# Examinations (OSCEs) for Medical School Graduation: A Comparison of the borderline and Hofstee methods

*Marcelo R. García Diéguez[1], Marta P. del Valle[1], Alejandro G. Cragno[1]*

## ABSTRACT

***Introduction.*** Setting the cut-off point in objective structured clinical examinations (OSCEs) is a controversial aspect of assessment. In resource-limited settings, the Hofstee method requires additional tasks from other teachers outside the examination time. In contrast, the borderline group method is applied during the assessment, allowing for a more efficient use of time and resources.

***Objective.*** To compare the reliability of the borderline group and Hofstee methods applied in a graduation OSCE at an Argentine public university, providing local evidence to an internationally relevant debate.

***Population and methods.*** Cross-sectional study of 56 students in a 12-station OSCE. Two standard-setting methods were applied: borderline group (using observers during the exam) and Hofstee (electronic consultation with expert judges). Cut-off points, failure rates, and reliability (phi coefficient λ) were compared using generalizability theory.

***Results.*** The average score was 66.1 (SD 4.7). The cut-off point using the borderline group method was 54 (overall) with a reliability of 0.89 and no failures. The Hofstee method defined cut-off points of 60.7 (overall), with 3 and 1 students failing, respectively, and reliability of 0.68 and 0.82.

***Conclusions.*** Both methods show adequate reliability; however, they differ in their practical consequences, as the borderline group method was more lenient, generating a higher number of passing students.

***Keywords:*** *professional competence; clinical competence; educational assessment; medical students; undergraduate medical education.*

doi: http://dx.doi.org/10.5546/aap.2025-10758.eng

**To cite:** García Diéguez MR, del Valle MP, Cragno AG. Examinations (OSCEs) for Medical School Graduation: A Comparison of the borderline and Hofstee methods. *Arch Argent Pediatr.* 2025;e202510758. Online ahead of print 28-AUG-2025.

*[1] Center for Studies in Health Professional Education (CEEProS), Department of Health Sciences, Universidad Nacional del Sur, Bahía Blanca, Argentina.*

***Correspondence to*** *Marcelo R. García Diéguez: mgdieguez@uns.edu.ar*

***Funding**: None.*

***Conflict of interest**: None.*

***Received**: 5-26-2025*
***Accepted**: 7-3-2025*

## INTRODUCTION

The assessment of clinical competencies in medicine has evolved significantly in recent decades, consolidating the objective structured clinical examination (OSCE) as a reliable and valid strategy for evaluating student performance.[1] Through standardized clinical stations that reproduce everyday situations in medical practice, the OSCE allows for the assessment of skills such as physical examination, communication, and clinical reasoning through direct observation and structured checklists,[2] thus overcoming the limitations of more subjective traditional assessments.[3]

A central aspect of its implementation is the determination of the cut-off point, i.e., the threshold that establishes whether a student has reached the minimum acceptable level of competence. The methods for setting this threshold are classified into three types: normative (group-based), empirical or examinee-centered (such as the borderline group –BGM–), and compromise, judgment, or consensus (such as Hofstee, combining empirical and normative criteria).[4,5]

The Hofstee method requires additional tasks and planning before or after the exam.[6] In contrast, the BGM can be applied simultaneously with the OSCE, which promotes efficiency in contexts with limited human resources, such as those where part-time and multi-employed teachers predominate.[7] Despite mixed psychometric results, both methods have proven helpful for high-impact decisions such as accreditation or graduation.[8]

This study seeks to compare the reliability and practical implications of the BGM and Hofstee methods in a graduation OSCE at an Argentine public university, providing local evidence to an international debate.

## POPULATION AND METHODS

A cross-sectional study was conducted with 56 medical students in their final year. The OSCE, used as a final comprehensive exam, included 12 clinical stations distributed across four circuits and two shifts. Each station had a maximum score of 100, using dichotomous checklists. The final score was the average of all stations, applying a system of total compensation between stations and dimensions.

We used two methods to determine cut-off points: BGM and Hofstee. In the 48 teachers observed the performance and, in addition to completing the checklist, assigned an overall grade, classifying students as outstanding, satisfactory, borderline, or unsatisfactory. The average of the borderline scores determined the cut-off point for each term; their average established the overall standard.

For the Hofstee method, 17 judges completed an electronic survey with four key questions: minimum and maximum acceptable percentages of passes, and minimum and maximum scores for passing. A graph was constructed showing the intersection between the student score distribution curve and the parameters established by the judges, defining the cut-off point for each station and the overall cut-off point as the sum of these.

The generalizability theory (“G-Theory”) was used to estimate the reliability of the exam and the consistency of the cut-off points. First, a generalizability study was conducted with a student × station (SdC/St) design, equivalent to Cronbach’s alpha as an estimate of the relative consistency of student scores across stations. Then, the phi lambda coefficient [$\varphi(\lambda)$] was calculated for each method, which estimates the reliability of pass decisions, using the EduG. Method software, which estimates the proportion of systematic variance in pass/fail decisions attributable to actual student performance rather than measurement error.[9] The resulting cut-off points for each method were compared, as well as the percentage of students who failed and the values of the coefficient [$\varphi(\lambda)$].

The OSCE is mandatory for graduation. Previously, a criteria-based system was used. The Career Evaluation Committee and the Final Career Examination Committee authorized the implementation of both methods on the condition that the most lenient cut-off point for students be adopted. This decision was communicated to students, ensuring that the study would have no negative consequences.

This work was carried out following the methodological recommendations of Patricio et al. for communicating research on OSCE, ensuring a clear and transparent description.[2]

## RESULTS

The average score was 66.1 (SD = 4.7; range: 56.4-77.5). The analysis by station is presented in *Figure 1*, where the distribution of individual scores for each station can be observed.

The Hofstee method defined cut-off points of 60.7 (overall) and 57.6 (by season), with reliability [$\varphi(\lambda)$] of 0.68 and 0.82, respectively.

The BGM defined a global cut-off point of 54, with reliability 0.89 and no failures. No cut-off point was estimated by season due to the lack of sufficient borderline observations in all seasons.

The generalizability study showed adequate internal consistency. In the facet analysis (circuit and shift), no significant differences were observed in either facet.

Regarding the reliability of the instrument, the generalizability study (G Study), based on the student × station (SdC/St) measurement design, yielded a relative G coefficient—equivalent to Cronbach's alpha—indicating an adequate level of internal consistency of the assessment. For the analysis focused on the approval or disapproval decisions, the phi lambda coefficient [φ(λ)] was calculated, which allowed the reliability of criterion-referenced decisions to be estimated. This coefficient was particularly useful for comparing the consistency of the cut-off points obtained using the borderline group and Hofstee methods. The BGM had the highest decision reliability value ([φ(λ)] = 0.89), indicating high consistency in the classification of students concerning the cut-off point. In comparison, the Hofstee method with global scoring showed lower reliability ([φ(λ)] = 0.68), while the Hofstee method by station achieved an intermediate value ([φ(λ)] = 0.82).

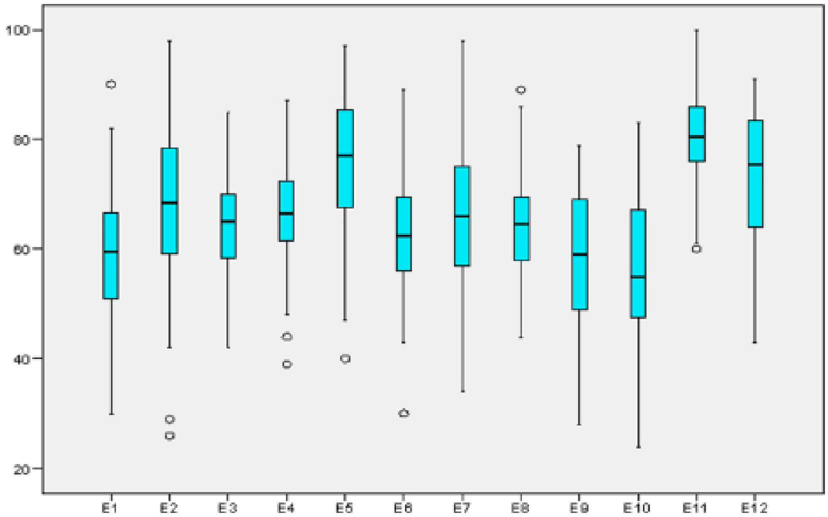
During the administration of the exam, 98 student-station combinations (student and station combinations) with borderline performance were identified, allowing specific cut-off points to be established for each station based on this subgroup. The comparative results between the two standard-setting methods, including the cut-off values and associated failure rates, are summarized in *Table 1*.

## DISCUSSION

Our results show that both the BGM and the Hofstee method yielded cut-off points with high reliability (0.89 and 0.82, respectively), which is in line with other studies that used generalizability theory to evaluate the stability of approval decisions in the OSCE.[8,10] In particular, the BGM demonstrated greater stability in the classification of students, reflected in a higher coefficient [φ(λ)], indicating more reliable decisions when determining the minimum expected clinical competence.

The comparison between methods shows

**FIGURE 1. Summary of results by station**



*E: station.*

**TABLE 1. Comparative summary of the cut-off point for boths methods**

| | Borderline group method | Hofstee overall score | Hofstee by season |
|---|---|---|---|
| Cut-off score (max. 100) | 54 | 60.7 | 57.6 |
| Number of students who failed (n = 56) | 0 | 3 | 1 |
| Coefficient φ(λ) | 0.89 | 0.68 | 0.82 |

significant differences in terms of educational impact. The BGM, in line with previous studies,[11,12] tended to generate more lenient cut-off points: in this study, no student failed using this method. In contrast, the Hofstee method resulted in one or three failures depending on whether it was applied to each station or the overall score. This finding reflects the report by Cusimano and Rothman (2003), who pointed out that compromise methods, while improving perceptions of fairness, can lead to more demanding decisions.[13] In neither case was correction for measurement error considered, which would undoubtedly increase the number of failures but would not affect the differences observed between the methods.

Some methods, such as Ansoff or Ebel, require preparatory tasks before the exam. The Hofstee method allows for asynchronous application, as was done in this study using electronic surveys, which could facilitate its implementation in distributed or virtual contexts, but requires extra effort from teachers, ideally different from the station evaluators, which increases logistical and resource needs. Unlike these other methods, the BGM has the operational advantage of being carried out simultaneously with the assessment, without the need for additional meetings. This feature makes it particularly efficient in contexts where teachers have part-time commitments and availability for extracurricular tasks is limited.[14,15]

The BGM also introduces the evaluator's overall impression without losing the advantages of the checklist, as both methods are combined, one for scoring and the other for establishing the cut-off point. Several studies in the literature have highlighted the value of these observations, since, although checklists offer a detailed and objective structure, focusing on the observation of specific actions performed by the student, global scales allow evaluators to make a holistic judgment about the student's performance, considering broader aspects of clinical competence. Some studies have indicated a significant correlation between the two methods.[16] However, the use of global scales may be influenced by the evaluator's overall impression, which could introduce subjective biases into the evaluation.[17]

Previous studies have shown that commitment methods such as Hofstee can improve students' perception of fairness and reduce individual bias among evaluators.[13] However, they may also involve stricter decisions that affect students' academic trajectories, so they should be chosen carefully.[18]

Although some authors have pointed out limitations of the BGM in small cohorts, recent studies have shown that it can be reliable even in contexts with fewer than 50 students, provided that there is adequate performance dispersion and well-designed rating scales are used.[19] In our study, the reliability achieved (G index = 0.89 for the BGM) supports this observation and reinforces the applicability of the method in contexts such as ours. On the other hand, our resource-constrained context reinforces the need to consider the relationship between the feasibility and robustness of decisions. In line with the discussion by Cole and Dupre, in this setting, methods such as the BGM, which take advantage of direct interaction between judges and students without requiring additional instances, are particularly valuable because of their low cost and adaptability to the local environment.[20]

A widely used method in universities is an absolute cut-off point based on a percentage of the maximum expected score or a percentage of the best score obtained on that exam.[21,22] This practice has limitations, as it generally produces stricter cut-off points whose reliability we cannot establish, and fundamentally does not consider the difficulty of the exam or the overall performance of the group.[15] Our findings support the use of methods based on observed performance to define more contextually appropriate standards.

A limitation of this study is the relatively small sample size, although consistent with other studies in similar contexts. In addition, the perceptions of students or teachers regarding the methods used were not explored, which could be addressed in future studies.

Furthermore, recent literature has questioned the use of solely compensatory standards in OSCEs such as those used in our research, pointing out that, in the absence of additional criteria such as a minimum number of passing grades ("conjunctive standards"), some students could pass an overall exam without having demonstrated competence in key domains.[23] In this sense, the decision not to include an additional criterion of this type in our exam may have favored a more lenient interpretation of overall competence. However, it also reflects a focus on aggregate performance, which is consistent with the foundations of the BGM.

## CONCLUSIONS

Both methods showed acceptable levels of reliability for establishing cut-off points in a graduation OSCE. However, relevant differences were observed in their practical impact: BGM offered greater consistency in student classification. It was more lenient in terms of the number of passes than the Hofstee method. ■

## Acknowledgments

The authors would like to thank María Gabriela Serralunga for her contribution to the psychometric analysis; Prof. Carlos Brailovsky (†) for his support in implementing valid assessment practices at our university; and the teaching staff, students, and the Centro de Estudios en Educación de Profesionales de la Salud for their support and logistical and academic contributions.

## REFERENCES

1. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: An historical and theoretical perspective. *Med Teach*. 2013;35(9):e1437-46.
2. Patrício M, Julião M, Fareleira F, Young M, Norman G, Vaz Carneiro A. A comprehensive checklist for reporting the use of OSCEs. *Med Teach.* 2009;31(2):112-24.
3. Vincent SC, Arulappan J, Amirtharaj A, Matua GA, Al Hashmi I. Objective structured clinical examination vs traditional clinical examination to evaluate students' clinical competence: A systematic review of nursing faculty and students' perceptions and experiences. *Nurse Educ Today*. 2022;108:105170.
4. McKinley DW, Norcini JJ. How to set standards on performance-based examinations: AMEE Guide No. 85. *Med Teach*. 2014;36(2):97-110.
5. Ben-David MF. AMEE Guide No. 18: Standard setting in student assessment. *Med Teach.* 2000;22(2):120-30.
6. Jalili M, Hejri SM, Norcini JJ. Comparison of two methods of standard setting: the performance of the three-level Angoff method. *Med Educ.* 2011;45(12):1199-208.
7. Chaz Sardi MC, Martinez CK, Mirofsky MA, Lopez FJ, Garzaniti R, Gubilei ES, et al. Multiempleo en salud en provincia de Buenos Aires: estudio transversal de profesiones afectadas al cuidado de pacientes con COVID-19. *Rev Argent Salud Pública*. 2023;15:e89.
8. Kramer A, Muijtjens A, Jansen K, Dusman H, Tan L, van der Vleuten C. Comparison of a rational and an empirical standard setting procedure for an OSCE. *Med Educ.* 2003;37(2):132-9.
9. Gempp R. Coeficiente Phi(Lambda) y la fiabilidad de las decisiones sobre selección de personal. *Rev Psicol (Santiago)*. 2014;23(1):12-20.
10. Brennan RL, Kane MT. An index of dependability for mastery tests. *J Educ Meas*. 1977;14(3):277-89.
11. Shulruf B, Turner R, Poole P, Wilkinson T. The Objective Borderline method (OBM): a probability-based model for setting up an objective pass/fail cut-off score in medical programme assessments. *Adv Health Sci Educ Theory Pract*. 2013;18(2):231-44.
12. Boursicot KAM, Roberts TE, Pell G. Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Med Educ*. 2007;41(11):1024-31.
13. Cusimano MD, Rothman AI. The Effect of Incorporating Normative Data into a Criterion-Referenced Standard Setting in Medical Education. *Acad Med.* 2003;78(10 Suppl):S88-90.
14. Malau-Aduli BS, Teague PA, D'Souza K, Heal C, Turner R, Garne DL, et al. A collaborative comparison of objective structured clinical examination (OSCE) standard setting methods at Australian medical schools. *Med Teach.* 2017;39(12):1261-7.
15. Kaufman DM. Applying educational theory in practice. *BMJ.* 2003;326(7382):213-6.
16. Ilgen JS, Ma IWY, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Med Educ.* 2015;49(2):161-73.
17. del Valle M. Desarrollo de un cuestionario de indagación del proceso cognitivo de los docentes de medicina en la evaluación basada en observaciones [Tesis de Doctorado]. [Buenos Aires]: Instituto Universitario Hospital Italiano de Buenos Aires; 2023. [Accessed on: May 10, 2025]. Available at: https://trovare.hospitalitaliano.org.ar/ descargas/tesisytr/20240618145036/tesis-valle-marta.pdf
18. Kamal D, Sallam M, Gouda E, Fouad S. Is There a "Best" Method for Standard Setting in OSCE Exams? Comparison between Four Methods (A Cross-Sectional Descriptive Study). *J Med Educ.* 2020;19(1):e106600.
19. Homer M, Fuller R, Hallam J, Pell G. Setting defensible standards in small cohort OSCEs: Understanding better when borderline regression can 'work.' *Med Teach.* 2020;42(3):306-15.
20. Cole G, Dupre J. Constraints on reducing the costs of high-stakes OSCEs. *Med Teach.* 2016;38(11):1182.
21. Pitarque R. OSCE: Teoría y experiencia práctica: evaluación de competencias clínicas en la Facultad de Ciencias de la Salud, UNICEN. Olavarría: UNICEN; 2019.
22. Di laila S, Manjarin M, Torres F, Ossorio MF, Wainztein R, Ferrero F. Empleo del examen clínico objetivo estructurado (OSCE) en diversos niveles de educación de la pediatría. *Rev Fac Cienc Med Cordoba*. 2014;71(2):94-7.
23. Homer M, Russell J. Conjunctive standards in OSCEs: The why and the how of number of stations passed criteria. *Med Teach*. 2021;43(4):448-55.