BMC Medical Education

RESEARCH Open Access

Experts' prediction of item difficulty of multiple-choice questions in the Ethiopian Undergraduate Medicine Licensure Examination

Shewatatek Gedamu Wonde[1,2*], Tefera Tadesse[3,4], Belay Moges[5] and Stefan K. Schauber[6]

Abstract

Background The ability of an expert's item difficulty ratings to predict test-taker actual performance is an important aspect of licensure examinations. Expert judgment is used as a primary source of information for users to make prior decisions to determine the pass rate of test takers. The nature of raters involved in predicting item difficulty is central to set credible standards. Therefore, this study aimed to assess and compare raters' prediction and actual Multiple-Choice Questions' difficulty of the undergraduate medicine licensure examination (UGMLE) in Ethiopia.

Method 815 examinees' responses to 200 Multiple-Choice Questions (MCQs) were used in this study. The study also included experts' item difficulty ratings of seven physicians who participated in the standard settings of UGMLE. Then, analysis was conducted to understand experts' rating variation in predicting the actual difficulty levels of examinees. Descriptive statistics was used to profile the mean rater's and actual difficulty value for MCQs, and ANOVA was used to compare the mean differences between raters' prediction of item difficulty. Additionally, regression analysis was used to understand the interrater variations in item difficulty predictions compared to the actual difficulty. The proportion of variance of actual difficulty explained from rater prediction was computed using regression analysis.

Results In this study, the mean difference between raters' prediction and examinees' actual performance was inconsistent across the exam domains. The study revealed a statistically significant strong positive correlation between the actual and predicted item difficulty in exam domains eight and eleven. However, a non-statistically significant very weak positive correlation was reported in exam domains seven and twelve. The multiple comparison analysis showed significant differences in mean item difficulty ratings between raters. In the regression analysis, experts' item difficulty ratings of the UGMLE had 33% power in predicting the actual difficulty level. The regression model also showed a moderate positive correlation ($R = 0.57$) that was statistically significant at F (6, 193) = 15.58, $P = 0.001$.

Conclusion This study demonstrated the complex process for assessing the difficulty level of MCQs in the UGMLE and emphasized the benefits of using experts' ratings in advance. To ensure the exams maintain the necessary reliable

*Correspondence:
Shewatatek Gedamu Wonde
gedamuwonde@gmail.com; shewataw@uio.no

Full list of author information is available at the end of the article

and valid scores, raters' accuracy on the UGMLE must be improved. To achieve this, techniques that align with the evolving assessment methodologies must be developed.

Keywords Licensure examination, Undergraduate medicine, Expert judgment, Ethiopia

Background

To increase patient safety and public trust, medical graduate licensure examinations aimed to assess physicians' minimal performance believed to be necessary to practice medicine safely and effectively [1–3]. In this regard, the assessment of a graduate's competency level with a defensible licensing examination and its valid exam scores from a well-defined standard-setting procedure has become the prerequisite for healthcare practices [4].

The validity of professional readiness examination scores mostly depends on the cut-off score determined by experts' judgments on each item, and a full exam against the predicted performance of exam takers' [5, 6]. To this end, expert judgments should be made with great emphasis on a reasonable degree of predicting the actual scores of examinees. Therefore, considering the exam item's complexity against the expected test-taker performance has become a mandatory prior judgments [7].

Generating evidence about the level of item difficulty in licensure examinations has become a challenging task for raters because of the need for accurate and reliable information [8]. Therefore, the item difficulty prediction ability of experts is an important aspect of licensure examinations and depends on expert's knowledge, skill, and professional passion. This is because expert judgment are used as the primary source of information for predicting examinees' actual performance in making decisions congruent with the intended purpose of patient safety and public trust [9–11]. Studies also showed that raters' variation of examinees' performance prediction was due to their differences in educational background, professional role, socioeconomic status, knowledge, and experience on standard-setting [12–14].

A significant number of studies have been conducted to explore the consistency among experts' ratings and its alignment with the actual examinees' performance in line with the intended purpose of the examination [15]. For instance, studies have shown variations in raters' judgment accuracy due to the difference in their experience and skill in item difficulty ratings [10, 16]. A study also conducted in Taiwan on registered Nurse Licensure exam has shown a high frequency of inconsistency between examinees' actual performance and expert item difficulty ratings [17]. However, a research finding has also shown agreement between raters' prediction of item difficulty and examinees' actual performance scores [18].

Therefore, this study aimed to explore experts' prediction of MCQ item difficulty as compared to the examinees' actual difficulty score of Undergraduate Medical Licensure Examination (UGMLE) in Ethiopia. Specifically, this study aimed to answer the following research questions.

1. How much do the experts' ratings of the MCQs' difficulty levels agree on Undergraduate Medical Licensure Examination in Ethiopia?
2. Do experts' rating of the MCQ items on the Undergraduate Medical Licensure Examination in Ethiopia predict the actual difficulty?

Methods and study design

Study design

This study used a cross-sectional survey design to explore experts' prediction of item difficulty in the UGLME. This study aimed to assess and compare raters' prediction of item difficulty and examinees' actual performance in UGMLE.

Study participant selection and data sources

The study data comprise seven physicians item difficulty ratings of 200 MCQ. These physicians were selected purposefully from different clinical specialties for standard settings of UGMLE. They were selected based on the national licensure examination criteria, which stated that experts involved in item rating for standard setting should have an MSc or clinical specialty with a minimum of 4 years of higher education teaching experience [19]. Additional data was collected from exam booklets of 815 (100%) examinees of UGMLE. The licensure examination examinees' data and rater prediction of item difficulty values were obtained from the Federal Ministry of Health, health professional competency assessment, and licensing directorate exam database.

Data analysis

200 MCQs' actual difficulty values were computed from the existing dataset and compared against the seven raters' difficulty ratings. The variation between the rater prediction and the actual item difficulty values was computed to assess the effect of the experts' judgment on the standard setting of the UGMLE cut-off score. The mean difference of the rater's prediction across exam domains was compared against the actual item difficulty to identify the effect of prediction on actual performance. The overall rater prediction agreement and the mean differences between raters were also computed in this study

to assess the quality of the experts' judgments in the UGMLE.

The statistical analysis was performed using SPSS version 27.0 (IBM Corp, Armonk, NY, United States of America), and descriptive statistics were presented as the frequency, mean, and standard error (mean ± SE). The results of this study findings are presented in tables. Regression analysis was performed to measure the linear relationship between raters' item difficulty rating in the prediction of the actual difficulty across exam domains. One-way ANOVA was used to examine the mean difference between rater's ratings in the UGMLE. $P<0.05$ was considered to indicate statistical significance.

Ethical considerations

The study protocol was approved by the Jimma University, Health Institute Research Review Board (Reference No. IHRPG/868/20) and the Norwegian Data Protection Authority (Reference number 321099) for the use of UGMLE data for research. All study participants' data and examinees' identifiers were mentioned anonymously for confidentiality.

Results

This study presents the findings in four sections. Section 1 presented the overall descriptive statistics and the study participants. Section 2 the proportion of the variance in the dependent variable predicted from independent variables. Section 3 examines the comparison between inter-rater difficulty and actula difficulty correlations. Finally, Section 4 summarizes the result of the regression analysis on the raters' ability to predict the actual difficulty levels.

Study participant profile

This study analysed the actual difficulty of 200 MCQ and item difficulty ratings of seven experts of the UGMLE. These seven experts were carefully chosen based on their specialization and years of teaching experience. These experts had five to twelve years of teaching experience and two to four years of taking part in the UGMLE process. However, out of the seven experts, rater four was excluded in subsequent analysis after being presented in the descriptive result table due to his contribution only in ratings of 30 MCQs out of 200 MCQs of the UGLME. Therefore, six raters and the score-recorded data from 815 examinees were used to compute the statistical analyses of this study.

Table 1 Descriptive statistics of the mean experts' difficulty rating and the actual difficulty values in the UGMLE

Variable	*N*	$\bar{X}$	SD	Min	Max
Rater one	200	59.39	8.549	30	80
Rater two	200	61.09	8.188	30	80
Rater three	200	62.08	9.873	40	85
Rater four	30	63.67	14.967	30	90
Rater five	200	64.63	8.152	40	90
Rater six	200	64.37	8.753	40	85
Rater seven	200	61.07	7.256	40	80
Average raters rating		62.04	7.89	37	83
Actual Difficulty	200	60.93	23.826	4	99

Descriptive analysis of the UGMLE item difficulty ratings

In this study, experts rated 200 MCQs from fourteen exam domains to determine the examinees' cut-off scores, and the study included all item data of the exam booklet. Table 1 presents the descriptive statistics summary of the raters' average ratings and actual difficulty scores.

As shown in Table 1, the mean expert's rating of the item difficulty of the UGMLE was found between 59.39 and 64.67 and the mean value was 62.04. Furthermore, the actual mean examinees' performance mean score was 60.93.

Analysis of the predictive power of raters' on the actual exam difficulty values across exam domains

Table 2 presents the analysis of the experts' rating predictive power on the actual performance across the fourteen exam domains of UGMLE. The study identified inconsistencies in the mean difference between raters' prediction and actual performance. For instance, in exam domain thirteen, the mean actual performance (Mean = 41.74, SE = 6.19) was 30.4% lower than the mean rater prediction (Mean = 59.93; SE = 1.71). Whereas in domain three the mean actual performance (Mean = 70.99, SE = 4.34) was greater than the mean rater prediction (Mean = 64.46; SE = 1.14) by 10.1%. Overall, in this study, the maximum differences were observed in exam domain thirteen (30.4%), domain seven (26.4%), and domain eleven (25.5%). On the other hand, the minimum differences were observed in exam domain four (0.1%), domain one (1.2%), and domain two (2.4%) (Table 2).

The relationship between raters' perceived difficulty and examinees' actual difficulty value was computed using linear regression analysis for each of the fourteen exam domains. The proportion of variance in actual difficulty explained by raters' prediction identified a wide difference in degrees of determination across exam domains. As depicted in Table 2, the regression model showed a statistically significant relationship in exam domain eight, F (1, 4) = 16.05, $P=0.016$, and exam domain eleven F (1, 6) = 17.86, $P=0.006$. The value of the regression analysis explained 80% ($R^2=0.8$) and 74.8% ($R^2=0.748$) of the variance in actual difficulty explained by the raters' prediction in domains eight and eleven respectively. The correlation coefficient of $R=0.895$, and $R=0.865$ in these domains indicated a strong positive

Table 2 The proportion of variance of rater prediction in determining the examinees' actual performance across the fourteen exam domains of UGMLE

Domain	No. of MCQs	Rater difficulty level		Actual difficulty level		Model			
		Mean	SE	Mean	SE	*R* Square	SE	F	Sig.
One	24	62.86	4.61	62.06	2.56	0.48	16.59	20.53	0.000
Two	24	63.3	1.77	64.83	3.75	0.25	16.32	7.16	0.014
Three	26	64.46	1.14	70.99	4.34	0.17	20.55	5.01	0.035
Four	24	64.84	1.42	64.78	5.44	0.37	21.56	13.17	0.001
Five	24	65.34	0.94	72.54	4	0.42	15.27	15.96	0.001
Six	8	60.31	1.07	55.34	9.26	0.13	26.41	0.89	0.382
Seven	6	59.17	3	43.54	9.9	0.001	27.1	0.003	0.961
Eight	6	58.19	3.96	55.42	11.46	0.8	14.02	16.05	0.016
Nine	10	58.65	1.99	62.73	7.06	0.29	27.05	3.18	0.112
Ten	8	58.6	5.6	62.7	19.9	0.39	16.82	3.87	0.097
Eleven	8	54.69	2.7	40.74	7.34	0.75	11.25	17.86	0.006
Twelve	12	59.65	1.84	51.51	5.24	0.03	18.79	0.25	0.625
Thirteen	12	59.93	1.71	41.74	6.19	0.13	20.96	1.49	0.249
Fourteen	8	59.06	1.79	63.22	7.4	0.096	21.49	0.64	0.455
Total	200	62.04	0.564	60.93	1.68	0.32	19.66	16.55	0.000

Table 3 The Mean difference between two raters (I-J) using post hoc multiple comparisons

	Mean difference (I-J)					
Rater (I)	Rater (J) 1	2	3	5	6	7
1	0.00					
2	1.75	0.00				
3	2.73**	0.98	0.00			
5	4.95**	3.20**	2.23	0.00		
6	5.00**	3.25**	2.28	0.50	0.00	
7	1.58	-0.17	-1.15	-3.37**	-3.42**	0.00

**The mean difference is significant at 0.05

correlation between the actual and predicted difficulty values. However, the correlation coefficient of $R=0.026$, and $R=0.157$ in domain seven and domain twelve respectively, indicated no correlation between the actual and predicted difficulty values. These correlations were not statistically significant in domain seven, $F(1, 4)=0.003$, $P=0.961$, and domain twelve $F(1, 10)=0.25$, $P=0.625$. The value of the regression analysis also explained 0.1% ($R^2=0.001$) and 3% ($R^2=0.03$) of the variance in the actual difficulty explained by the raters' prediction in these domains respectively.

In general, a non-statistically significant relationship with notable mean differences observed in a few domains of UGMLE indicated that the observed differences were due to random chance rather than a true predictive relationship. It also indicated that to set a defensible cut score, linear regression analyses are more important than the mean difference to provide a comprehensive understanding of how actual performance is related to raters' prediction.

Raters' mean differences in item difficulty level ratings

A one-way ANOVA was conducted to compare the mean differences across the six raters of UGMLE. The analysis showed a significant difference in mean raters' scores, $F(5, 1188)=10.61$, $p<0.001$. The post-hoc comparisons using Tukey's HSD test were performed to determine the differences between raters' mean difficulty values. The analysis of the multiple comparisons is summarized in Table 3.

The analysis of multiple comparisons revealed that there were statistically significant ($p<0.05$) mean differences between Rater 1 and Rater 5 (mean difference=5.23), between Rater 1 and Rater 6 (mean difference=4.97), between Rater 2 and Rater 5 (mean difference=3.53), between Rater 2 and Rater 6 (mean difference=3.27), between Rater 3 and Rater 6 (mean difference=2.28) and between Rater 5 and Rater 7 (mean difference = -3.56). However, there were statistically non-significant mean difference observed between the difficulty value ratings of the other raters.

Correlation analysis

The agreement between expert ratings and actual difficulty for MCQ requires a thorough comparison between interrater and actual difficulty scores. To this end, the correlations between raters' rating values and the actual difficulty were computed to know whether or not

statistically significant relationships exist. Table 4 presents the summary results of the comparison between inter-rater ratings of MCQs' difficulty score and the actual difficulty values.

As indicated in Table 4, the Pearson correlation coefficient of inter-rater ratings ranged between $r=0.725$ and $r=0.838$. These inter-rater correlation coefficient values indicated a strong positive correlation among raters' prediction in the UGMLE that were statistically significant at $p<0.001$. Similarly, a statistically significant correlation value at $p<0.001$ was found between the actual difficulty values and individual raters' ratings in 200 MCQs of the UGMLE. These correlation coefficient values were ranged between $r=0.455$ and $r=0.545$ that indicated moderate positive correlations between raters' perceived difficulty and examinees' actual difficulty.

Regression analysis

In this study, the six experts' item difficulty ratings of the UGMLE had 33% power in predicting the actual difficulty level. This means that the remaining 67% of the variance was determined by other variables beyond the raters' ratings of difficult values. The regression analysis showed that the model was statistically significant at F (6, 193)=15.58, $P=0.001$, and the proportion of variance explained in the model was $R^2=0.33$ ($R=0.57$, and adjusted $R^2=0.31$). The correlation coefficient of $R=0.57$ indicated a moderate positive correlation between the actual examinees' performance and experts predicted difficulty value.

Discussion

This study focused on item analysis of 200 MCQ recorded data from of 815 UGMLE examinees and item difficulty ratings of seven physicians from different clinical specialties. These experts who participated in the item difficulty rating for standard settings had more than five years of teaching experience in higher education institutions. The experts' selection criteria aligned with the national exam development directorate guideline for experts to participate in the standard settings, except the required number of experts involved in the process was 7–15. In addition, a similar high-stakes examination also involved experts with proven subject matter expertise and more than two years of teaching experience for item difficulty level ratings [19, 20]. The percentage of borderline examinees' correct answers was used to apply the Angoff method of standard settings to determine the cut-off score. This method is widely used for standard settings of high-stakes examinations, including licensure examinations, and has been deemed to be the method that best balances technical suitability and practicality [21, 22]. The experts were presented with a total of 200 MCQs from fourteen exam domains for item difficulty ratings to set the cut-off score for the examinees' pass mark decision. Commonly, 100–200 MCQs were used for such exams and studies also have shown a similar number of MCQs for licensure examination of physicians and other health professionals [23, 24]. This assumption is considering many items to improve the reliability and validity of the exam scores.

The mean and standard deviation of the raters' difficulty ratings ranged between 59.4±8.5 and 64.6±8.2, with a minimum score of 30 and a maximum score of 90. However, the mean and standard deviation of the actual difficulty was 60.9±23.8. This finding was consistent with the difficulty values of similar exams reported in Ethiopia, Pakistan, and India [25–27]. This might be raters have a similar understanding of the minimum performance expected from borderline students in the licensure examination. The other possible reason might be the similarities in the experts' qualifications, experience in item rating, teaching, and clinical experience [28]. However, studies carried out in Mongolia and India have reported higher mean difficulty values of experts' rating [23, 29].

Again, this study analyzed the variance of raters' difficulty ratings to predict the actual difficulty across exam domains and the full exam items. The study finding noted a statistically significant positive linear relationship at $p<0.05$ between mean raters' difficulty ratings and the actual difficulty in seven of the fourteen exam domains of UGLME. This result indicates the MCQs assembled across these domains are found to be sound in attaining the intended purpose of the exam [30]. In particular, this study demonstrated a statistically significant strong positive linear relationship in the UGLME domains eight

Table 4 The comparison between inter-rater difficulty and actual difficulty values agreement in the UGMLE

Raters	R_1	R_2	R_3	R_5	R_6	R_7	Actual Diff
Rater One	1						
Rater Two	0.833**	1					
Rater Three	0.830**	0.768**	1				
Rater Five	0.816**	0.791**	0.725**	1			
Rater Six	0.833**	0.793**	0.796**	785**	1		
Rater Seven	0.838**	0.784**	0.794**	0.779**	0.830**	1	
Actual Diff	0.536**	0.509**	0.545**	0.455**	0.499**	0.489**	1

**. Correlation is significant at the 0.01 level (2-tailed)

and eleven. This significant relationship indicates that the MCQs in these domains are valid to measure the minimal performance of physicians required to deliver healthcare services. This further explained that the UGLME items under these domains reflect the cognitive demands expected from the examinees [31].

The non-significant relationship predictions of actual difficulty in other domains showed that the raters' assessments differed, which may have highlighted the subjectivity of item difficulty prediction, the experts' specialization consistency in predicting difficulty levels in particular domains, and the rater's selection criteria based on required experiences and domain expertise [20, 32, 33].

Additionally, this study examined the relationship between the interrater difficulty prediction and the actual difficulty values. The analysis revealed significant interrater correlations and between raters' and actual examinees' performance. These notable relationships between raters were in line with the findings reported the significance of developing a shared understanding among raters to improve the accuracy of predicting the actual difficulty of assessments [34].

The significant correlations between the raters' difficulty and the actual difficulty values, indicated that the raters' difficulty ratings are consistent with the actual difficulty of the exam. This consistency is essential for the validity of licensure examination scores, to reflect the cognitive demands expected from the examinees [31]. The finding of this study explored that experts' item difficulty prediction accounted for only 33% of the actual difficulty values, but the remaining 67% of the variability was attributed to other unexplored variables. The observed limited power (33%) of raters in predicting actual difficulty values evidenced the challenge of capturing the details of exam items' difficulty through expert judgment alone. This aligned with the scholar's suggestion on the need for working on continuous improvement in licensure examination design and administration [34] and the importance of adequate training to enhance raters' ability to ensure meaningful contributions to examinees' performance prediction [34].

Investigating the potential sources of the observed relationship between the expert ratings and the examinee's performance score in the UGMLE prediction is necessary. The findings and cause of concerns raised in this study have called researchers to further explore the limitations of relying on experts' judgment to make exam decisions and the need for designing possible mitigating mechanisms [31, 35]. In addition, these unexplained variabilities might be rooted in the multifaceted nature of exam-related factors and other constructs, such as item format, examinee characteristics, and non-test-related attributes [36, 37]. The regression model further confirmed that as raters perceived item difficulty increased, there was a corresponding increase in the actual difficulty levels. This finding was consistent with the notion that expert judgments hold value in predicting item difficulty but there is a need to explore additional variables contributing to the overall construct [38]. The statistically significant differences among the experts' ratings were consistent with other similar study findings of rating variations among experts [10, 16]. This implies that some raters' judgments are aligned with the examinees' actual score, whereas others' predictions deviate from reality [39].

Berk (1996) noted several factors that can cause variations in the experts' mean item difficulty ratings differences, including the types of items being rated, differences in experts' backgrounds, and a lack of clarity in performance standards perceptions [5].For instance, rater one and five, rater one and six, rater two and five, rater two and six, rater three and six, and rater five and seven demonstrated significant item difficulty mean rating differences. Studies have also reported that exam item rater experts' overestimate or underestimate the examinees' level of achievement [40–43].

Conclusions

The purpose of this study was to evaluate how well experts' predictions match the actual of difficulty of MCQs on the Ethiopian UGMLE. The results revealed statistically significant mean differences in item difficulty ratings of certain raters, while the other raters did not differ significantly. Additionally, the results of the correlation analysis showed a statistically significant moderate correlation between the raters' average ratings and the actual difficulty scores and a strong correlation between raters. In conclusion, this study highlighted the complex nature of rating MCQs' difficulty level in medical licensure examinations and revealed the potential of using expert ratings if considerations are given to improve the capabilities of expert.

Implications for theory and practice

In the realm of education and assessment, it is beneficial to compare the prediction of experts with the actual difficulty of MCQs on undergraduate medical licensing exams. This kind of analysis can shed light on how well assessments match learning objectives, how well questions are designed, and how well the exam is done overall.

The results of this study point to several potential explanations for the difference that exists between the perceived and actual levels of difficulty in the MCQ of the Undergraduate Medicine Licensure Examination. Realizing that evaluating MCQ's difficulty is a complex process and is essential. The examination procedure needs to be updated often to take these factors into account.

This includes stringent review procedures, continuous statistical analysis, frequent adjustments to account for shifts in medical education and practice, and the hiring and training of item writers. Expert evaluations can be aligned with examinees' actual levels of difficulty using of consistent feedback mechanisms and a commitment to continuous improvement.

MCQs may appear more difficult than planned from the perspective of the examines' if curriculum changes have taken place and are not adequately reflected in the test. Additionally, the assumptions and biases of experts may affect how difficult an MCQ is perceived. It is possible that this subjectivity does not always match the actual difficulty that examinees' face. If there is no feedback loop in place where MCQ results are reviewed and improvements are made in response to actual examinees' performance, there may also be persistent discrepancies between the experts' perceived and actual levels of difficulty.

The techniques and standards for peer review of MCQs can improve the psychometric qualities of items. Programs for continuous professional development (CPD) that seek to increase the quality of MCQs should take note of these findings [44]. Asking teachers or subject matter experts to rate MCQs based on their areas of expertise is a quick and easy way to determine the level of difficulty in advance [45].

This study highlights the need for ongoing efforts to enhance rater training to promote consistency prediction among raters across exam domains and ensure the validity of the licensure examination standard-setting process. In this regard, licensure examination directorates can improve the ability of experts to rate the difficulty levels of MCQs in medical licensure examinations in Ethiopia by adopting a multidimensional strategy that includes expert training, monitoring and evaluation of MCQ item development, application of quality assurance mechanisms, and item banking. This approach helps to ensure the validity and reliability of licensure examination item difficulty rating approaches and thereby the quality and fairness of MCQs in medical licensure examinations.

Acknowledgements
The authors express sincere gratitude to the Health Professional Competency Assessment and Licensing Directorate of the National Ministry of Health for generously providing the entire dataset used in this study.

Author contributions
SGW and SKS conceived the concept and have contributed sufficiently to the development and preparation of this paper. TT and BM were involved in the data collection and carried out the analyses. All authors reviewed and approved the final manuscript.

Funding
The funding for this research was provided by the University of Oslo (UiO) under the NORPART and EXCEL SMART project, supporting the corresponding author's PhD research.

Data availability
The licensure examination examines performance and experts judgment data have been deposited in Ethiopian Licensure examination database. The Author will present the anlysed data based on the request from the publisher. The data not available openly on any link due to the privacy of the participant and the high stake nature of the licensure examination data. Point of contact is gedamuwonde@gmail.com.

Declarations

Ethical considerations
The study protocol for this cross-sectional study was approved by the Jimma University Health Institute Research Review Board (Reference No. IHRPG/868/20) and the Norwegian Data Protection Authority (Reference number 321099) to use the data for research. All study participants' data and examinee identifiers were mentioned anonymously for confidentiality.

Competing interests
The authors declare no competing interests.

Author details
[1]Institute of Health, Jimma University, Jimma, Ethiopia
[2]Faculty of Medicine, Institute of Health and Society, University of Oslo, Oslo, Norway
[3]Institute of Educational Research (IER), Addis Ababa University, Addis Ababa, Ethiopia
[4]Educational Development and Quality Center, University of Global Health Equity, Kigali, Rwanda
[5]Institute of Education and Behavioural Sciences, Department of Psychology, Dilla University, Dila, Ethiopia
[6]Faculty of Educational Sciences, Faculty of Medicine, University of Oslo, Oslo, Norway

Received: 16 February 2024 / Accepted: 10 September 2024
Published online: 16 September 2024

References

1. Swanson DB, Roberts TE. Trends in national licensing examinations in medicine. Med Educ. 2016;50(1):101–14.
2. Breithaupt K. Medical Licensure Testing, White Paper for the Assessment Review Task Force of the Medical Council of Canada. 2011.
3. Archer DJ, Lynn DN, Roberts MM, Lee D, Gale DT. A Systematic Review on the impact of licensing examinations for doctors in countries comparable to the UK. 2019.
4. Castle RA. Developing a Certification or Licensure Exam. 2002.
5. Berk RA. Standard setting: the Next Generation (where few psychometricians have gone before!). Appl Measur Educ. 1996;9(3):215–25.
6. Kane M. Validating the performance standards Associated with passing scores. Rev Educ Res. 1994;64(3):425–61.
7. Clauser BE, Swanson DB, Harik P. Multivariate Generalizability Analysis of the impact of training and Examinee Performance Information on judgments made in an Angoff-Style Standard-setting Procedure. J Educ Meas. 2002;39(4):269–90.
8. Bramley T, Wilson F. Maintaining test standards by expert judgement of item difficulty. 2016.
9. Südkamp A, Kaiser J, Möller J. Teachers' judgments of students' academic achievement: results from field and experimental studies. In: Krolak-Schwerdt S, Glock S, Böhmer M, editors. Teachers' Professional Development. Rotterdam: Sense; 2014. pp. 5–25.
10. Südkamp A, Kaiser J, Möller J. Accuracy of teachers' judgments of students' academic achievement: a meta-analysis. J Educ Psychol. 2012;104(3):743–62.
11. Ready DD, Wright DL. Accuracy and inaccuracy in teachers' perceptions of Young Children's cognitive abilities: the role of child background and Classroom Context. Am Educ Res J. 2011;48(2):335–60.
12. Mortaz Hejri S, Jalili M. Standard setting in medical education: fundamental concepts and emerging challenges. Med J Islam Repub Iran. 2014;28:34.
13. Norcini JJ. Setting standards on educational tests. Med Educ. 2003;37(5):464–9.

14. Barman A. Standard setting in student assessment: is a defensible method yet to come? Ann Acad Med Singap. 2008;37(11):957–63.
15. Meissel K, Meyer F, Yao ES, Rubie-Davies CM. Subjectivity of teacher judgments: exploring student characteristics that influence teacher judgments of student ability. Teach Teacher Educ. 2017;65:48–60.
16. Machts N, Kaiser J, Schmidt FTC, Möller J. Accuracy of teachers' judgments of students' cognitive abilities: a meta-analysis. Educational Res Rev. 2016;19:85–103.
17. Lin LC, Tseng HM, Wu SC. Item analysis of the registered nurse license exam by nurse candidates from vocational nursing high schools in Taiwan. Proc-Natl Sci Counc ROC(D). 1999;9(1):24–30.
18. Attali Y, Saldivia L, Jackson C, Schuppan F, Wanamaker W. Estimating item Difficulty with comparative judgments: estimating Item Difficulty. ETS Res Rep Ser. 2014;2014(2):1–8.
19. MOH. Federal Democratic Republic of Ethiopia, Health professinals Licensure examination Development Manual. 2019.
20. Brunk I, Schauber S, Georg W. Do they know too little? An inter-institutional study on the anatomical knowledge of upper-year medical students based on multiple choice questions of a progress test. Ann Anat. 2017;209:93–100.
21. Bandaranayake RC. Setting and maintaining standards in multiple choice examinations: AMEE Guide 37. Med Teach. 2008;30(9–10):836–45.
22. Berk RA. A consumer's guide to setting performance standards on Criterion-Referenced tests. Rev Educ Res. 1986;56(1):137–72.
23. Gomboo A, Gomboo B, Munkhgerel T, Nyamjav S, Badamdorj O. Item analysis of multiple-choice questions in medical licensing examination. Cent Asian J Med Sci. 2019;5(2):141–8.
24. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. BMC Med Educ. 2009;9:40.
25. Hingorjo MR, Jaleel F. Analysis of one-best MCQs: the Difficulty Index, discrimination index and distractor efficiency. J Pak Med Assoc. 2012;62(2):142–7.
26. Date AP, Borkar AS, Badwaik RT, Siddiqui RA, Shende TR, Dashputra AV. Item analysis as tool to validate multiple choice question bank in pharmacology. Int J Basic Clin Pharmacol. 2019;8(9):1999–2003.
27. Belay LM, Sendekie TY, Eyowas FA. Quality of multiple-choice questions in medical internship qualification examination determined by item response theory at Debre Tabor University, Ethiopia. BMC Med Educ. 2022;22(1):635.
28. Yim MK, Shin S. Using the Angoff method to set a standard on mock exams for the Korean nursing licensing examination. J Educ Eval Health Prof. 2020;17:14.
29. Mehta G, Mokhasi V. Item analysis of multiple-choice Questions- An Assessment of the Assessment Tool. Int J Health Sci. 2014;4(7).
30. Kuncel NR, Hezlett SA, Ones DS. Academic performance, career potential, creativity, and job performance: can one construct predict them all? J Pers Soc Psychol. 2004;86(1):148–61.
31. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ. 2003;37(9):830.
32. Ben-David MF. AMEE Guide 18: standard setting in student assessment. Med Teach. 2000;22(2):120–30.
33. Cusimano MD. Standard setting in medical education. Acad Med. 1996;71(10 Suppl):S112–120.
34. Clauser BE, Mazor KM. Using statistical procedures to identify differentially functioning test items. Educational Meas. 1998;17(1):31–44.
35. Lamé G, Dixon-Woods M. Using clinical simulation to study how to improve quality and safety in healthcare. BMJ STEL. 2020;6(2):87–94.
36. Cook DA. Twelve tips for evaluating educational programs. Med Teach. 2010;32(4):296–301.
37. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. Med Teach. 2011;33(6):478–85.
38. Van Der Vleuten CPM, Schuwirth LWT, Driessen EW, Govaerts MJB, Heeneman S. Twelve Tips for programmatic assessment. Med Teach. 2015;37(7):641–6.
39. Karst K, Bonefeld M. Judgment accuracy of preservice teachers regarding student performance: the influence of attention allocation. Teach Teacher Educ. 2020;94:103099.
40. Leucht M, Tiffin-Richards S, Vock M, Pant HA, Koeller O. English teachers' diagnostic skills in judging their students' competencies on the basis of the common European Framework of Reference. Zeitschrift für Entwicklungspsychologie Und Pädagogische Psychologie. 2012;44(4):163–77.
41. Feinberg AB, Shapiro ES. Teacher accuracy: an examination of teacher-based judgments of students' reading with differing achievement levels. J Educational Res. 2009;102(6):453–62.
42. Martin SD, Shapiro ES. Examining the accuracy of teachers' judgments of DIBELS performance. Psychol Sch. 2011;48(4):343–56.
43. Zhu M, Urhahne D. Teachers' judgements of students' foreign-language achievement. Eur J Psychol Educ. 2015;30(1):21–39.
44. Abozaid H, Park YS, Tekian A. Peer review improves psychometric characteristics of multiple choice questions. Med Teach. 2017;39(sup1):S50–4.
45. Huang Z, Liu Q, Chen E, Zhao H, Gao M, Wei S et al. Question Difficulty Prediction for reading problems in standard tests. AAAI. 2017;31(1).

Publisher's note