Advances in Medical Education and Practice

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Comparison of Certainty-Based Marking (CBM) and Number Right Scoring (NRS) in Multiple-Choice Question (MCQ) Assessments: A Prospective Cohort Study of Second-Year Medical Students

Farhang Rashidi[1,2], Mohammad Sina Rezaei[1], Alipasha Meysamie[3]

[1]School of Medicine, Tehran University of Medical Sciences, Tehran, Iran; [2]Iranian Center of Neurological Research (ICNR), Neuroscience Institute, Tehran University of Medical Sciences, Tehran, Iran; [3]Department of Preventive and Community Medicine, School of Medicine, Tehran University of Medical Sciences, Tehran, Iran

Correspondence: Alipasha Meysamie, Preventive and Community Medicine, Medical School, Tehran University of Medical Sciences, Tehran, Iran, Email Meysamie@gmail.com

Background: Multiple-choice questions (MCQs) with number right scoring (NRS) are now the most preferred assessment method in medical education, which requires program holders to ensure their fairness for profound decision-making and policy arrangements. Including certainty-based marking (CBM) in MCQ exams is suggested to help better distinguish students and thus increase the validity of assessments. This study aims to integrate CBM into MCQ assessments using a relatively new scoring matrix to assess pass/fail rates and exam scores.

Methods: Students in their second year of medical school participated in ten different CBM-MCQ exams. Questions were scored twice: first, traditionally using the NRS, and second, using the CBM (based on correctness and a self-report certainty scale), yielding two different final scores for each student. Exam scores and pass/fail rates were compared using paired t-tests and McNemar tests, respectively.

Results: A total of 935 students in 10 different exams were included in the study. CBM scores were significantly lower than NRS scores (0.82 points; $P<0.001$). There was a significant shift in the pass/fail classifications ($P<0.001$). Overall, 34 out of 141 NRS failures (24.1%) passed under CBM, while 85 (10.7%) of the 794 students who passed under NRS failed under CBM. This significant shift was also reported in 5 of 10 exams ($P<0.05$). Additionally, there was a trend towards worsening scores in CBM compared to NRS. Students with "A" grades decreased from 8.4% to 4.7% while students with "D" grades increased from 15.1% to 20.5%.

Conclusion: CBM yields significantly different scores compared to NRS, as evidenced by distinct pass/fail rates, suggesting the potential for a better assessment tool. Further studies with different types of assessments are needed to validate this method in terms of reducing cheating and guesswork. However, replacing CBM with conventional scoring methods requires further evidence and consideration.

Keywords: medical education, assessment, multiple choice question, MCQ, number right scoring, NRS, certainty base marking, CBM

Introduction

Profound decision-making is among the most critical skills of any physician. Considering the benefits and consequences, clinical decisions should be made with the appropriate confidence and certainty to minimize adverse outcomes and patient harm.[1–4] Clinicians should be able to learn profound decision-making during their medical training, and thus, medical schools are responsible for training doctors who are both knowledgeable and capable. They should be able to educate and assess students carefully. Assessment is a critical step in medical education that requires validity and fairness.[5] It allows instructors to provide necessary feedback to modify ongoing learning and teaching and help learners achieve their intended educational goals.[6] Assessment is shifting toward systems of assessment, rather than relying on

Received: 31 August 2025
Accepted: 15 November 2025
Published: 19 November 2025

summative/formative measurements or stand-alone evaluations.[7] There are two main types of assessment: formative, which often occurs during the course, and summative, which typically appears as final exams and usually provides learners with a numerical score and limited feedback at the end of the course.[8] To date, multiple-choice questions (MCQs) are preferred for assessing knowledge and applied knowledge, typically with positive markings and no penalties for incorrect answers, which is known as the number right scoring (NRS) method.[9–13]

Numerous studies on student assessment and MCQ exams have tried to find the most qualified types. Based on a recent study, MCQ exams have low discriminative power in high-score students.[14] Additionally, one critical challenge in MCQ exams is the presence of guesswork or partial knowledge, which means possessing some correct information about a given question but being unable to identify the correct response.[15] Partial knowledge can compromise the reliability of exams if we do not care enough.[16–18] All together, these issues risk MCQ's validity and were the initial driver of the notion of certainty-based marking (CBM).[19] Asking students to state their certainty in their answers changes the forced-choice assessment to measure self-awareness.[20] Considering the certainty level in the scoring process during examinations may better assess the actual capacity of students, and it is helpful during self-monitoring and clinical practice assessment.[21] Thus, some experts believe that program holders should integrate CBM into MCQ exams in medical schools.

Dressel and Schmid were the first to suggest certainty assessment in exams, although Gardner-Medwin was the first to officially integrate CBM into MCQs in 1994.[22,23] Many authors, including Gardner, have employed a three-level certainty rate; however, the scoring matrix has been a matter of discussion.[19,23–25] A study by Schoendorfer et al investigating the use of CBM in formative assessments in Australia revealed that students have positive attitudes toward CBM and the decision-making process. However, they still did not prefer to take the exam in CBM format, despite not counting in the final results. Additionally, they did not investigate or compare the results with the standard method.[26] In a different study by Ghadermarzi et al, using certainty-based assessment, student satisfaction was estimated to be acceptable, and certainty-based assessment was shown to be more fair, effective, authentic, and precise in estimation, as well as having higher construct validity, compared to conventional MCQ exams. They concluded that this method simulates the reflection for deeper learning among the students.[27] Evaluating the certainty of responses, along with their correctness, through CBM has been shown to accurately distinguish between partial knowledge and the need for guesswork in MCQs in similar studies.[28–30]

Although many studies have attempted to evaluate various aspects of CBM, the effect of this method on pass/fail rates has not been extensively studied recently and remains a topic of debate. In the current study, we attempted to use a relatively new scoring matrix to analyze trends in pass/fail rates and exam scores, which may provide an additional discriminating factor for distinguishing students based on their competencies. Additionally, we aimed to conduct a brief student survey to examine participants' perceptions of CBM, focusing on its alignment with the study's educational objectives, specifically self-assessment accuracy and exam-related confidence.

Materials and Methods

Study Design and Setting

The current study was designed as a quantitative, longitudinal cohort study. Ten exams were designed, each containing a different number of questions, all with four options, only one of which was correct. The total time assigned to each exam was 90 seconds per question. All questions for the ten exams differed and took place in the Preventive and Community Medicine Department at Tehran University of Medical Sciences.

To apply CBM, certainty levels are categorized as low, moderate, or high. Due to the potential need for extra time to estimate certainty, we added 15 seconds per question to the total exam time. We scored each question differently based on the correct answer and certainty levels demonstrated in Table 1. All students were clearly informed of the scoring matrix at the beginning of the semester and again before each exam, as recommended by Rippey et al.[31]

Ethical Considerations

The Ethics Committee of the Tehran University of Medical Sciences reviewed and approved the study (approval number: IR.TUMS.MEDICINE.REC.1400.345). Participation was voluntary, with no academic rewards or consequences linked to involvement. Prior to examinations, all students received a comprehensive lecture covering the study's objectives, procedures,

Table 1 Scoring Matrix

Answer	Certainty Level			No Response
	Low	Moderate	High	
Mark, if correct	+1	+2	+3	0
Penalty, if incorrect	0	−1	−2	0

confidentiality measures, and data use. Informed consent was initially obtained orally at the beginning of the semester, allowing for questions before consenting to participation. A second instance of consent was obtained at the beginning of each exam. Students were allotted a minimum of five minutes to reevaluate their participation. Data were anonymized prior to analysis to maintain participant confidentiality.

Participants

We conducted this study with different groups of second-year medical students at Tehran University of Medical Sciences, who were enrolled in the "Principles of Epidemiology" course during the fall semester of 2021 (from September 22, 2021, to January 19, 2022). We trained them all from the beginning of the semester, and informed consent was obtained through the software. Every student had the option to take the exam with a recording of certainty or in the regular form. In the window where this consent appeared before entering the exam page, extra time for the exam with certainty recording was emphasized, and the aims and study protocol were also explained. Students could choose to participate in the certainty method, agree to the terms, or participate in the regular format; however, all of the students chose the certainty method for each exam. Eventually, the results of these assessments did not contribute to progression decisions, and the information on students' responses, as well as certainty levels, was reported to each student as feedback. We kept all personal information confidential and used only exam results in the data analysis, without disclosing any details about each student. The Tehran University of Medical Sciences ethics committee approved this study by the "IR.TUMS.MEDICINE.REC.1400.345" code of ethics. All obtained data was stored on the online platform and used anonymously, without disclosure of students' identities or any other personal information. Only the three authors of the current manuscript had access to these data, which were restricted.

Assessment Platform

We designed a specific online assessment platform to meet the requirements of our study, including assessing participants' certainty levels in answering questions while addressing the security issues associated with online exams. Each student had a unique username and password. As mentioned earlier, the questions had four options; one was correct. After choosing an answer and before going to the next question, a pop-up query appears in the window asking about the certainty level with the three options mentioned earlier. According to the software design, after answering a question or running out of time for each question, there was no way to return and review it again. The order of the questions and the choices for each question were randomly generated in the software for each student in each exam. With no permission to review previously answered questions and a random order of questions and answers, the probability of cheating during the exam could be decreased. These regulations were all made in accordance with TUMS's administrative protocols for virtual examinations during the pandemic. We recorded all students' login and exam details on the server for further evaluation. All exams were taken in person, using computers at the Faculty of Medicine, under the supervision of trained staff.

Student Survey

At the end of each exam, a voluntary survey was administered to participants using the designed software, comprising 20 items that assessed students' attitudes toward this new method. A group of experts at the university's medical education center first designed this survey. Then, several other experts were asked to review the survey questionnaire for face validity and to assess the necessity of the primary version. After implementing the suggestions, a pilot study involving 26 volunteer students was conducted, and Cronbach's alpha coefficient was calculated for the final questionnaire, which exceeded 0.6, indicating reliability. The pilot's response data was not included in the study results. In total, students' responses to the survey questions

ranged from 401 to 422, with minimum and maximum response rates for each item. The response rate to the survey items was approximately 60%. Participating in this survey was not mandatory in the software.

Data Collection and Statistical Analysis

Unlike regular scores that are calculated out of 20, the range, mean, and standard deviation (SD) of the CBM scores depend on the number of questions. Therefore, the scores were computed based on the following formula on a scale of 20 (Formula 1- Calculating CBM scores).

$$\frac{(Scores\ calculated\ using\ the\ scoring\ matrix) + [2 \times (Number\ of\ Questions)]}{(Number\ of\ Questions)} \times 4$$

Thus, both scores were scaled from 0 to 20 and were comparable, considering "10" as the pass/fail threshold. This was due to the traditional passing threshold for final exams at TUMS; thus, according to university regulations, we had to apply it to the current study as well. To address this traditional limitation, grades were classified for further analysis into four categories: "A" for scores of 17≤, "B" for scores between 14≤ and <17, "C" for scores between 10≤ and <14, and "D" for scores <10.

We compared mean scores using the paired *t*-test and pass/fail and A-D grade rates using the McNemar test. Kappa statistics were used to find the agreement between the two methods. We analyzed all data using Statistical Package for Social Sciences (SPSS) version 25.0, and P values less than 0.05 were considered statistically significant.

Results

In total, during the semester, 1078 Iranian and International students took the exams. Data from students who answered more than half of the questions were used in the analysis to draw robust conclusions. Data from students who answered fewer than half of the exam or survey questions were omitted to avoid potential student ignorance and possible bias in the results. Thus, 935 students across the 10 exams were included in the data synthesis. The number of students who participated in each exam varied from 28 to 144. This was due to the different number of students in each presented class. All students of each class participated voluntarily, and no selection bias occurred.

Exam Scores

The mean scores obtained in the NRS and CBM were 13.09 (2.89) and 12.27 (2.73), respectively, and the difference was statistically significant ($P < 0.001$). Except for one exam (number 10, P=0.568), all exams showed statistically significantly lower mean scores in the CBM than in NRS. In addition, a positive and strong correlation was observed in the total ($r = 0.885$) and each of the ten exams (correlation range: 0.522–0.913) (Table 2).

Pass and Fail Rates

Out of 935 students in 10 different exams, 794 (84.4%) passed using the NRS method, while 85 (9.1%) failed using the CBM method. On the other hand, of the 141 students who failed the NRS method, 34 (3.6%) passed using the CBM method. This result was statistically significant ($P < 0.001$). Five of the 10 exams also showed statistically significant differences regarding changes in pass/fail rates (Table 3). Kappa statistics indicated positive agreement in the total (56.7%) and in each of the ten exams, except for exam number two (−5%) and exam number five (8.7%).

Grade Classifications

Grades in both CBM and NRS were categorized as mentioned in the methods above. There is an overall trend toward worsening students' grades in CBM score categorization. According to Kappa statistics, there was a positive agreement in the total (44%) and each of the ten exams, except for exam number two (−8.5%) and five (−2.2%), which is similar to previous results regarding the pass/fail analysis [(P<0.001) except for exams number two and five (P=0.618 and 0.851, respectively)]. The detailed comparison is presented in Table 4.

Table 2 Comparison of CBM and NRS Scores

Exam		Score		Difference	P value	Correlation	P value
		CBM	NRS				
Total	Number	935	935	−0.82	<0.001	0.885	<0.001
	Mean (SD)	12.27 (2.73)	13.09 (2.89)				
	Median (IQR)	12.2 (10.4–14.12)	13.33 (11.2–14.91)				
	Mode (Range)	12 (4.16–20)	12 (5.33–20)				
1	Number	129	129	−1.07	<0.001	0.934	<0.001
	Mean (SD)	11.57 (2.71)	12.64 (3.02)				
	Median (IQR)	11.47 (9.6–13.07)	12 (10.67–14.67)				
	Mode (Range)	12 (5.6–19.47)	12 (6.67–20)				
2	Number	28	28	−0.6	0.048	0.522	0.004
	Mean (SD)	9.89 (1.44)	10.49 (1.66)				
	Median (IQR)	9.76 (8.84–10.96)	10.4 (9.6–11.2)				
	Mode (Range)	9.44 (7.36–12.96)	11.2 (5.6–14.4)				
3	Number	28	28	−0.97	0.001	0.851	<0.001
	Mean (SD)	10.29 (2.59)	11.26 (2.53)				
	Median (IQR)	10.88 (8.72–11.8)	12 (9.6–13.4)				
	Mode (Range)	11.52 (4.16–15.04)	12 (6.4–15.2)				
4	Number	29	29	−1.36	<0.001	0.795	<0.001
	Mean (SD)	8.66 (2.04)	10.01 (1.94)				
	Median (IQR)	8.64 (7.52–9.68)	9.6 (8.8–11.2)				
	Mode (Range)	4.8 (4.8–12.96)	8.8 (7.2–14.4)				
5	Number	32	32	−1.28	<0.001	0.837	<0.001
	Mean (SD)	9.61 (1.81)	10.88 (1.71)				
	Median (IQR)	9.19 (8.6–10.52)	10.67 (9.95–11.62)				
	Mode (Range)	9.14 (5.24–13.62)	10.67 (7.33–14.95)				
6	Number	144	144	−1	<0.001	0.814	<0.001
	Mean (SD)	11.69 (1.84)	12.69 (2.1)				
	Median (IQR)	11.73 (10.47–13.07)	13.33 (12–14.67)				
	Mode (Range)	12 (5.6–16.27)	12 (5.33–17.33)				
7	Number	144	144	−0.68	<0.001	0.913	<0.001
	Mean (SD)	13.48 (2.18)	14.16 (2.39)				
	Median (IQR)	13.4 (12–14.8)	14 (13–16)				
	Mode (Range)	12 (6.8–20)	15 (7–20)				

(*Continued*)

Table 2 (Continued).

Exam		Score		Difference	P value	Correlation	P value
		CBM	NRS				
8	Number	140	140	−1.09	<0.001	0.843	<0.001
	Mean (SD)	13.48 (1.92)	14.57 (2.02)				
	Median (IQR)	13.41 (12.29–14.35)	14.59 (13.41–15.76)				
	Mode (Range)	14.12 (6.82–19.29)	14.59 (8.71–20)				
9	Number	129	129	−0.97	<0.001	0.912	<0.001
	Mean (SD)	11.44 (2.73)	12.4 (3.08)				
	Median (IQR)	11.2 (9.33–13.07)	12 (10.67–14.67)				
	Mode (Range)	9.33 (5.6–19.47)	12 (6.67–20)				
10	Number	132	132	0.1	0.568	0.829	<0.001
	Mean (SD)	14.17 (2.68)	14.07 (3.43)				
	Median (IQR)	14.18 (12–16)	13.09 (11.27–16.73)				
	Mode (Range)	15.27 (5.09–20)	13.09 (5.82–20)				

Table 3 Pass-Fail Rates of the Students Based on CBM and NRS

Exam	Score		CBM			P value (McNemar)	Kappa
			Pass	Fail	Total		
Total	NRS	Pass	709 (75.8%)	85 (9.1%)	794 (84.9%)	<0.001	0.567
		Fail	34 (3.6%)	107 (11.4%)	141 (15.1%)		
		Total	743 (79.5%)	192 (20.5%)	935 (100%)		
1	NRS	Pass	89 (69%)	13 (10.1%)	102 (79.1%)	0.007	0.694
		Fail	2 (1.6%)	25 (19.4%)	27 (20.9%)		
		Total	91 (70.5%)	38 (29.5%)	129 (100%)		
2	NRS	Pass	8 (28.6%)	10 (35.7%)	18 (64.3%)	0.302	−0.05
		Fail	5 (17.9%)	5 (17.9%)	10 (35.7%)		
		Total	13 (46.4%)	15 (53.6%)	28 (100%)		
3	NRS	Pass	16 (57.1%)	4 (14.3%)	20 (71.4%)	0.375	0.607
		Fail	1 (3.6%)	7 (25%)	8 (28.6%)		
		Total	17 (60.7%)	11 (39.3%)	28 (100%)		
4	NRS	Pass	6 (20.7%)	7 (24.1%)	13 (44.8%)	0.016	0.486
		Fail	0 (0%)	16 (55.2%)	16 (55.2%)		
		Total	6 (20.7%)	23 (79.3%)	29 (100%)		

(*Continued*)

Table 3 (Continued).

Exam	Score		CBM			P value (McNemar)	Kappa
			Pass	Fail	Total		
5	NRS	Pass	8 (25%)	15 (46.9%)	23 (71.9%)	0.002	0.087
		Fail	2 (6.3%)	7 (21.9%)	9 (28.1%)		
		Total	10 (31.3%)	22 (68.8%)	32 (100%)		
6	NRS	Pass	114 (79.2%)	16 (11.1%)	130 (90.3%)	0.052	0.34
		Fail	6 (4.2%)	8 (5.6%)	14 (9.7%)		
		Total	120 (83.3%)	24 (16.7%)	144 (100%)		
7	NRS	Pass	136 (94.4%)	4 (2.8%)	140 (97.2%)	0.125	0.654
		Fail	0 (0%)	4 (2.8%)	4 (2.8%)		
		Total	136 (94.4%)	8 (5.6%)	144 (100%)		
8	NRS	Pass	134 (95.7%)	1 (0.7%)	135 (96.4%)	1	0.793
		Fail	1 (0.7%)	4 (2.9%)	5 (3.6%)		
		Total	135 (96.4%)	5 (3.6%)	140 (100%)		
9	NRS	Pass	85 (65.9%)	14 (10.9%)	99 (76.7%)	0.013	0.673
		Fail	3 (2.3%)	27 (20.9%)	30 (23.3%)		
		Total	88 (68.2%)	41 (31.8%)	129 (100%)		
10	NRS	Pass	113 (85.6%)	1 (0.8%)	114 (86.4%)	0.001	0.307
		Fail	14 (10.6%)	4 (3%)	18 (13.6%)		
		Total	127 (96.2%)	5 (3.8%)	132 (100%)		

Table 4 Comparison of the Score Categories

Exam	Score		CBM					Kappa	P value
			A	B	C	D	Total		
Total	NRS	A	36 (3.9%)	41 (4.4%)	2 (0.2%)	0 (0%)	79 (8.4%)	0.44	<0.001
		B	8 (0.9%)	133 (14.2%)	147 (15.7%)	0 (0%)	288 (30.8%)		
		C	0 (0%)	26 (2.8%)	316 (33.8%)	85 (9.1%)	427 (45.7%)		
		D	0 (0%)	0 (0%)	34 (3.6%)	107 (11.4%)	141 (15.1%)		
		Total	44 (4.7%)	200 (21.4%)	499 (53.4%)	192 (20.5%)	935 (100%)		
1	NRS	A	4 (3.1%)	10 (7.8%)	0 (0%)	0 (0%)	14 (10.9%)	0.486	<0.001
		B	0 (0%)	8 (6.2%)	17 (13.2%)	0 (0%)	25 (19.4%)		
		C	0 (0%)	1 (0.8%)	49 (38%)	13 (10.1%)	63 (48.8%)		
		D	0 (0%)	0 (0%)	2 (1.6%)	25 (19.4%)	27 (20.9%)		
		Total	4 (3.1%)	19 (14.7%)	68 (52.7%)	38 (29.5%)	129 (100%)		

(*Continued*)

https://doi.org/10.2147/AMEP.S564289

Table 4 (Continued).

Exam	Score		CBM					Kappa	P value
			A	B	C	D	Total		
2	NRS	A	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	−0.085	0.618
		B	0 (0%)	0 (0%)	1 (3.6%)	0 (0%)	1 (3.6%)		
		C	0 (0%)	0 (0%)	7 (25%)	10 (35.7%)	17 (60.7%)		
		D	0 (0%)	0 (0%)	5 (17.9%)	5 (17.9%)	10 (35.7%)		
		Total	0 (0%)	0 (0%)	13 (46.4%)	15 (53.6%)	28 (100%)		
3	NRS	A	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0.536	0.001
		B	0 (0%)	1 (3.6%)	1 (3.6%)	0 (0%)	2 (7.1%)		
		C	0 (0%)	1 (3.6%)	13 (46.4%)	4 (14.3%)	18 (64.3%)		
		D	0 (0%)	0 (0%)	1 (3.6%)	7 (25%)	8 (28.6%)		
		Total	0 (0%)	2 (7.1%)	15 (53.6%)	11 (39.3%)	28 (100%)		
4	NRS	A	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0.421	0.006
		B	0 (0%)	0 (0%)	1 (3.4%)	0 (0%)	1 (3.4%)		
		C	0 (0%)	0 (0%)	5 (17.2%)	7 (24.1%)	12 (41.4%)		
		D	0 (0%)	0 (0%)	0 (0%)	16 (55.2%)	16 (55.2%)		
		Total	0 (0%)	0 (0%)	6 (20.7%)	23 (79.3%)	29 (100%)		
5	NRS	A	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	−0.022	0.851
		B	0 (0%)	0 (0%)	3 (9.4%)	0 (0%)	3 (9.4%)		
		C	0 (0%)	0 (0%)	5 (15.6%)	15 (46.9%)	20 (62.5%)		
		D	0 (0%)	0 (0%)	2 (6.3%)	7 (21.9%)	9 (28.1%)		
		Total	0 (0%)	0 (0%)	10 (31.3%)	22 (68.8%)	32 (100%)		
6	NRS	A	0 (0%)	2 (1.4%)	0 (0%)	0 (0%)	2 (1.4%)	0.272	<0.001
		B	0 (0%)	10 (6.9%)	26 (18.1%)	0 (0%)	36 (25%)		
		C	0 (0%)	1 (0.7%)	75 (52.1%)	16 (11.1%)	92 (63.9%)		
		D	0 (0%)	0 (0%)	6 (4.2%)	8 (5.6%)	14 (9.7%)		
		Total	0 (0%)	13 (9%)	107 (74.3%)	24 (16.7%)	144 (100%)		
7	NRS	A	8 (5.6%)	13 (9%)	1 (0.7%)	0 (0%)	22 (15.3%)	0.505	<0.001
		B	0 (0%)	41 (28.5%)	25 (17.4%)	0 (0%)	66 (45.8%)		
		C	0 (0%)	2 (1.4%)	46 (31.9%)	4 (2.8%)	52 (36.1%)		
		D	0 (0%)	0 (0%)	0 (0%)	4 (2.8%)	4 (2.8%)		
		Total	8 (5.6%)	56 (38.9%)	72 (50%)	8 (5.6%)	144 (100%)		

(*Continued*)

Table 4 (Continued).

Exam	Score		CBM					Kappa	P value
			A	B	C	D	Total		
8	NRS	A	5 (3.6%)	2 (1.4%)	0 (0%)	0 (0%)	7 (5%)	0.391	<0.001
		B	0 (0%)	41 (29.3%)	47 (33.6%)	0 (0%)	88 (62.9%)		
		C	0 (0%)	2 (1.4%)	37 (26.4%)	1 (0.7%)	40 (28.6%)		
		D	0 (0%)	0 (0%)	1 (0.7%)	4 (2.9%)	5 (3.6%)		
		Total	5 (3.6%)	45 (32.1%)	85 (60.7%)	5 (3.6%)	140 (100%)		
9	NRS	A	3 (2.3%)	10 (7.8%)	0 (0%)	0 (0%)	13 (10.1%)	0.488	<0.001
		B	1 (0.8%)	8 (6.2%)	14 (10.9%)	0 (0%)	23 (17.8%)		
		C	0 (0%)	1 (0.8%)	48 (37.2%)	14 (10.9%)	63 (48.8%)		
		D	0 (0%)	0 (0%)	3 (2.3%)	27 (20.9%)	30 (23.3%)		
		Total	4 (3.1%)	19 (14.7%)	65 (50.4%)	41 (31.8%)	129 (100%)		
10	NRS	A	16 (12.1%)	4 (3%)	1 (0.8%)	0 (0%)	21 (15.9%)	0.372	<0.001
		B	7 (5.3%)	24 (18.2%)	12 (9.1%)	0 (0%)	43 (32.6%)		
		C	0 (0%)	18 (13.6%)	31 (23.5%)	1 (0.8%)	50 (37.9%)		
		D	0 (0%)	0 (0%)	14 (10.6%)	4 (3%)	18 (13.6%)		
		Total	23 (17.4%)	46 (34.8%)	58 (43.9%)	5 (3.8%)	132 (100%)		

Student Survey

A total of 422 out of 935 participants answered at least one question in an anonymous online survey containing 20 items that assessed students' attitudes toward the NRS and CBM examination methods regarding the accuracy and fairness of the assessment. The majority of students were somewhat optimistic about the current study and CBM method, indicating the following agree or strongly agree with the items:

1. Importance of accurately assessing students' knowledge (85.5%)
2. Reflecting students' competencies in their exam scores (88.6%)
3. The necessity of using novel methods in assessment (74.1%)
4. Increasing the quality of the study by using novel assessment methods (72.1%)
5. Correlation of the certainty level in the answers to the extent of knowledge (72.0%)
6. Correlation of uncertainty to the guesswork in MCQs exams (65.6%)
7. The importance of assessing the certainty levels of students in MCQs exams (63.5%)
8. Dependence of higher scores on students' eligibility (51.2%)
9. Simplicity of determination of certainty level in the exams (59.2%)
10. Fairness of CBM in final scores of MCQ exams (48.8%)

The majority of students' opinions, as agree or strongly agree, about the scoring matrix in CBM were the following:

1. Assigning no penalties to incorrect high-, moderate-, and low-certainty answers (67.1%, 77.2%, and 85.5%, respectively)
2. Assigning one positive mark to correct answers with low certainty (56.5%)

3. Assigning two positive marks to correct answers with moderate certainty (38.5%)
4. Assigning three positive marks to correct answers with high certainty (53.8%)

Discussion

In the current study, we scored answers based on students' self-report certainty levels using a relatively novel scoring matrix. Although several studies have investigated CBM and its outcomes, the effect on pass/fail classification has received little attention. Unlike a recent study,[32] CBM scores were 0.82 points lower than NRS scores in total, and this difference was statistically significant ($P < 0.001$). Similar results were observed in nine out of ten exams, with only one exam showing a CBM score that was better than the NRS score, but the difference was insignificant. These differences could be attributed to the students' varying levels of uncertainty about their answers. However, despite these significant differences, their importance and indications are still debated, as the differences are mostly less than a single point on a scale of 20. As previously mentioned in the results, there was a strong, positive correlation in the total and in each of the ten exams ($r > 0.8$), except for exam number two, which may be due to the small number of participants and possibly greater student uncertainty when answering the questions.

In addition, the findings suggest that CBM better distinguished the students' pass/fail rates over NRS ($P < 0.001$). The agreement was positive overall and for each of the exams, except for exam number two. As mentioned previously, this might be possible due to the limited number of participants and the students' uncertainty in answering the questions. In addition, exam five had a relatively low Kappa, unlike the other exams. Similar results were observed when categorizing scores into A-D, with a positive, significant Kappa ($P < 0.001$), except for exams 2 and 5, which were negative and insignificant. This trend further supports the previous results, suggesting possible cheating and guesswork in MCQ exams. These aligned results highlight the nature of CBM results, which are independent and tend to favor large-scale competitive examinations.

Although these results cannot be robustly interpreted as an advantage of CBM, since no gold-standard assessment tool was used, our analyses suggest that, in standard exams, some students who pass may rely on guesswork and may not know the correct answer. On the other hand, some students who fail probably have the necessary qualifications to pass. Therefore, if we mark them based on certainty levels, it appears that more qualified students tend to pass, while those who might not have the correct answer and try to guess tend to fail more often. However, to prove this, we will need at least another assessment of the students' competencies.

We also attempted to minimize guesswork for students by assigning no negative penalties for incorrect answers with low certainty. Otherwise, choosing answers randomly would be irrational, as it would result in negative marks. Smrkolj et al applied this idea, which was absent in several other CBM studies.[32] For instance, Wu et al assign no positive or negative marks for questions without certainty; we believe this might lead to leaving questions blank if candidates do not know the answers for sure.[33] Gardner-Medwin also suggested this idea in a chapter in Bryan's book. However, his scoring matrix contained more negative penalties for incorrect high-certainty answers than positive marks for correct high-certainty answers, which we believe can lead students to avoid high-certainty levels and make our methodology in the current study[34] uncertain. A similar strategy was deployed in a study by Barr et al,[34] who assigned no negative marks to incorrect low-confidence answers but two and four negative penalties for medium- and high-confidence wrong answers, respectively. They concluded that the correlation between the number of correct answers and the CBM score (high, medium, low) was significant. We believe this correlation is mainly due to the exponentially high negative penalties, which may result in overly conservative confidence choices, leading to bias.

In a similar study in 2013, Curtis et al categorized incorrect answers into two groups: misinformed and uninformed, corresponding to high- and low-confidence responses, respectively. The results indicated that students were more misinformed than uninformed (22% versus 8%), and misinformed responses were more common in complex than factual questions.[20] Similar results were obtained by Tweed et al, who found that incorrect answers with high certainty were more likely to be unsafe in real-world practice than those with low or moderate certainty. They also indicated a positive correlation between high certainty and correct answers.[35] Tabibzadeh et al agreed with previous results that higher-performing students were more confident in their answers and more efficient at providing correct answers than lower-performing students. However, their efficacy was worse in incorrect answers. In addition, male students tend to be less confident.[21]

In several studies, CBM has been used as a formative assessment to evaluate learning and teaching, and has been found beneficial for student-instructor interaction. The participants engaged more deeply with the quality of the course.[36,37] However, in the current study, students were less likely to participate in CBM and considered it too strict, according to the student survey. Similar results were reported in previous studies, suggesting that program holders must better demonstrate the necessity and benefits of innovative assessment methods in medical schools in the future.[26,38]

CBM and the scoring matrix used in the current study can be applied to other exam formats, such as practical or structured written assessments, with some adaptations. For instance, in practical exams, examiners could define some "must not answer" questions, resulting in greater negative marks due to the potential dangers in clinical care. Similar methods could be implemented in other exam formats.

Additionally, students' perceptions in this cohort offer critical insights into the feasibility of large-scale implementation of CBM. Students found CBM acceptable when its rationale and scoring criteria were clearly articulated, and when practice opportunities were offered before the summative assessment. Several perceived benefits were identified: CBM promoted reflection on confidence calibration rather than default guessing; it rendered uncertainty explicit, which students noted was more aligned with clinical decision-making; and it provided a transparent framework that rewarded well-justified answers without disproportionately penalizing honest uncertainty. Students appreciated the direct interpretability of certainty levels while assessing their performance, which aligns with a recent study.[39]

Concerns focused on equity and cognitive demand. Certain students expressed concern that miscalibrated confidence might lead to lower scores despite sufficient content knowledge, particularly during the initial stages of learning. Participants reported heightened time pressure due to the additional requirement of selecting a certainty level. At the same time, a small subset voiced concerns about the risk of "double jeopardy", which involves the possibility of being both incorrect and overly confident. The observed concerns lessened with increased exposure and a focused orientation, indicating that the quality of implementation, especially in briefing, exemplars, and formative practice, is crucial.

Collectively, these perceptions indicate that adoption is viable under specific conditions; preparation involves concise, structured training that includes worked examples and low-stakes practice; clarity requires simple, prominently displayed rules and consequences associated with certainty options; support entails providing feedback that indicates where confidence was accurately or inaccurately calibrated; and gradual roll-out consists of initial formative use or partial summative weighting to alleviate transition anxiety. Institutions must monitor for unintended effects, such as systematic under- or over-confidence, and adjust guidance as necessary. In this context, student feedback indicates that CBM serves as a valid and educationally significant enhancement to MCQ assessment, contingent upon intentional implementation and adequate support.

Limitations and Future Directions

Despite using a relatively novel methodology and scoring matrix, the present study has several limitations. Initially, despite comprising 10 MCQ examinations, variations in exam content, topic coverage, and difficulty may have contributed to the differences observed between CBM and NRS. This heterogeneity, along with the lack of power in some subgroup analyses, although indicative of actual assessment conditions, may diminish the comparability of outcomes at the exam level. The study was conducted within a single cohort of second-year medical students at a single institution, potentially limiting the generalizability of the findings to other contexts and educational settings. This study evaluated performance solely through written multiple-choice questions, omitting assessment of other competencies, including clinical reasoning and practical skills.

Also, differences in scores can be attributed to baseline knowledge of the participating students. Since we had no basic information (such as GPAs or scores on their comprehensive exams), we could not compare their baseline knowledge or randomly allocate them. Additionally, the traditional scoring system, which uses a 20-point scale, is another limitation in the current study. To address this issue, we categorized scores. Furthermore, there was limited data on the detailed certainty levels and answer rates of questions, which, separately, prevented us from conducting more detailed and complex analyses.

Future studies with larger samples and different question types, rather than MCQs, are needed to evaluate CBM's validity, reliability, discrimination power, and its correlation with higher-performing students, and to identify the optimal scoring matrix for CBM. CBM also provides a clearer view of the pass/fail threshold, which warrants further investigation in the future, at the very least by considering an assessment of competencies for comparison.

Conclusion

CBM yielded superior mean scores, pass rates, and discrimination indices relative to traditional NRS in the present study. Our findings suggest that integrating certainty judgments into MCQ assessments could improve the diagnostic value and interpretive depth of exam results by incentivizing well-calibrated confidence and deterring random guessing. Nonetheless, due to the variability noted across examinations and the single-institution framework, the results warrant careful interpretation. Additional research, preferably incorporating multiple institutions and comparisons with external competency measures (eg, OSCEs or standardized progress tests), is necessary to validate the effectiveness and generalizability of CBM in medical education. In addition, examiners can utilize the platform designed for the current study in various sections of university assessments, including class quizzes and final exams. Even comprehensive exams, such as University Entrance Exams and USMLE steps, could be taken for credit in the future. However, replacing CBM with conventional scoring methods requires further evidence and consideration.

Data Sharing Statement

The online assessment platform, exam questions, and student survey used in the current study are available from the first and corresponding author upon reasonable request.

Acknowledgments

The authors would like to thank all second-year medical students at Tehran University of Medical Sciences in 2021 who participated in this research.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This study was supported by a grant from Tehran University of Medical Sciences/School of Medicine (grant number 1400.52557). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Disclosure

All authors certify that they have no conflicts of interest.

References

1. Williams BW. The prevalence and special educational requirements of dyscompetent physicians. *J Contin Educ Health Prof*. 2006;26(3):173–191. doi:10.1002/chp.68
2. Hays RB, Jolly BC, Caldon LJ, et al. Is insight important? Measuring capacity to change performance. *Med Educ*. 2002;36(10):965–971. doi:10.1046/j.1365-2923.2002.01317.x
3. Hardy D, Smith BE. Decision making in clinical practice. *Br J Anaesth Recovery Nurs*. 2008;9:19–21. doi:10.1017/S1742645608000028
4. Hajjaj FM, Salek MS, Basra MK, Finlay AY. Non-clinical influences on clinical decision-making: a major challenge to evidence-based practice. *J R Soc Med*. 2010;103(5):178–187. doi:10.1258/jrsm.2010.100104
5. Kinnear B, St-Onge C, Schumacher DJ, Marceau M, Naidu T. Validity in the next era of assessment: consequences, social impact, and equity. *Perspect Med Educ*. 2024;13(1):452–459. doi:10.5334/pme.1150
6. Wilson M, Sloane K. From principles to practice: an embedded assessment system. *Appl Meas Educ*. 2000;13(2):181–208. doi:10.1207/S15324818AME1302_4
7. Schuwirth LWT, van der Vleuten CPM. A history of assessment in medical education. *Adv Health Sci Educ*. 2020;25(5):1045–1056. doi:10.1007/s10459-020-10003-0
8. Glazer N. Formative plus summative assessment in large undergraduate courses: why both? *Int J Teach Learn High Educ*. 2014;26:276–286.
9. Epstein RM. Assessment in medical education. *N Engl J Med*. 2007;356(4):387–396. doi:10.1056/NEJMra054784
10. Vanderbilt A, Feldman M, Wood I. Assessment in undergraduate medical education: a review of course exams. *Med Educ Online*. 2013;18(1):20438. doi:10.3402/meo.v18i0.20438

11. Tabish SA. Assessment methods in medical education. *Int J Health Sci*. 2008;2(2):3–7.
12. Considine J, Botti M, Thomas S. Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian*. 2005;12(1):19–24. doi:10.1016/S1322-7696(08)60478-3
13. Kilgour JM, Tayyaba S. An investigation into the optimal number of distractors in single-best answer exams. *Adv Health Sci Educ Theory Pract*. 2016;21(3):571–585. doi:10.1007/s10459-015-9652-7
14. Iñarrairaegui M, Fernández-Ros N, Lucena F, et al. Evaluation of the quality of multiple-choice questions according to the students' academic level. *BMC Med Educ*. 2022;22(1):779. doi:10.1186/s12909-022-03844-3
15. Coombs CH, Milholland JE, Womer FB. The assessment of partial knowledge. *Educ Psychol Meas*. 1956;16(1):13–37. doi:10.1177/001316445601600102
16. Burton RF. Quantifying the effects of chance in multiple choice and true/false tests: question selection and guessing of answers. *Assess Eval Higher Educ*. 2001;26:41–50. doi:10.1080/02602930020022273
17. Frary RB. Formula scoring of multiple-choice tests (Correction for Guessing). *Educ Meas*. 1988;7(2):33–38. doi:10.1111/j.1745-3992.1988.tb00434.x
18. Golvardi Yazdi MS, Haghighat Shoar SM, Sobhani G, et al. Factors affecting students' guesswork in multiple choice questions and corrective strategies. *Med Educ Bull*. 2021;2(4):297–305.
19. Hevner K. A method of correcting for guessing in true-false tests and empirical evidence in support of it. *J Soc Psychol*. 1932;3:359–362. doi:10.1080/00224545.1932.9919159
20. Curtis DA, Lind SL, Boscardin CK, Dellinges M. Does student confidence on multiple-choice question assessments provide useful information? *Med Educ*. 2013;47(6):578–584. doi:10.1111/medu.12147
21. Tabibzadeh N, Mullaert J, Zafrani L, et al. Knowledge self-monitoring, efficiency, and determinants of self-confidence statement in multiple choice questions in medical students. *BMC Med Educ*. 2020;20(1):445. doi:10.1186/s12909-020-02352-6
22. Dressel PL, Schmid J. Some modifications of the multiple-choice item. *Educ Psychol Meas*. 1953;13(4):574–595. doi:10.1177/001316445301300404
23. Gardner-Medwin AR. Optimisation of certainty-based assessment scores. 2013 [cited September 4, 2023]. Available from: https://tmedwin.net/~ucgbarg/tea/IUPS_2013a.pdf. Accessed November 17, 2025.
24. Little E, Creaser J. Uncertain responses on multiple-choice examinations. *Psychol Rep*. 1966;18(3):801–802. doi:10.2466/pr0.1966.18.3.801
25. Abu-Ghazalah RM, Dubins DN, Poon GMK. Dissecting knowledge, guessing, and blunder in multiple choice assessments. *Appl Meas Educ*. 2023;36(1):80–98. doi:10.1080/08957347.2023.2172017
26. Schoendorfer N, Emmett D. Use of certainty-based marking in a second-year medical student cohort: a pilot study. *Adv Med Educ Pract*. 2012;3:139–143. doi:10.2147/AMEP.S35972
27. Ghadermarzi M, Yazdani S, Pooladi A, Bahram-Rezaei M, Hosseini F. A comparative study between the conventional MCQ scores and MCQ with the CBA scores at the standardized clinical knowledge exam for clinical medical students. *J Med Educ*. 2015;14:6–12.
28. Bush M. Reducing the need for guesswork in multiple-choice tests. *Assess Eval Higher Educ*. 2015;40(2):218–231. doi:10.1080/02602938.2014.902192
29. Snow E. *Effects of a Confidence-Based, Individualized Remediation Strategy on Student Learning and Final Grades in a Multi-Campus Human Anatomy Curriculum* [Theses and Dissertations]. 2019.
30. Snow E, Brown M. Frequency of unconventional certainty-based remediation performances on high-stakes OT anatomy examinations. Medical Science Educator: Oral Presentation Abstracts, 25th Annual Meeting of the International Association of Medical Science Educators; 2021:111–151.
31. Rippey RM, Voytovich AE. Adjusting confidence tests for realism: the favorable consequences. *Eval Health Prof*. 1982;5(1):71–85. doi:10.1177/016327878200500105
32. Smrkolj Š, Bančov E, Smrkolj V. The reliability and medical students' appreciation of certainty-based marking. *Int J Environ Res Public Health*. 2022;19(3):1706. doi:10.3390/ijerph19031706
33. Wu Q, Vanerum M, Agten A, et al. Certainty-based marking on multiple-choice items: psychometrics meets decision theory. *Psychometrika*. 2021;86(2):518–543. doi:10.1007/s11336-021-09759-0
34. Bryan C, Clegg K, editors. *Innovative Assessment in Higher Education: A Handbook for Academic Practitioners*. 2nd ed. Routledge; 2019.
35. Tweed MJ, Stein S, Wilkinson TJ, Purdie G, Smith J. Certainty and safe consequence responses provide additional information from multiple choice question assessments. *BMC Med Educ*. 2017;17(1):106. doi:10.1186/s12909-017-0942-z
36. Luetsch K, Burrows J. Certainty rating in pre-and post-tests of study modules in an online clinical pharmacy course - A pilot study to evaluate teaching and learning. *BMC Med Educ*. 2016;16(1):267. doi:10.1186/s12909-016-0783-1
37. Hendriks WJAJ, Bakker N, Pluk H, et al. Certainty-based marking in a formative assessment improves student course appreciation but not summative examination scores. *BMC Med Educ*. 2019;19(1):178. doi:10.1186/s12909-019-1610-2
38. Blanar V, Pospichal J. I Know that I Know: a certainty based marking tests designed for evaluating knowledge of healthcare students (English and Czech version of the article). *Profese Online*. 2016;9:1–8. doi:10.5507/pol.2016.001
39. Suryavanshi C, Nayak KR. Certainty-based marking in multiple-choice assessments in physiology: a web-based implementation using an AI assistant. *Adv Physiol Educ*. 2025. doi:10.1152/advan.00087.2025