Transition From Normative to Criterion-Based Grading in the Obstetrics and Gynecology Clerkship

Cynthia Abraham[0]

Department of Obstetrics, Gynecology and Reproductive Science, Department of Medical Education at the Icahn School of Medicine at Mount Sinai, Mount Sinai Health System, New York, NY, USA.

Journal of Medical Education and Curricular Development
Volume 11: 1–5

DOI: 10.1177/23821205241239201

ABSTRACT

OBJECTIVES: To compare grades, National Board of Medical Examiners (NBME) Shelf Exam scores, and student satisfaction with the Obstetrics and Gynecology (OB/GYN) clerkship after transitioning from normative to criterion-based grading.

METHODS: Between July 2021 and July 2022, the Icahn School of Medicine at Mount Sinai (ISMMS) adhered to a normative grading scheme in which ~60% of students achieved a grade of Honors, 30% achieved a grade of High Pass and 10% achieved a grade of Pass for the OB/GYN clerkship. In July 2022, ISMMS transitioned to a criterion-based scheme. In this scheme, 6 competencies were created. Criteria were determined for each competency, delineating achieving a score of "Pass" versus "Honors" for the specific objective. Students needed to meet the criteria for Honors for 4 out of 6 of the competencies in order to ultimately receive a grade of Honors for the clerkship. The number of students achieving Honors, NBME shelf exam scores, and student clerkship satisfaction ratings between the normative and criterion-based schemes were compared.

RESULTS: The number of students studying in academic year (AY) 2021–2022 and AY 2022–2023 were 134 and 137, respectively. A significantly lower percentage of students received Honors in AY 2021–2022 than in AY 2022–2023 (66% vs. 96%, $P < .01$). Mean exam scores were significantly higher for those receiving Honors in AY 2021–2022 than in AY 2022–2023 ($P < .05$); scores for AY 2021–2022 and AY 2022–2023 were 78.9, 95% CI [77.6, 80.1] and 76.7, 95% CI [75.6, 77.8], respectively. Mean exam scores for all students were not significantly different between the 2 academic cohorts (77.8 vs. 76.2, $P = .06$). Clerkship satisfaction rating was significantly higher in AY 2022–2023 than in AY 2021–2022 (4.1 vs. 3.7, $P < .05$).

CONCLUSIONS: These findings support a paradigm that compares learner performance to predefined measures as opposed to peer performance.

KEYWORDS: undergraduate medical education, competency, grading

RECEIVED: October 25, 2023. **ACCEPTED:** February 27, 2024.

TYPE: Original Research Article

FUNDING: The author received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Cynthia Abraham, Department of Obstetrics, Gynecology and Reproductive Science, Department of Medical Education at the Icahn School of Medicine at Mount Sinai, Mount Sinai Health System, 1176 Fifth Avenue, Ninth Floor, New York, NY 10029, USA.
Email: cynthia.abraham@mssm.edu

Introduction

Traditionally, undergraduate medical education grades have been determined using a normative-based scheme. However, this model contributes to the perception of a learning environment as being one that is highly stressful. Accreditation bodies have acknowledged the influence that the learning environment has on learner performance.[1] A normative-based grading scheme leads to comparison between learners causing them to be driven by the pursuit of creating favorable impressions by supervisors and ultimately high grades.[2] Prior to the elimination of numerical scoring for the United States Medical Licensing Examination (USMLE) Step 1, USMLE Step 1 scores were used to screen applicants in the residency selection process. The use of USMLE Step 1 scores eventually proved to be problematic as scores that differed by up to 20 points were, in actuality, not significantly different given that the standard error of measurement and standard error of difference were 6 and 9 points, respectively.[3] This also hindered the ability to evaluate residency candidates holistically. In addition, the emphasis on USMLE Step 1 scores appeared to be aligned with the aim of sorting learners into different residency programs as opposed to the goal of undergraduate medical education which is to train future doctors to meet the needs of society.[4] This shift corresponds with the changes that have taken place in the graduate medical education realm over the past 2 decades.

In 2013, the Accreditation Council for Graduate Medical Education launched the Next Accreditation System, which transitioned the focus of resident performance reporting practices toward assessing specific behavioral criteria linked to an underlying competency continuum. Consequently, given this transition in graduate medical education, initiatives were constructed for undergraduate medical education to move toward a criterion-based evaluation system.[5] However, many medical schools continue to utilize normative-based grading and rely on proxy assessments, rather than direct observations, and summative, rather than formative assessments.[6] Nonetheless, as health care is becoming increasingly interprofessional, it is critical that our learners become adept in communicating with patients and members of the health care team.

In July 2022, the Icahn School of Medicine at Mount Sinai (ISMMS) transitioned from a normative-based grading scheme

to a criterion-based one. Hence, the purpose of this study was to compare grades, NBME scores, and student satisfaction with the clerkship in Obstetrics and Gynecology (OB/GYN) after this shift. There are no studies that have evaluated the use of a competency-based grading scheme solely in the OB/GYN clerkship.

Methods

Between July 2020 and July 2022, the ISMMS adhered to a normative-based grading scheme. Table 1 outlines the items used for determining the final grade for the clerkship in this model: National Board of Medical Examiners (NBME) Shelf Exam score, clinical evaluations, oral and written case presentations, direct observations scores, and completion of clinical skills assessment card. The total score was based on the number of points accrued with each item. The top 60% of scorers received a grade of Honors. Precisely, 30% of scorers below this first group received a grade of High Pass and the bottom 10% received a grade of Pass unless there were criteria that the student met warranting receipt of a grade of Fail.

In July 2022, ISMMS transitioned to a criterion-based scheme (based on competencies) which is outlined in Table 2. In this scheme, 6 objectives were created. Criteria to delineate achieving a score of "Pass" versus "Honors" for each specific competency were created. Students needed to meet the criteria for Honors for 4 out of 6 of the competnecies in order to receive a grade of Honors for the clerkship. The criterion-based scheme retained the same items for grade determination as the normative-based grading scheme. All evaluators were trained by the OB/GYN clerkship directors on the components of this grading paradigm and all associated assessments.

In this retrospective study, the following were compared between the 2 grading schemes: shelf exam scores, oral and written case presentation scores, direct observation pass rates on first attempt, and student clerkship satisfaction rating (measured on a 1–5 Likert scale with 1 being not satisfied and 5 being very satisfied). The clerkship satisfaction rating was the only question asked of students. All students included in this review were third-year medical students who were enrolled in the OB/GYN clerkship in academic years (AYs) 2021–2022 and 2022–2023. Students not enrolled during these AYs were excluded. This study was exempt from Institutional Review Board review as this was an educational research study and no sensitive student information was being obtained. Neither verbal nor written consent for the review of grades was mandated by ISMMS.

Statistical Analysis

To compare the 2 academic cohorts (AY 2021–2022 and AY 2022–2023), a student *t*-test was used. A p value of <.05 designated a statistically significant difference.

Table 1. Grading components.

Grading component	Description
National Board of Medical Examiners (NBME) Shelf Exam	Assessed student knowledge base. Student received 15 points if achieved NBME shelf exam score of at least 5th percentile.
Clinical Evaluation Form	Components: knowledge, history-taking skills, physical examination skills, ability to form a differential diagnosis, ability to construct plans for follow-up, patient interaction, team interaction, dependability, and engagement. Five grading categories: points of 1, 2, 3, 4, or 5 allotted for does not meet expectations, below expectations, meets expectations, exceeds expectations, and greatly exceeds expectations, respectively **In a normative-based grading scheme, the number of clinical evaluation points was calculated after summation of assigned component points.*
Written Presentation	Components: presentation, assessment, quality of topic review and organization, maximum of 2, 2, 4, and 2 points, respectively. Maximum number of points: 10.
Oral Presentation	Components: presentation, assessment, quality of topic review and organization, maximum of 2, 2, 4, and 2 points, respectively. Maximum number of points: 10.
Direct Observation—History	Assessed learner's ability to evaluate abnormal uterine bleeding. Standardized patient encounter. Maximum number of points: 10.
Direct Observation—Physical Exam	Assessed learner performance of Papanicolaou smear on a pelvic exam trainer. Standardized observer used. Maximum number of points: 10.
Clinical Skills Assessment Card	For satisfactory completion, students attested to completing online quizzes, watching educational videos, and obtained signatures after being observed obtaining a history and performing a female genitourinary exam. Student received 10 points on confirmation of successful completion of card.

Results

One hundred and thirty-four students were enrolled in the OB/GYN clerkship in the academic year (AY) 2021–2022; and 137 in AY 2022–2023. Mean shelf exam scores for all students were not significantly different between the 2 academic cohorts (77.8 vs. 76.2, $P = .06$). Mean shelf exam scores were significantly higher for those receiving Honors in AY 2021–2022 than in AY 2022–2023 ($P < .05$); scores for AY 2021–2022 and AY 2022–2023 were 78.9, 95% CI [77.6, 80.1] and 76.7, 95% CI [75.6,

Table 2. Criterion-based grading scheme.

Competency	Assessment	Criteria—Pass	Criteria—Honors
Identify risk factors and prevention strategies for common medical conditions occurring throughout the lifespan.		Complete the required activities on the CSA.	
Describe and develop a differential diagnosis and plan for common obstetric conditions (including routine prenatal, intrapartum, and postpartum problems), applying principles of maternal physiology and anatomy.	CE: Knowledge CE: Differential diagnosis CE: Assessment CE: Plans CE: Follow-up NBME exam	On average, meets expectations **AND** NBME SHELF EXAM ≥ 5th percentile first or second attempt **AND** Oral presentation ≥ 6 out of 10 **AND** Written presentation ≥ 6 out of 10	On average, exceeds expectations **OR** NBME SHELF EXAM ≥15th percentile on the first attempt **AND** Oral presentation ≥ 8 out of 10 **OR** Written presentation ≥ 8 out of 10
Describe and develop a differential diagnosis and plan for common gynecologic conditions.	CE: Knowledge CE: Differential diagnosis CE: Assessment CE: Plans CE: Follow-up NBME exam	On average, meets expectations **AND** NBME SHELF EXAM ≥fifth percentile first or second attempt **AND** Oral presentation ≥ 6 out of 10 **AND** Written presentation ≥ 6 out of 10	On average, exceeds expectations **OR** NBME SHELF EXAM ≥15th percentile on the first attempt **AND** Oral presentation ≥ 8 out of 10 **OR** Written presentation ≥ 8 out of 10
Complete a comprehensive women's medical interview	CE: Knowledge DO: History	CE: On average, meets expectations **AND** passing DO score on the first or second attempt	CE: On average, exceeds expectations **OR** passing DO score on the first
Develop rapport with patients, taking into account patients' social and cultural contexts.	CE: Communication—patient	CE: On average, meets expectations	CE: On average, exceeds expectations
Work cooperatively with the healthcare team	CE: Communication—team CE: Dependability, engagement CE: Responsiveness to feedback		
Perform basic skills and procedures relevant to the practice of obstetrics and gynecology (including the complete female genitourinary exam) in an accurate and sensitive manner.	CE: Knowledge DO: Physical exam	CE: On average, meets expectations **AND** passing DO score on the first or second attempt	CE: On average, exceeds expectations **OR** passing DO score on the first

Abbreviations: CE, Clinical Evaluation; DO, Direct Observation; NBME, National Board of Medical Examiners; CSA, Clinical Skills Assessment Card.

77.8], respectively. Confidence intervals for these means overlapped. Written and oral presentation scores were significantly higher in AY 2021–2022 than in AY 2022–2023; 8.1 versus 7.4 ($P < .05$) and 9.5 versus 9.1 ($P < .05$), respectively. A significantly lower percentage of students received Honors in AY 2021–2022 ($n = 89$, 66.4%) than in AY 2022–2023 ($n = 132$, 96.4%; $P < .01$. Clerkship satisfaction rating was significantly higher in AY 2022–2023 than in AY 2021–2022 (4.1 vs. 3.7, $P < .05$).

Table 3 outlines grading component scores for all students and grade distribution. Tables 4 and 5 outline grading component scores for those who achieved a grade of Honors and those who achieved a grade of Pass, respectively.

Discussion

After the implementation of criterion-based grading, there was a statistically higher percentage of students who achieved a clerkship grade of Honors. Overall, NBME shelf exam scores and first-time pass rates for the Direct Observations (two objectively scored grading components) did not differ significantly between the two cohorts. Shelf exam scores were significantly higher for those who achieved Honors for the clerkship preimplementation than for those who achieved Honors postimplementation. The confidence intervals for mean shelf exam scores for those who achieved Honors in the two cohorts overlapped. There was no significant difference in overall mean shelf exam scores between the cohorts which the possibility of had been entertained given multiple pathways for students to achieve a grade of Honors. Written and oral presentation scores (two subjectively scored grading components) were significantly higher in the normative-based grading cohort than in the criterion-based grading cohort. Student clerkship satisfaction ratings were significantly higher after the introduction of criterion-based

Table 3. All students.

	Academic year 2021–2022 ($n = 134$)	Academic year 2022–2023 ($n = 137$)	*P* value
National Board of Medical Examiners Shelf Exam Score	78.9 (76.7, 80.1)	76.2 (75.1, 77.4)	.06
Written Presentation Score	8.3 (8.1, 8.5)	7.4 (7.1, 7.8)	<.01
Oral Presentation Score	9.6 (9.4, 9.7)	9.0 (8.9, 9.3)	<.01
Direct Observation—History	130 (97.1%)	137 (100%)	—
Number of students passing on first attempt			
Direct Observation—Physical Exam	133 (99.2%)	135 (98.5%)	.32
Number of students passing on first attempt			
Number of students with a grade of:			
Honors	89 (66.4%)	132 (96.4%)	<.01
High pass	38 (28.4%)	N/A	–
Pass	7 (5.2%)	5 (3.6%)	.22

Scores are presented as means with associated confidence intervals. Number of students is also presented with the associated percentage.

Table 4. Students receiving a grade of Honors.

	Academic year 2021–2022 ($n = 89$)	Academic year 2022–2023 ($n = 132$)	*P* value
National Board of Medical Examiners Shelf Exam Score	78.9 (77.6, 80.1)	76.7 (75.6, 77.8)	<.05
Written Presentation Score	8.3 (8.1, 8.5)	7.4 (7.1, 7.7)	<.05
Oral Presentation Score	9.6 (9.4, 9.7)	9.1 (8.9, 9.3)	<.05
Direct Observation—History			
Number of students passing on first attempt	88 (98.9%)	132 (100%)	—
Direct Observation—Physical Exam			
Number of students passing on first attempt	89 (100%)	130 (98.5%)	—

Scores are presented as means with associated confidence intervals. Number of students is also presented with the associated percentage.

grading than before. These findings are reassuring as undergraduate medical education appears to be transitioning to a model based on the procurement of objective measurements to then determine the attainment of competencies.

Table 5. Students receiving a grade of Pass.

	Academic year 2021–2022 ($n = 7$)	Academic year 2022–2023 ($n = 5$)	*P* value
National Board of Medical Examiners Shelf Exam Score	73.5 (66.8, 80.2)	64.8 (60.3, 69.2)	.18
Written Presentation	7.6 (6.3, 8.9)	7.6	<.05
Oral Presentation	9.0 (8.3, 9.7)	9.0 (8.3, 9.7)	.75
Direct Observation—History			
Number of students passing on first attempt	6 (85.7%)	5 (100%)	—
Direct Observation—Physical Exam			
Number of students passing on first attempt	7 (100%)	5 (100%)	—

Scores are presented as means with associated confidence intervals. Number of students is also presented with the associated percentage.

The United States (US) is currently in the nascent stages of developing a criterion-based framework for undergraduate medical education. However, criterion-based assessment systems have been in place in Europe for over a decade. In the European Union, all undergraduate and graduate schools are required to base their curricula on a clear and well-defined set of competencies. Although medical educators in Europe were slow to accept this system, they have grown to appreciate this system's dedication to quality improvement and to creating transparency.[7]

The number of studies on the implementation of criterion-based grading in the US is scant. Although there has been a multitude of commentaries published in the US addressing the benefits and pitfalls of criterion-based grading, there has been only one published study to date that assessed outcomes after transitioning to this grading scheme. In a study by Ryan et al,[8] Virginia Commonwealth University enacted a 4-tiered (Honors, High Pass, Pass, and Fail) criterion-based grading model in which students were evaluated across 4 domains: "Patient Care," "Professionalism," Communication/teamwork," and "Medical Knowledge." Students achieved Honors if they were deemed competent in all domains, exemplary in the "Patient Care" domain, and exemplary in at least 1 other domain. With this scheme, compared to previous AYs, more students (40% vs. 15%) received a grade of Honors, while substantially fewer (20% vs. 50%) received a grade of Pass. The authors found that this grading model was successful in providing various pathways for students to achieve Honors and allowing for recognition of students who excelled in patient care, professionalism, and communication/teamwork but not necessarily medical knowledge, similar to this study. In this study, 2 students achieved a grade of Honors despite having to

remediate one of the Direct Observations and 1 achieved Honors despite not achieving a shelf exam score of greater than or equal to the fifth percentile on the first attempt.

This study is limited by its retrospective nature and its restriction to a single medical school. Extending the study period (as this study included 2 years of data) would substantiate the conclusions of this study. Another weakness is the absence of a specific question inquiring about satisfaction specifically with the grading scheme. Student clerkship satisfaction rating was based on many components such as quality of didactics and clinical exposure. An additional weakness was the unclear delineation between those students who performed significantly better than others (ie higher NBME shelf exam score or higher score on Direct Observations) given the high rate of students receiving a grade of Honors. Nevertheless, this study paves the way for a multitude of future initiatives and studies to further assess the utility of criterion-based grading as this study brings to light the benefits and downsides of this grading paradigm.

Although there has been support for criterion-based learning, educators have identified several disadvantages and challenges. One is the diminished capability of a competency-based scheme to set one learner apart from another. This can impact a learner's ability to match into their preferred residency program as well as affect an institution's reputation.[9] Precisely, 96% of the students in this study achieved a grade of Honors. A future study would entail comparing institutional OB/GYN residency match rates before and after the implementation of a criterion-based scheme. A study assessing this issue would be very informative given that the competitiveness of OB/GYN residency programs has increased significantly from 2003–2012 to 2013–2022.[10]

Another challenge is the dearth of assessments to assist in determining learner competency level. Most medical schools (including ISMMS) rely on proxy assessments such as written and oral presentations which are prone to bias. In this study, written and oral case presentation scores were significantly higher in the normative-based grading scheme than in the criterion-based one. This is an impetus to create additional objective assessments and assess learner satisfaction with the clerkship after incorporation.

A third challenge involves misconceptions about the purpose of competency-based assessment. Criterion-based assessment is a formative tool, not a summative one. However, in a study that evaluated the operationalization of an assessment to measure the performance of core Entrustable Professional Activities (EPAs) in the pediatrics clerkship, it was apparent that students self-selected completion of EPAs that they thought would favorably affect their summative evaluation.[11] In this study, on receipt of informal feedback, there were learners who requested more sessions on learning how to obtain histories and perform physical examinations in order to master these skills. There were several learners though who requested more sessions in order to perform successfully on the Direct Observations. One initiative would be to not only evaluate the attainment of a particular competency over multiple time points but also assess learner perception of the assessment.

A fourth challenge is in creating a harmonious undergraduate medical education (UME) competency taking into account the following: (a) composition of committee members (combination of faculty and students who are knowledgeable with the curriculum vs. faculty members only who are not as familiar as students but who could potentially be more objective), (b) assessments to be used, and (c) learners to be reviewed (all or only problem ones).[12] Given the high percentage of students who received Honors, the OB/GYN clerkship grading committee at ISMMS elected to review learners who did not meet the criteria for a grade of Honors. A future study would require obtaining input from institutions that have transitioned over to criterion-based grading. A compilation of best practices for OB/GYN clerkship grading committees can ultimately be created.

Conclusion

In conclusion, the findings from this study support a model that compares learner performance to predefined measures as opposed to peer performance. These findings are especially important as our future physicians will be evaluated on a criterion-based scheme during residency. Hence, it is imperative that those involved in UME prepare them. Additional studies will have to be conducted to further elucidate the efficacy of criterion-based grading.

REFERENCES

1. Bierer SB, Dannefer EF. The learning environment counts: longitudinal qualitative analysis of study strategies adopted by first-year medical students in a competency-based educational program. *Acad Med*. 2016;91(11 Association of American Medical Colleges Learn Serve Lead: Proceedings of the 55th Annual Research in Medical Education Sessions):S44-S52.
2. Smith JF Jr, Piemonte NM. The problematic persistence of tiered grading in medical school. *Teach Learn Med*. 2023;35(4):467-476.
3. United States Medical Licensing Examination. USMLE score interpretation guidelines. http://www.usmle.org/pdfs/transcripts/USMLE_Step_Examination_Score_Interpretation_Guidelines.pdf
4. Dhaliwal G, Hauer KE. Excellence in medical training: developing talent-not sorting it. *Perspect Med Educ*. 2021;10(6):356-361.
5. Scielzo SA, Abdelfattah K, Ryder HF. Is it all about the form? Norm- vs criterion-referenced ratings and faculty inter-rater reliability. *Ochsner J*. 2023;23(3):206-221.
6. Pereira AG, Woods M, Olson APJ, van den Hoogenhof S, Duffy BL, Englander R. Criterion-based assessment in a norm-based world: how can we move past grades? *Acad Med*. 2018;93(4):560-564.
7. Patrício M, Harden RM. *Medical education and the Bologna process in EUA Bologna. Handbook: Making bologna work*. Berlin: Raabe Academic Publishers; Brussels: European University Association; 2009. http://www.bologna-handbook.com/
8. Ryan MS, Haley KE, Helou M, Ryan MH, Rigby F, Santen SA. Bringing clerkship grading back to the bedside. *Clin Teach*. 2021;18(3):274-279.
9. McDonald JA, Lai CJ, Lin MY, O'Sullivan PS, Hauer KE. "There is a lot of change afoot": a qualitative study of faculty adaptation to elimination of tiered grades with increased emphasis on feedback in core clerkships. *Acad Med*. 2021;96(2):263-270.
10. Michelotti AM, Stansbury N, Treffalls RN, Page-Ramsey SM. Competitiveness of obstetrics and gynecology residency programs and applicants. *Obstet Gynecol*. 2023;142(2):364-370.
11. Rodgers V, Tripathi J, Lockeman K, Helou M, Lee C, Ryan MS. Implementation of a workplace-based assessment system to measure performance of the core entrustable professional activities in the pediatric clerkship. *Acad Pediatr*. 2021;21(3):564-568.
12. Monrad SU, Mangrulkar RS, Woolliscroft JO, et al. Competency committees in undergraduate medical education: approaching tensions using a polarity management framework. *Acad Med*. 2019;94(12):1865-1872.