
5	Biochemistry	37	16	15	25	0	11	52
6	Physiology	14	21	32	19	3	9	49
7	Anatomy	19	31	49	28	2	11	70
8	Physiology	26	29	53	30	5	25	84
9	Anatomy	14	17	31	18	1	11	46
10	Biochemistry	12	25	54	29	0	12	66
11	Biochemistry	0	19	61	30	1	13	62
12	Physiology	17	36	51	29	11	14	79
13	Pathology	6	9	30	14	4	17	40
14	Pharmacology	3	8	17	7	0	5	20
15	Forensic Medicine	49	36	59	28	0	44	108
16	Pharmacology	2	18	37	8	0	13	39
17	Pathology	15	15	29	21	6	14	50
18	Pathology	1	11	16	5	3	4	20
19	Pharmacology	1	5	42	15	0	23	43
20	Pathology	6	9	21	18	3	3	30
TOTAL		**307**	**438**	**771**	**420**	**56**	**292**	**1150**

There was a weak positive (r=0.26) but statistically significant correlation (p = 0.000) between faculty perception of the difficulty of items and their assigned cognitive levels implying that more difficult questions were labeled as Application cognitive level.

In table 2, all of the comparisons show no agreement between the two measurements, and of all the writers, writer 12 and 17 showed the highest level of agreement.However, there was seen a weak positive but statistically significant correlation between faculty perception of difficulty of items and Rightmark analysis (r = 0.11).

Qualitative strand

Of the twenty item developers, eighteen agreed to participate in the study. There were 2 (20%) male and 16 (80%) female participants with an age range of 30-68 years. We explored the thought process of the item developers for the categorization of the items on a scale of difficulty. In response to the first question, four themes and eight subthsemes were identified. The themes were categorized under two heads; related to students and related to items (Table 3)

Two factors emerged in response to the second question: 1; the interactions of the test developers with the students in class and 2; their teaching experience. When we asked them why their item ranking did not match the post-test analysis, some attributed the students (for not studying well or being stressed during the exam); some acknowledged their errors and some claimed that they usually ranked the items correctly.

We interpreted the qualitative findings in light of the DiaCoM framework (Figure 1), which stands for Diagnostic Judgements-by Cognitive Modeling Framework, designed by Loibl and colleagues.[7]

Mixed method strand

A review of the qualitative data in the perspective of quantitative results led to the following insights:

1. The experts were too focused on didactic teaching, they thought that students learn only through lectures and student-centered learning methods are of little use.

2. Almost all experts thought that the recall questions were easy, yet they tagged the majority of the recall items as moderate.

Table 2: Agreement between Item writers' perception and Rightmark difficulty analysis

ITEM SUBMITTER	Agreement (Cohen's Kappa, k)	p
1	0.125	0.120
2	0.009	0.846
3	0.027	0.721
4	-0.030	0.642
5	0.042	0.622
6	0.035	0.728
7	-0.104	0.215
8	-0.007	0.918
9	0.108	0.870
10	0.152	0.43
11	-0.022	0.352
12	0.189	0.010
13	-0.21	0.809
14	0.066	0.538
15	0.070	0.176
16	-0.50	0.157
17	0.178	0.068
18	-0.145	0.103
19	-0.027	0.226
20	0.085	0.543

Table 3: Themes and subthemes for assigning an item easy, moderate, or difficult by the experts

Themes and subthemes	Illustrative Quotes
Theme "Academic performance"	
Subtheme "Average student "	"A moderate question is the one that an average student will be able to solve" "An easy question is the one that a below-average student can answer correctly"
Theme "Learning habits"	
Subtheme "Superficial learners"	"Most of the students study superficially to answer the MCQs"
Theme "Construction of the Item"	
Subtheme "Cognitive level" (Application or recall)	"The recall questions are easy because there is not so much mental work". "We categorize recall questions under the easy category and usually, application-based questions are categorized as moderate and difficult"
Subtheme "The phrasing of the items"	"When something is asked indirectly they are categorized as difficult"
Theme "Content"	
Subtheme "Repetition"	"Easy questions are the ones that target content repeated many times "
Subtheme "Teaching Strategy"	"Difficult ones are the ones targeting content that is left to the students for self-study. For example, some learning objectives are covered through small group discussions with students knowing that they are not that important"
Subtheme "Difficult concepts"	The difficult items have difficult concepts that are difficult to retain

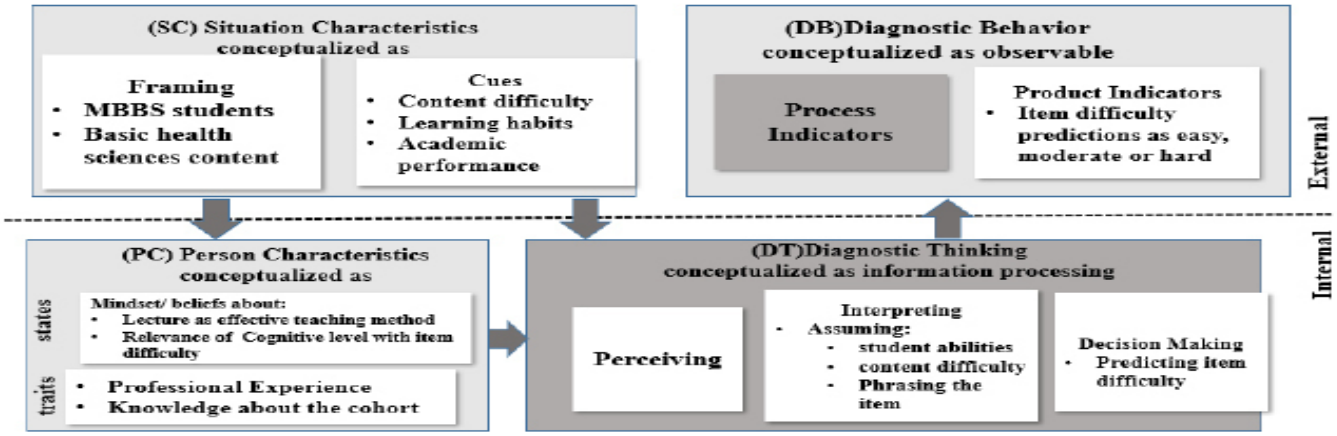

Figure 1: The DiaCom Framework

Discussion

In this explanatory sequential mixed method study, the expert prediction of item difficulty was compared with the one obtained from psychometric analysis for agreement between the two (the quantitative analysis). Later the experts were interviewed to explain the quantitative findings, which led to the unveiling of four themes (the qualitative analysis); Academic performance, learning habits, the content targeted, and the item's construction. The quantitative analysis revealed no agreement between the item writers' perceptions and the actual item difficulty. This is contrary to the expectation as the item writers are not only experts with postgraduate qualifications but with teaching experience ranging from three to thirty years. Our observation conflicted with the findings of Witat Fakcharoenphol and colleagues who had observed accurate predictions of test item difficulty by the content expert test-makers[9] and attributed this to their experience. The qualitative strand showed that they have varied reasons for their classification. It is reasonable to state that assigning the item difficulty is very subjective and it is not unusual that the two measurements do not match.[5] Urhahne and Wijnia in their meta-analysis on the accuracy of school teachers' judgment concluded that teachers' experience is only weakly associated with judgment accuracy.[10] We can also attribute these findings to the "Experts' blind spot" that leads teachers to inaccurately judge student abilities especially while assessing the difficulty of the task.[11] Interesting facts were revealed while exploring the reasoning behind their perception of an item's difficulty. It was influenced by two main factors; the characteristics of the students and the item suggesting that it is not a fixed attribute, but a relative and situational one.

Our experts compared the difficulty of an item to the expected level of performance of the students and adjusted their judgment accordingly. For example, they characterize a student as an "average performer" if his or her assessment scores are close to the mean. They would classify the questions expected to be answered correctly by such students of moderate difficulty. This assumption about the competence of the examinees while designing test items aligns with the guidelines by Gerard and Janine who argue that it is important to consider the "level of student population" while composing the assessment for a given performance standard so that it can be associated with the difficulty of the assessment. This is a usual practice in standard-setting procedures to think about an average or below-average student while assigning difficulty to test items.[3] The second theme refers to the faculty members' perception of the learning habits of their students. They generally had the view that most of the students were superficial learners. Thus while ranking they classify the items that require deeper thinking as difficult compared to the ones targeting lower thinking levels like recall and comprehension. This notion is linked to the other subtheme "cognitive level" identified under the theme "construction of the item" where the same reasoning is repeated that an item targeting a higher cognitive level like application is ranked as moderate or difficult as compared to the one targeting testing of rote memory. This was in line with the findings of the quantitative strand where a weak positive but statistically significant correlation was found between the cognitive level and difficulty of items. However, there is conflicting evidence in the comparable literature. The findings of Rush and colleagues are closer to our observations that increasing the cognitive complexity directly increased the difficulty of the items.[12] The percentage of recall (60% vs 66%) and application (30% vs 34%) items in both studies is also similar. However, the works of Pedro et al and Kibble and Johnson did not find any correlation between the cognitive level of the item and the difficulty of the item on psychometric analysis.[13,14] The difference might be due to the difference in sample size and characteristics of the cohort or the delivery of content; whatever the case maybe our quantitative results are validating the perceptions of the experts in this case. The experts also shared that phrasing of the item is important because indirectly asked questions were termed as difficult and directly asked questions as easy, by them. This premise aligns directly with the theory of affordances which implies that the way information is presented influences the way the human mind processes it.[15] Different ways of presenting the same problem lead to distinct problem-solving approaches.[15] Thus indirectly asked questions require additional cognitive processing as compared to direct ones. Analogously, they lack clear affordances, making them more challenging. Consequently, the directly asked questions are perceived as easier because the affordances guide the respondents. The most repeated word in the transcribed text was content and it emerged as a separate theme. It influenced the item writer in different ways; they thought of content in terms of its difficulty, its repetition in class, and the teaching strategy used. They based their judgment of the difficulty of items on the concepts the students struggled with[16] which they identified during formal or informal student interaction.[17] Similarly, they deemed that content or information easy that was repeated or reiterated in various academic sessions. Another interesting subtheme was the teaching strategy. Some of the experts thought that the content learned through learner-centered strategies like small group discussions and problem-based learning were difficult and the ones taught by the experts in the lectures were relatively easy. Hence, they ranked the item based on the teaching method used as well. This perception is contrary to the empirical evidence where learner-centered strategies are proven to improve students' critical thinking and application of content.[18]The researchers endorse the teachers' ability to identify learners' level of understanding and predict their performance is crucial for effective teaching. Although the research struggles to precisely measure or link it directly to better learning outcomes.[19] Application of DiaCom framework to our qualitative findings (Figure 1) depicted how Situational Characteristics (SC), such as basic health sciences context and framing biases, influence Perceptual Characteristics (PC), including mindset and professional experience. These characteristics impact Diagnostic Thinking (DT), which involves assessing situations and decision-making processes. The framework suggests that these elements collectively contribute to Diagnostic Behavior (DB), observable through various product and process indicators. This model highlights the complexity of diagnostic decision-making and the interplay between different factors affecting it.

Limitation

The limitation of our study is that it focused only on the preclin-

ical years in a single institute. Valuable insights would have been gathered by including clinical teachers in the study and analyzing the item statistics from clerkship exams.

Conclusion

The expert prediction is influenced by many factors which are usually contextual and lead to discrepancies and errors in judgments. In contrast to that the psychometric analysis is solely based on statistical properties like response patterns. Despite the soundness of reasoning, these findings raised questions about the quality of expert judgments because of the stark disagreement between the two rankings. While experts' insights are valuable and educationists suggest a combination of both perspectives (experts' opinion and psychometric history) of items for predictability, we advise otherwise and propose reliance on item history.

Implications of the study

Many variables have been suggested to explain item difficulty, but predicting item difficulty is still a difficult problem in educational assessment. This study has the potential to affect different areas; it can inform how we weigh expert predictions versus data-driven approaches (psychometric analysis) in decision-making, it can contribute to the development of AI systems combining expert knowledge with data analysis leading to more accurate predictions, information can be used to develop training programs for experts for improved judgment.

Conflict of Interest: The authors declare no conflict of interest

Disclosure: This study was presented in ICME Baku in October 2023

Funding sources: Not applicable

Authors' contribution: MM is the principal investigator who conceived the idea and collected the data (psychometric analysis and interviews), transcription and thematic analysis, and manuscript writing. SI collected the data (psychometric analysis and interviews), transcription and thematic analysis, and intellectual input to the manuscript. AT developed the Interview script, statistical analysis and interpretation, transcription and thematic analysis, and intellectual input to the manuscript. RS supervised the project, transcription and thematic analysis, and gave intellectual input to the manuscript

Consent form for the study

https://docs.google.com/document/d/10GD2OQc6GAo0WsUZ4VwF-HKD4Kq0Q9NYo/edit

References

1. Franzen D, Cuddy MM, Ilgen JS. Trusting Your Test Results: Building and Revising Multiple-Choice Examinations. Journal of Graduate Medical Education. 2018;10(3):337–338. doi.org 10.4300/JGME-D-18-00265.1

2. Kurdi G, Leo J, Matentzoglu N, Parsia B, Sattler U, Forge S, et al. A comparative study of methods for a priori prediction of MCQ difficulty. Gromann D, editor. Semantic Web. 2021;12(3):449–465. doi.org/10.3233/SW-200390

3. Van de Watering G, Van der Rijt J. Teachers' and students' perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items. Educational Research Review. 2006;1(2):133–147. doi.org/10.1016/j.edurev.2006.05.001

4. Falcão F, Costa P, Pêgo JM. Feasibility assurance: a review of automatic item generation in medical assessment. Advances in Health Sciences Education Theory and Practice. 2022;27(2):405-425. doi:10.1007/s10459-022-10092-z

5. Bramley T, Wilson F. Maintaining test standards by expert judgment of item difficulty. Research Matters: A Cambridge Assessment Publication [internet]. 2016 [date accessed-2023;21:48–54.Available from: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-matters. accessed on 15th Feb. 2024

6. Herppich S, Praetorius AK, Förster N, Glogger-Frey I, Karst K, Leutner D, et al. Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. Teaching and Teacher Education. 2018;76(4):181–193. doi.org/10.1016/j.tate.2017.12.001

7. Loibl K, Leuders T, Dörfler T. A Framework for Explaining Teachers' Diagnostic Judgements by Cognitive Modeling (DiaCoM). Teaching and Teacher Education. 2020;91(5):1-10. doi.org/10.1016/j.tate.2020.103059

8. Rezigalla AA, Eleragi AMESA, Elhussein AB, Alfaifi J, ALGhamdi MA, Al Ameer AY, et al. Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. BMC Medical Education. 2024;24(1):1–7. doi.org/10.1186/s12909-024-05433-y

9. Fakcharoenphol W, Morphew JW, Mestre JP. Judgments of physics problem difficulty among experts and novices. Physical Review Special Topics - Physics Education Research. 2015;11(2):1–14. doi.org/10.1103/PhysRevSTPER.11.020128

10. Urhahne D, Wijnia L. A review on the accuracy of teacher judgments. Educational Research Review.2021;32(4):1-26. doi.org/10.1016/j.edurev.2020.100374

11. Ostermann A, Leuders T, Nückles M. Improving the judgment of task difficulties: prospective teachers' diagnostic competence in the area of functions and graphs. Journal of Mathematics Teacher Education. 2018;21(6):579–605. doi.org 10.1007/s10857-017-9369-z

12. Rush BR, Rankin DC, White BJ. The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. BMC Medical Education. 2016;16(1):1–10. doi.org/10.1186/s12909-016-0773-3

13. Hamamoto Filho PT, Silva E, Ribeiro ZMT, Hafner M de LMB, Cecilio-Fernandes D, Bicudo AM. Relationships between Bloom's taxonomy, judges' estimation of item difficulty and psychometric properties of items from a progress test: A prospective observational study. Sao Paulo Medical Journal. 2020;138(1):33–39. doi.org/10.1590/1516-3180.2019.0459.r1.19112019

14. Kibble JD, Johnson T. Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations? American Journal of Physiology - Advances in Physiology Education. 2011;35(4):396–401. doi.org/10.1152/advan.00062.2011

15. Moon JA, Keehner M, Katz IR. Affordances of Item Formats and Their Effects on Test-Taker Cognition under Uncertainty. Educational Measurement: Issues and Practice. 2019;38(1):54–62. doi/10.1111/

emip.12229

16. Mansoor M, Aly SM, Javaid A. Effect of team-based learning on second-year students' academic performance. Journal of the College of Physicians and Surgeons Pakistan. 2019;29(9):860–864. doi.org/10.29271/jcpsp.2019.09.860

17. Mansoor M, Tayyab A, Shah SS, Sarfraz R. Threshold concepts encountered by second-year medical students in a Basic Health Science module; a qualitative study. Journal of the Pakistan Medical Association. 2022;72(5):901–907. doi.org/10.47391/jpma.3348

18. Dehghanzadeh S, Jafaraghaee F. Comparing the effects of traditional lecture and flipped classroom on nursing students' critical thinking disposition: A quasi-experimental study. Nurse Education Today. 2018;71(12):151–156. doi.org/10.1016/j.nedt.2018.09.027

19. Hill HC, Chin M. Connections Between Teachers' Knowledge of Students, Instruction, and Achievement Outcomes. American Educational Research Journal. 2018;55(5):1076–1112. doi.org/10.3102/0002831218769614