Arch Argent Pediatr. 2025;e202510758

Estableciendo estándares en Evaluación Clínica Objetiva Estructurada (ECOE) de graduación en medicina: comparación de los métodos de grupo límite y Hofstee

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

RESUMEN

Introducción. Establecer el punto de corte en exámenes clínicos objetivos estructurados (ECOE) es un aspecto controvertido de la evaluación. En contextos de recursos limitados, el método Hofstee requiere tareas adicionales de otros docentes fuera del momento del examen, mientras que el método de grupo límite se aplica durante la evaluación, lo que permite un uso más eficiente del tiempo y los recursos.

Objetivo. Comparar la confiabilidad de los métodos de grupo límite y Hofstee aplicados en un ECOE de graduación en una universidad pública argentina, aportando evidencia local a un debate de relevancia internacional.

Población y métodos. Estudio transversal sobre 56 estudiantes en un ECOE de 12 estaciones. Se aplicaron dos métodos de fijación de estándares: grupo límite (mediante observadores durante el examen) y Hofstee (consulta electrónica a jueces expertos). Se compararon los puntos de corte, porcentaje de desaprobación y fiabilidad (coeficiente phi λ) mediante teoría de la generalizabilidad.

Resultados. El puntaje promedio fue 66,1 (DE 4,7). El método de grupo límite arrojó un punto de corte de 54 (global) con fiabilidad 0,89 y ningún desaprobado. El método Hofstee definió puntos de corte de 60,7 (global), con 3 y 1 estudiantes desaprobados respectivamente, y fiabilidad 0,68 y 0,82.

Conclusiones. Ambos métodos presentan una fiabilidad adecuada; no obstante, difieren en sus consecuencias prácticas, ya que el método de grupo límite resultó más benévolo al generar un mayor número de estudiantes aprobados.

Palabras clave: *competencia profesional; competencia clínica; evaluación educacional; estudiantes de medicina; educación de pregrado en medicina.*

doi (español): http://dx.doi.org/10.5546/aap.2025-10758
doi (inglés): http://dx.doi.org/10.5546/aap.2025-10758.eng

Cómo citar: García Diéguez MR, del Valle MP, Cragno AG. Estableciendo estándares en Evaluación Clínica Objetiva Estructurada (ECOE) de graduación en medicina: comparación de los métodos de grupo límite y Hofstee. *Arch Argent Pediatr.* 2025;e202510758. Primero en Internet 28-AGO-2025.

[1] *Centro de Estudios en Educación de Profesionales de la Salud (CEEProS), Departamento de Ciencias de la Salud, Universidad Nacional del Sur, Bahía Blanca, Argentina.*

Correspondencia para *Marcelo R. García Diéguez: mgdieguez@uns.edu.ar*

Financiamiento*: Ninguno.*

Conflicto de intereses*: Ninguno que declarar.*

Recibido*: 26-5-2025*
Aceptado*: 3-7-2025*

INTRODUCCIÓN

La evaluación de competencias clínicas en medicina ha evolucionado significativamente en las últimas décadas, consolidando al examen clínico objetivo estructurado (ECOE, u OSCE por sus siglas en inglés) como una estrategia confiable y válida para evaluar el desempeño de los estudiantes.[1] A través de estaciones clínicas estandarizadas que reproducen situaciones habituales de la práctica médica, el ECOE permite evaluar habilidades como el examen físico, la comunicación o el razonamiento clínico mediante observación directa y listas de cotejo estructuradas,[2] superando así limitaciones de evaluaciones tradicionales más subjetivas.[3]

Un aspecto central en su implementación es la determinación del punto de corte, es decir, el umbral que establece si un estudiante alcanza el nivel mínimo aceptable de competencia. Los métodos para fijarlo se clasifican en tres tipos: normativos (basados en el grupo), empíricos o centrados en el examinado (como el método de grupo límite), y de compromiso, juicio o consenso (como Hofstee, combinando criterios empíricos y normativos).[4,5]

El método Hofstee requiere tareas adicionales y planificación previa o posterior al examen,[6] mientras que el de grupo límite puede aplicarse en simultáneo con la administración del ECOE, lo que favorece la eficiencia en contextos con recursos humanos limitados, como aquellos donde predominan docentes con dedicación parcial y multiempleo.[7] A pesar de resultados psicométricos dispares, ambos métodos han demostrado utilidad para decisiones de alto impacto como la acreditación o graduación.[8]

Este estudio busca comparar la confiabilidad y las implicancias prácticas de los métodos de grupo límite y Hofstee en un ECOE de graduación en una universidad pública argentina, aportando evidencia local a un debate internacional.

POBLACIÓN Y MÉTODOS

Se realizó un estudio transversal con 56 estudiantes de Medicina en su último año. El ECOE, utilizado como examen final integrador, incluyó 12 estaciones clínicas distribuidas en cuatro circuitos y dos turnos. Cada estación tenía un puntaje máximo de 100, utilizando listas de cotejo dicotómicas. El puntaje final fue el promedio de todas las estaciones, aplicándose un sistema de compensación total entre estaciones y dimensiones.

Se aplicaron dos métodos de determinación de puntos de corte: grupo límite y Hofstee. En el método de grupo límite, 48 docentes observaron el desempeño y asignaron, además de completar la lista de cotejo, una calificación global, clasificando a los estudiantes como sobresalientes, satisfactorios, límites ("*borderline*") o insatisfactorios. La media de las puntuaciones de los puntajes límites determinó el punto de corte por estación; su promedio estableció el estándar global.

Para el método Hofstee, 17 jueces completaron una encuesta electrónica con cuatro preguntas clave: porcentajes mínimo y máximo aceptables esperado de aprobados, y puntuaciones mínima y máxima para aprobar. Se construyó un gráfico de intersección entre la curva de distribución de puntuaciones del estudiante y los parámetros establecidos por los jueces, definiendo el punto de corte por estación; y global, por la suma de aquellos.

Se utilizó la teoría de la generalizabilidad ("*G-Theory*") para estimar la fiabilidad del examen y la consistencia de los puntos de corte. Primero, se realizó un estudio de generalizabilidad con diseño estudiante × estación (SdC/St), equivalente al alfa de Cronbach como estimación de la consistencia relativa de las puntuaciones de los estudiantes a lo largo de las estaciones. Luego, se calculó el coeficiente phi lambda [φ(λ)] para cada método, que estima la fiabilidad de las decisiones de aprobación, usando el *software* EduG. Método, que estima la proporción de varianza sistemática en las decisiones de aprobación/desaprobación atribuible al desempeño verdadero del estudiante, en lugar de al error de medición.[9] Se compararon los puntos de corte resultantes de cada método, así como el porcentaje de estudiantes reprobados y los valores del coeficiente[φ(λ)].

El ECOE analizado es obligatorio para graduarse. Previamente, se usaba un sistema basado en criterio. El Comité de Evaluación de la Carrera y el Comité de Examen Final de Carrera autorizaron la implementación de ambos métodos con la condición de que se adoptara como punto de corte el más benévolo para los estudiantes. Esta decisión fue informada a los estudiantes, garantizando que el estudio no tendría consecuencias negativas.

Este trabajo se desarrolló siguiendo las recomendaciones metodológicas de Patricio y col. para comunicar investigaciones sobre ECOE, asegurando una descripción clara y transparente.[2]

RESULTADOS

La puntuación promedio fue de 66,1 (DE = 4,7; rango: 56,4-77,5). El análisis por estaciones se presenta en la *Figura 1*, donde se puede observar la distribución de las puntuaciones individuales para cada estación.

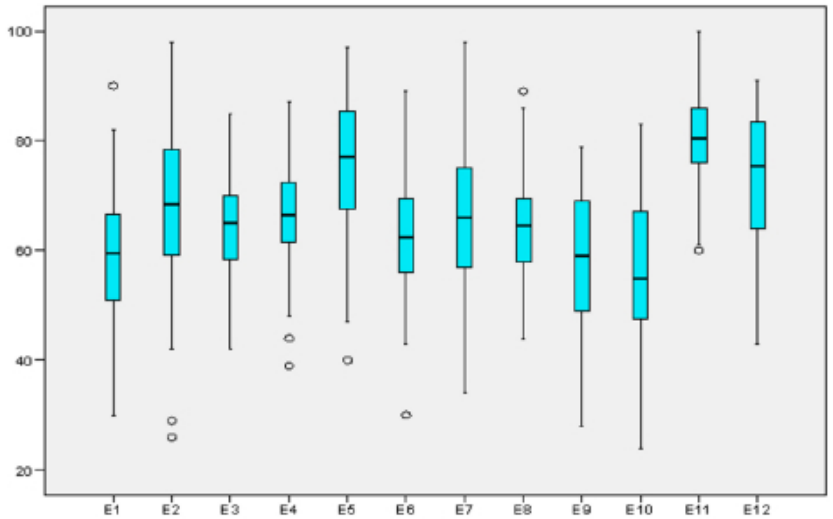
El método Hofstee definió puntos de corte de 60,7 (global) y 57,6 (por estación), con fiabilidad [φ(λ)] de 0,68 y 0,82, respectivamente. El grupo límite definió un punto de corte global de 54, con fiabilidad 0,89 y ningún desaprobado. No se estimó punto de corte por estación debido a la falta de suficientes observaciones límite en todas las estaciones.

El estudio de generalizabilidad mostró consistencia interna adecuada. En el análisis por facetas (circuito y turno), no se observaron diferencias significativas en ninguno de los dos.

En relación con la confiabilidad del instrumento, el estudio de generalizabilidad (*G Study*), basado en el diseño de medición estudiante × estación (SdC/St), arrojó un coeficiente G relativo –equivalente al alfa de Cronbach– que indica un nivel adecuado de consistencia interna de la evaluación en su conjunto. Para el análisis centrado en las decisiones de aprobación o desaprobación, se calculó el coeficiente phi lambda [φ(λ)], lo que permitió estimar la fiabilidad de las decisiones referidas a un criterio (*criterion-referenced*). Este coeficiente fue particularmente útil para comparar la consistencia de los puntos de corte obtenidos mediante los métodos grupo límite y Hofstee. El método de grupo límite presentó el valor más alto de fiabilidad decisional ([φ(λ)] = 0,89), lo que indica una alta consistencia en la clasificación de los estudiantes respecto al punto de corte. En comparación, el método Hofstee con puntuación global mostró una fiabilidad menor ([φ(λ)] = 0,68), mientras que el Hofstee por estación alcanzó un valor intermedio ([φ(λ)] = 0,82).

Durante la administración del examen, se identificaron 98 estaciones-estudiante (combinaciones de estudiante y estación) con desempeño clasificado como límite, lo que permitió establecer los puntos de corte específicos por estación con base en ese subgrupo. Los resultados comparativos entre ambos métodos de establecimiento de estándar, incluidos los valores de corte y los porcentajes de desaprobación asociados, se resumen en la *Tabla 1*.

FIGURA 1. Resumen de los resultados por estación



E: estación.

TABLA 1. Resumen comparativo del punto de corte para cada uno de los métodos

	Método Grupo límite	Hofstee puntuación global	Hofstee por estación
Puntaje de corte (máx. 100)	54	60,7	57,6
Cantidad de estudiantes reprobados (n = 56)	0	3	1
Coeficiente φ(λ)	0,89	0,68	0,82

DISCUSIÓN

Nuestros resultados muestran que tanto el método de grupo límite como el de Hofstee arrojaron puntos de corte con alta confiabilidad (0,89 y 0,82 respectivamente), lo que se alinea con otros estudios que emplearon la teoría de la generalizabilidad para evaluar la estabilidad de las decisiones de aprobación en ECOE.[8,10] En particular, el método de grupo límite demostró una mayor estabilidad en la clasificación de los estudiantes, reflejada en un coeficiente [φ(λ)] más alto, lo que indica decisiones más fiables al determinar la competencia clínica mínima esperada.

La comparación entre métodos muestra diferencias relevantes en términos de impacto educativo. El método de grupo límite, en línea con estudios previos,[11,12] tendió a generar puntos de corte más indulgentes: en este estudio, ningún estudiante fue desaprobado por este método. En contraste, el método de Hofstee resultó en uno o tres desaprobados según se aplicara a cada estación o a la puntuación global. Este hallazgo refleja lo informado por Cusimano y Rothman (2003), quienes señalaron que los métodos de compromiso, si bien pueden mejorar la percepción de justicia, pueden producir decisiones más exigentes.[13] En ninguno de los dos casos se tuvo en cuenta la corrección por el error de medida que seguramente incrementaría el número de desaprobados, pero no afectaría las diferencias observadas entre los métodos.

Algunos métodos como el Ansoff o Ebel requieren tareas preparatorias previas al examen. El método Hofstee permite una aplicación asincrónica, como fue realizado en este estudio mediante encuestas electrónicas, lo que podría facilitar su implementación en contextos distribuidos o virtuales, pero requiere un esfuerzo extra de docentes, idealmente diferentes de los evaluadores de las estaciones, lo que incrementa las necesidades logísticas y de recursos. A diferencia de esos otros métodos, el método de grupo límite tiene la ventaja operativa de realizarse simultáneamente con la evaluación, sin necesidad de convocatorias adicionales. Esta característica lo hace particularmente eficiente en contextos donde los docentes tienen dedicaciones parciales y la disponibilidad para tareas extracurriculares es limitada.[14,15]

El método de grupo límite, además, introduce un aspecto como es la impresión global del evaluador sin perder las ventajas de la lista de cotejo, ya que se combinan ambos métodos, uno para el puntaje y el otro para establecer el punto de corte. Varios estudios en la literatura han rescatado el valor de estas observaciones, ya que, si bien las listas de cotejo ofrecen una estructura detallada y objetiva, enfocándose en la observación de acciones específicas realizadas por el estudiante, las escalas globales permiten a los evaluadores emitir un juicio holístico sobre el desempeño del estudiante, considerando aspectos más amplios de competencia clínica. Algunos estudios indicaron una correlación significativa entre ambos métodos.[16] Sin embargo, el uso de escalas globales puede estar influenciado por la impresión general del evaluador, lo que podría introducir sesgos subjetivos en la evaluación.[17]

Estudios previos han mostrado que los métodos de compromiso como Hofstee pueden mejorar la percepción de justicia por parte de los estudiantes y reducir el sesgo individual de los evaluadores.[13] No obstante, también pueden implicar decisiones más estrictas que afecten la trayectoria académica de los estudiantes, por lo que su elección debe ser cuidadosa.[18]

Si bien algunos autores han planteado limitaciones del método de grupo límite en cohortes pequeñas, estudios recientes han demostrado que este puede ser confiable incluso en contextos con menos de 50 estudiantes, siempre que exista una adecuada dispersión del rendimiento y se utilicen escalas de calificación bien diseñadas.[19] En nuestro estudio, la fiabilidad alcanzada (índice G = 0,89 para el método de grupo límite) respalda esta observación y refuerza la aplicabilidad del método en contextos como el nuestro. Por otro lado, nuestro contexto de recursos limitados refuerza la necesidad de considerar la relación entre la viabilidad y la robustez de las decisiones. En línea con lo discutido por Cole y Dupre, en este marco, métodos como el de grupo límite, que aprovechan la interacción directa entre jueces y estudiantes sin requerir instancias adicionales, resultan particularmente valiosos por su bajo costo y capacidad de adaptación al entorno local.[20]

Un método utilizado extendidamente en las instituciones universitarias es un punto de corte absoluto tomando un porcentaje del puntaje máximo esperado o un porcentaje del mejor puntaje obtenido en ese examen.[21,22] Esta práctica presenta limitaciones, ya que, en general, produce puntos de corte más estrictos cuya confiablidad no podemos establecer, y fundamentalmente no considera la dificultad

del examen ni el desempeño global del grupo.[15] Nuestros hallazgos respaldan el uso de métodos basados en el desempeño observado para definir estándares más contextualmente adecuados.

Una limitación de este trabajo es el tamaño muestral relativamente pequeño, aunque consistente con otros estudios en contextos similares. Además, no se exploró la percepción de los estudiantes o docentes sobre los métodos utilizados, lo cual podría abordarse en estudios futuros.

Además, la literatura reciente ha problematizado el uso de estándares únicamente compensatorios en los ECOE como los utilizados en nuestra investigación, señalando que, en ausencia de criterios adicionales como el mínimo de estaciones aprobadas ("*conjunctive standards*"), algunos estudiantes podrían aprobar un examen global sin haber demostrado competencia en dominios clave.[23] En este sentido, la decisión de no incluir un criterio adicional de este tipo en nuestro examen podría haber favorecido una interpretación más laxa de la competencia global, pero también refleja un enfoque centrado en el rendimiento agregado, lo cual es coherente con los fundamentos del método de grupo límite.

CONCLUSIONES

Ambos métodos mostraron niveles aceptables de fiabilidad para establecer puntos de corte en un ECOE de graduación. Sin embargo, se observaron diferencias relevantes en su impacto práctico: el método de grupo límite ofreció mayor consistencia en la clasificación de los estudiantes y fue más benévolo en términos de número de aprobados que el método Hofstee. ■

Agradecimientos

Los autores desean agradecer a María Gabriela Serralunga por su contribución en el análisis psicométrico; al Prof. Carlos Brailovsky (†) por su impulso a la implementación de prácticas de evaluación válidas en nuestra universidad; y al equipo docente, estudiantes y al Centro de Estudios en Educación de Profesionales de la Salud por su apoyo, así como al Centro de Estudios en Educación de Profesionales de la Salud por su aporte logístico y académico.

REFERENCIAS

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. [Consulta: 10 de mayo de 2025]. Disponible en: 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. ECOE: Teoría y experiencia práctica: evaluación de competencias clínicas en la Facultad de Ciencias de la Salud, UNICEN. Olavarría: UNICEN; 2019.
22. Di Lalla S, Manjarin M, Torres F, Ossorio MF, Wainztein R, Ferrero F. Empleo del examen clínico objetivo estructurado (ECOE) 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.