ĐÁNH GIÁ TÍNH TIN CẬY GIỮA CÁC GIẢNG VIÊN TRONG LƯỢNG GIÁ SINH VIÊN Y KHOA NĂM THỨ 2 TRONG KỲ THI OSCE VỀ KỸ NĂNG KHÁM PHỔI

1. TỔNG QUAN TÀI LIỆU

Thi chạy trạm hay Phương pháp lượng giá các kỹ năng lâm sàng có cấu trúc khách quan (Objective Structured Clinical Examination – OSCE) là một trong những hình thức lượng giá được sử dụng rộng rãi trong đào tạo sinh viên (SV) y khoa trên toàn thế giới, phương pháp này hướng đến việc lượng giá năng lực của SV một cách khách quan và theo những quy chuẩn nhất định. Tuy nhiên, có một số yếu tố ảnh hưởng đến mức độ tin cậy của phương thức lượng giá bằng OSCE như cách thức tổ chức, thời gian của mỗi trạm OSCE, hệ thống tính điểm, bảng kiểm chấm điểm và người tham gia lượng giá. Trong bối cảnh cần lượng giá OSCE đối với số lượng lớn SV, việc áp dụng bảng kiểm trong thực hành chấm điểm, cũng như việc tuân theo các quy trình chuẩn trong lượng giá OSCE có thể không được đảm bảo, đưa đến sự khác biệt trong việc chấm điểm của người lượng giá, ảnh hưởng đến mức độ tin cậy của hình thức thi chạy trạm này.

Đối với SV Y khoa năm thứ 2 tại khoa Y, Đại Học Y DượcThành phố Hồ Chí Minh, các bạn tham gia vào các buổi học mô phỏng lâm sàng trong học phần Kỹ năng Y khoa Cơ bản gồm 60 tiết tại Trung tâm Huấn luyện Nâng cao về Mô phỏng Lâm sàng (ATCS). Các kỹ năng được đào tạo bao gồm các kỹ năng giao tiếp hỏi bệnh sử, kỹ năng thăm khám, kỹ năng điều dưỡng, các kỹ năng sơ cứu ban đầu… Trong năm học 2022 - 2023, có 415 SV tham gia vào học phần Kỹ năng Y khoa Cơ bản và được sắp xếp thành các nhóm học theo thời khóa biểu năm học 2022 – 2023, sau đó các SV thực hành các kỹ năng trong module Thực hành Y khoa (POM) tại các bệnh viện tuyến quận, tuyến huyện.

Việc lượng giá OSCE đánh giá năng lực thực hành về kỹ năng hỏi bệnh sử, kỹ năng thăm khám được diễn ra vào bốn ngày tại trung tâm ATCS. Trong đó, mỗi ngày sẽ có 2 phiên lượng giá sáng, chiều với số lượng SV trong mỗi ngày trong khoảng gần 110 SV. Mỗi SV được lượng giá qua 06 trạm gồm 05 trạm OSCE và 1 trạm nghỉ, 06 phút cho mỗi trạm OSCE và 01 phút di chuyển, đọc yêu cầu của đề thi. Các trạm thi bao gồm: hỏi bệnh sử, khám tổng quát, khám tim, khám phổi và khám bụng. Tại trạm khám Phổi, SV thực hành kỹ năng và giảng viên (GV) lượng giá sẽ quan sát trực tiếp để chấm điểm SV theo Bảng kiểm Đánh giá kỹ năng khám phổi phía trước và Bảng kiểm Đánh giá kỹ năng khám phổi phía sau. Hai bảng kiểm đã được chuẩn hóa về mặt nội dung.

GV tham gia lượng giá đã được tập huấn để sử dụng bảng kiểm trong việc chấm điểm cho SV tại trạm thi khám phổi.

Trong lĩnh vực sử dụng các thang đo về sức khỏe, Intraclass Correlation Coefficient (ICC) được xem là một trong những tiêu chuẩn đo lường thường được sử dụng để đo lường mức độ tin cậy giữa những người tham gia lượng giá. ICC được sử dụng ngày càng rộng rãi vì sự đơn giản và dễ áp dụng khi diễn giải kết quả. ICC có giá trị từ 0 đến 1, và khi giá trị ICC nhỏ hơn 0.5 là mức độ tin cậy thấp, từ 0.5 đến 0.75 là mức độ tin cậy trung bình, từ 0.75 đến 0.9 là mức độ tin cậy tốt, và nếu giá trị ICC trên 0.9 là mức độ tin cậy xuất sắc. Bên cạnh đó, chỉ số tương quan Pearson cũng được sử dụng để đánh giá mức độ tin cậy giữa những người lượng giá khi đánh giá các biến số liên tục. Giá trị của chỉ số Pearson từ -1 đến 1 và thể hiện mối quan hệ tuyến tính giữa các điểm số từ những GV khác nhau. Giá trị nhỏ hơn 0 thể hiện mối quan hệ ngược chiều và giá trị tuyệt đối của chỉ số Pearson càng cao thì sự tương quan giữa hai điểm số càng mạnh. Chỉ số Pearson nhỏ hơn 0.3 là tương quan yếu, từ 0.3 đến 0.5 là tương quan trung bình và lớn hơn 0.5 là tương quan mạnh. Ngoài ra, trong nghiên cứu này, chúng tôi tính chỉ số Fleiss Kappa để đo lường mức độ đồng thuận giữa các GV khi đánh giá kết quả kỹ năng khám của SV là đạt/ ranh giới/ rớt. Chỉ số này thể hiện mức độ tin cậy giữa những GV khi tham gia lượng giá một đối tượng hay một kỹ năng. Chỉ số Fleiss Kappa < 0.4 thể hiện mức độ đồng thuận thấp; từ 0.41 – 0.6 là mức độ đồng thuận vừa; > 0.6 là mức độ đồng thuận cao[1].

Chúng tôi tiến hành nghiên cứu này để đánh giá về độ tin cậy giữa những GV tham gia lượng giá OSCE cho kỹ năng khám phổi của SV Y khoa năm thứ 2 tại Đại Học Y Dược Thành Phố Hồ Chí Minh. Kết quả nghiên cứu có thể góp phần trong việc hoàn thiện được quy trình và gia tăng được độ tin cậy giữa các GV tham gia lượng giá tại trung tâm ATCS.

2. PHƯƠNG PHÁP NGHIÊN CỨU

Thiết kế nghiên cứu

Nghiên cứu cắt ngang được thực hiện tại trung tâm ATCS vào tháng 07/2023 đến tháng 09/2023.

Đối tượng nghiên cứu

Chúng tôi chọn toàn bộ các SV năm thứ 2 tham gia vào kỳ thi OSCE cuối kỳ vào ngày thi thứ 2 của 4 ngày thi (ngày 17/05/2023) với cỡ mẫu là 99 SV (tổ 37 – tổ 48).

Kết quả thi thu thập từ 05 GV tham gia chấm thi kỹ năng khám phổi tại trung tâm ATCS.

Tiến trình thu thập số liệu

Chúng tôi thu thập số liệu từ các video bài thi và kết quả thi của SV Y năm thứ 2 tham gia trong nghiên cứu. Các video bài thi và kết quả thi được lưu trong hệ thống lưu trữ của Trung tâm để đảm bảo được tính bảo mật của thông tin thi từ SV. Chúng tôi mã hóa mỗi bài thi theo mã định danh nghiên cứu. Một GV trong Bộ môn Nội tổng quát, phân môn Hô Hấp được mời tham gia chấm điểm lần thứ 2 cho những video bài thi của SV tham gia nghiên cứu. Hình thức chấm trực tiếp trên nền tảng Google form, và lưu kết quả chấm thi của GV này được lưu theo mã định danh của các bài thi để được so sánh với kết quả thi đã có trong buổi thi OSCE trực tiếp tại ATCS.

Phương pháp thống kê và phân tích số liệu

Chỉ số Kappa, chỉ số Pearson và ICC được sử dụng để ước tính mức độ tương quan giữa kết quả chấm thi của hai GV đối với kỹ năng khám phổi của SV tham gia nghiên cứu. Kết quả thi của bốn GV chấm trực tiếp SV trong phòng thi được so sánh với kết quả thi của GV thứ 5 chấm qua video ghi hình bài thi của SV. Chỉ số Pearson và ICC thể hiện mức độ tương quan giữa điểm thi giữa hai GV khác nhau. Trong khi đó, chỉ số Kappa thể hiện mức độ tương đồng trong việc đánh giá đậu/rớt của SV đối với kỹ năng khám phổi.

KẾT QUẢ NGHIÊN CỨU

Đặc điểm dân số nghiên cứu

Kết quả thi của 99 SV Y đa khoa năm thứ 2 được thu thập. Có bốn GV tham gia chấm hai kỹ năng khám phổi trước và khám phổi sau trong ngày thi cuối kỳ 17/05/2023 và một GV chấm lại trên các video ghi hình các bài thi của 99 SV trong khoảng thời gian 05/09/2023 – 20/09/2023. Năm GV hiện tại là GV Bộ môn Nội tổng quát. Trong đó, chỉ có GV 2 và GV 5 là thuộc Phân môn Hô hấp, còn GV 1 và GV 4 thuộc Phân môn Tiêu hóa và GV 3 thuộc Phân môn Thận. Cả năm GV đều tham gia giảng dạy tại ATCS đối với hai kỹ năng khám phổi trước và sau. Các GV đều đã tham gia các khóa tập huấn về Phát triển GV của trường, và được tập huấn sử dụng bảng kiểm trong việc lượng giá kỹ năng khám phổi của SV. Độ tuổi trung bình của các GV là 35 tuổi, trong đó có một GV nữ và bốn GV nam.

Mức độ tin cậy giữa những GV khác nhau

Kết quả về chỉ số tương quan Pearson và ICC giữa GV 5 chấm thi qua video ghi hình và bốn GV chấm thi trực tiếp được thể hiện trong bảng 1. Mức độ tương quan giữa điểm thi từ các cặp GV có sự khác biệt. GV 1 và GV 5 có sự tương quan cao nhất (chỉ số Pearson = 0.804 và chỉ số ICC =

0.889). Trong khi đó, tuy cùng lượng giá kỹ năng khám phổi trước, nhưng điểm thi của GV 2 và GV 5 có sự tương quan thấp (chỉ số Pearson = 0.320 và chỉ số ICC = 0.467). Đối với kỹ năng khám phổi sau, điểm thi của GV 5 và GV 3 có mức độ tương quan tốt trong khi mức độ tương quan giữa GV 5 và GV 4 chỉ ở trung bình.

Bảng 1: *Mức độ tương quan của lượng giá kỹ năng khám phổi giữa bốn cặp GV*

	Pearson's	**ICC (95% CI)**	**Số SV**	**Kỹ năng**
GV5 với GV1	0.804	0.889 (0.753 – 0.889)	26	Khám phổi trước
GV5 với GV2	0.320	0.467 (-0.170 – 0.757)	27	Khám phổi trước
GV5 với GV3	0.750	0.828 (0.594 – 0.927)	23	Khám phổi sau
GV5 với GV4	0.443	0.613 (0.087 – 0.836)	23	Khám phổi sau

Mức độ đồng thuận giữa những GV khác nhau

Chỉ số Fleiss Kappa được sử dụng để đo lường mức độ đồng thuận của các đánh giá đậu rớt giữa GV 5 và từng GV còn lại. Giá trị Fleiss Kappa nhỏ hơn 0 đối với cặp GV 5 và GV 2, và cặp GV 5 và GV 3 thể hiện mức độ đồng thuận rất thấp của việc đánh giá đậu/ranh giới/rớt (đặc biệt với phân loại "ranh giới"). Đối với giá trị Fleiss Kappa > 0.6 của cặp GV 5 và GV 1 thể hiện mức độ đồng thuận tốt. Các đánh giá đậu/ranh giới/rớt của GV 5 và GV 4 hoàn toàn tương tự nhau nên tỉ lệ đồng thuận trong việc đánh giá là 100%.

Bảng 2: *Mức độ đồng thuận của kết quả lượng giá kỹ năng khám phổi giữa bốn cặp GV*

	Fleiss Multirater Kappa (95% CI)
GV5 với GV1	0.646 (0.634 – 0.659)
GV5 với GV2	-0.038 (-0.051 – -0.026)
GV5 với GV3	-0.022 ((-0.035 – -0.009)
GV5 với GV4	Kết quả đánh giá đậu/rớt giống nhau 100%

BÀN LUẬN

Mức độ tương quan và mức độ đồng thuận

Kết quả nghiên cứu cho thấy mức độ tin cậy (tương quan và đồng thuận) của điểm thi và lượng giá giữa các cặp GV có sự khác biệt khá lớn trong cả hai kỹ năng khám phổi trước và khám phổi sau. Kết quả này tương tự một số nghiên cứu khác khi dùng ICC để so sánh mức độ tương quan giữa từng cặp chuyên gia[2]. Kết quả ICC thấp cũng cho thấy có các yếu tố ảnh hưởng đến quá trình chấm thi do một số đặc điểm khác biệt giữa các GV hay giữa các SV[3]. Trong nghiên cứu của chúng tôi, chỉ có hai GV tham gia chấm thi là GV của phân môn Hô hấp (GV 2 và GV 5). Tuy nhiên chỉ số tương quan giữa hai GV này thấp nhất trong các cặp GV. Cặp số 1 (giữa GV 1 thuộc Phân môn Tiêu hóa và GV 5 thuộc Phân môn Hô hấp) thì có các chỉ số về mức độ tin cậy giữa các GV lượng giá tốt nhất. Chúng tôi chưa khảo sát các yếu tố nhân khẩu và đặc tính của GV tham gia lượng giá có thể ảnh hưởng đến mức độ tin cậy giữa các GV chấm thi. Một số nghiên cứu chỉ ra các yếu tố có thể ảnh hưởng đến chỉ số tương quan giữa các GV tham gia lượng giá SV, có thể kể đến như giới tính của GV, mức độ thâm niên của GV trong lĩnh vực mô phỏng lâm sàng, hay GV có tham gia giảng dạy lâm sàng không[2]. Tương tự, một số nghiên cứu chỉ ra các yếu tố có thể ảnh hưởng đến sự đồng thuận trong lượng giá OSCE của GV như sự đa dạng của SV, đặc điểm tính cách của GV, ấn tượng chung về SV hay kỳ vọng của GV về kỹ năng đang lượng giá SV[4-6].

Những điểm hạn chế của nghiên cứu

Nghiên cứu bước đầu cho thấy mức độ tin cậy trong việc lượng giá kỹ năng khám phổi trước và khám phổi sau từ các GV khác nhau. Số lượng SV đối với mỗi cặp GV cho mỗi kỹ năng còn thấp, nên mẫu nghiên cứu chưa đại diện được mức độ tin cậy chung của toàn bộ GV tham gia lượng giá tại ATCS. Tuy nhiên, khi so sánh với một GV được xác định là chấm theo biểu mẫu chuẩn của kỹ năng khám phổi qua video ghi hình các bài thi, mức độ đồng thuận thấp của GV 5 với GV 2 và GV 3 trong việc xác định đậu/ranh giới/rớt đưa ra vấn đề về việc lượng giá kết quả thi của SV có sự khác biệt lớn giữa các GV. Những nghiên cứu tiếp theo cần tìm hiểu những yếu tố ảnh hưởng đến việc chấm thi của GV khi chấm thi trực tiếp với SV hay chấm thi qua video ghi hình bài thi của SV. Từ đó, xác định các yếu tố có thể hoàn thiên thêm trong việc lượng giá kỹ năng khám phổi của SV tại ATCS.

Một điểm hạn chế khác của nghiên cứu là số lượng SV được chấm điểm bắt cặp trong cả bốn nhóm ít hơn số lượng mẫu khuyến nghị là 80 SV trong mỗi nhóm (trong nghiên cứu này, chúng tôi chỉ có nhỏ hơn 30 SV trong mỗi nhóm). Ngoài ra, chúng tôi chọn ngẫu nhiên toàn bộ SV tham gia

trong ngày thi thứ 2 để giảm bớt một số yếu tố gây nhiễu ảnh hưởng đến GV chấm thi. Tuy nhiên, kết quả có được từ 99 SV này không đại diện cho toàn bộ SV y khoa năm thứ 2. Sau nghiên cứu này, chúng tôi sẽ mở rộng nghiên cứu đến nhiều đối tượng và tăng số lượng SV trong mỗi nhóm bắt cặp GV chấm thi. Nghiên cứu tiếp theo cũng cần đánh giá các điểm khác biệt giữa các GV và các yếu tố liên quan đến quá trình lượng giá của GV (thời gian chấm, chấm một lần khi chấm trực tiếp/xem lại nhiều lần qua video, sự tập trung, các yếu tố khác liên quan đến tương tác của GV và SV).

KẾT LUẬN

Kết quả nghiên cứu đã thể hiện mức độ tương quan và mức độ đồng thuận khác nhau giữa năm GV trong lượng giá kỹ năng khám phổi của 99 SV tham gia thi OSCE cuối kỳ tại ATCS. Khi so sánh với cùng một GV chấm thi qua video ghi hình bài thi của SV, điểm thi và các đánh giá đậu/ranh giới/rớt của bốn GV còn lại là khác nhau về mức độ tương quan. Về mức độ đồng thuận, sự khác biệt cũng khác biệt giữa từng cặp GV. Nghiên cứu chỉ ra bên cạnh phát triển một bảng kiểm có tính giá trị về mặt nội dung và việc tập huấn sử dụng bảng kiểm đối với GV, việc cải thiện và chuẩn hóa quá trình lượng giá của GV có thể góp phần quan trọng trong cải thiện tính tin cậy về việc lượng giá giữa các GV. Các bước tiến hành nghiên cứu so sánh về độ tin cậy giữa các GV là khả thi và có thể thực hiện lại một cách hệ thống và quy mô lớn hơn. Tuy nhiên, cần chú ý đến các nguồn lực cần thiết về nhân lực chấm lại các video bài thi, hệ thống lưu trữ dữ liệu và thu thập dữ liệu.

TÀI LIỆU THAM KHẢO

1. Gwet KL. Handbook of Inter-Rater Reliability. *Gaithersburg : Advanced Analytics, LLC*. 4 ed. 2014:chap 6.
2. Mortsiefer A, Karger A, Rotthoff T, Raski B, Pentzek M. Examiner characteristics and interrater reliability in a communication OSCE. *Patient Education and Counseling*. 01/24 2017;100doi:10.1016/j.pec.2017.01.013
3. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of chiropractic medicine*. Jun 2016;15(2):155-63. doi:10.1016/j.jcm.2016.02.012
4. Boursicot K, Kemp S, Wilkinson T, et al. Performance assessment: Consensus statement and recommendations from the 2020 Ottawa Conference. *Medical Teacher*. 2021/01/02 2021;43(1):58-67. doi:10.1080/0142159X.2020.1830052

5. Schleicher I, Leitner K, Juenger J, et al. Examiner effect on the objective structured clinical exam – a study at five medical schools. *BMC Medical Education*. 2017/04/24 2017;17(1):71. doi:10.1186/s12909-017-0908-1
6. Kenny DA. PERSON: A General Model of Interpersonal Perception. *Personality and Social Psychology Review*. 2004/08/01 2004;8(3):265-280. doi:10.1207/s15327957pspr0803_3