

Đánh giá đề thi MCQ đầu vào và tương quan độ khó, độ phân cách các câu

Bùi Anh Tú^{1*}, Võ Đăng Khoa¹, Nguyễn Anh Vũ¹, Trần Thị Diệu², Phạm Dương Uyển Bình³,
Vĩnh Sơn³, Phạm Lê An⁴, Mai Phương Thảo⁵

¹Bộ môn Toán, Khoa Khoa học cơ bản, Đại học Y Dược Thành phố Hồ Chí Minh, Thành phố Hồ Chí Minh, Việt Nam

²Bộ môn Tin học, Khoa Khoa học cơ bản, Đại học Y Dược Thành phố Hồ Chí Minh, Thành phố Hồ Chí Minh, Việt Nam

³Phòng Đảm bảo chất lượng, Đại học Y Dược Thành phố Hồ Chí Minh, Thành phố Hồ Chí Minh, Việt Nam

⁴Trung tâm Y học Gia đình, Đại học Y Dược Thành phố Hồ Chí Minh, Thành phố Hồ Chí Minh, Việt Nam

⁵Phòng Đào tạo Sau Đại học, Đại học Y Dược Thành phố Hồ Chí Minh, Thành phố Hồ Chí Minh, Việt Nam

Tóm tắt

Mục tiêu: Xây dựng các câu hỏi trắc nghiệm (MCQs) phụ thuộc vào chương trình giảng dạy và cần toàn diện, đảm bảo tính ứng dụng để đánh giá kỹ năng nhận thức của sinh viên y khoa, việc sử dụng chỉ số phân cách D27%, tương quan điểm nhị phân r_{pbis} phù hợp câu có độ khó vừa phải p trong khoảng 0,6 – 0,7. Tuy nhiên đối với câu dễ $p > 0,8$ thì r_{pbis} có giá trị tốt nhưng D27% có giá trị thấp hơn. Do đó, nghiên cứu này nhằm phân tích mối quan hệ phức tạp của độ khó và các chỉ số phân cách D27%, độ phân cách tối đa $D_{max}27\%$, r_{pbis} của các câu hỏi dễ và rất dễ trong bài thi kiểm tra tuyển sinh đầu vào sau đại học môn Giải phẫu học tại Trường Đại học Y Dược Thành phố Hồ Chí Minh.

Đối tượng và phương pháp nghiên cứu: Nghiên cứu cắt ngang được áp dụng để phân tích bài kiểm tra đầu vào sau đại học môn Giải phẫu học gồm 120 câu, được thực hiện trong năm 2022. Nghiên cứu đã thu thập 325 kết quả kiểm tra từ một nhóm đa dạng thí sinh dự thi sau đại học. Phân tích câu đã được thực hiện để xác định độ khó theo lý thuyết tắc nghiệm cổ điển (CTT) cùng chỉ số phân cách D27% và r_{pbis} trên phần mềm phân tích Basicstat, $D_{max}27\%$ tính trên Excel, cũng như đếm tần số mỗi nhử không hoạt động (NPD) có số lựa chọn dưới 5%.

Kết quả: Chỉ số phân cách D27% thể hiện tương quan nghịch nhẹ với độ khó ($r = -0,34$, $p = 0,00012 < 0,01$). Các câu dễ/khó vừa phải ($p = 40\% - 80\%$) có tương thích phân cách tốt nhất theo D27% ($D27\% = 0,6 - 0,7$) và r_{pbis} . Tuy nhiên, trong nhóm câu hỏi dễ ($p > 0\%$), chỉ có 63% câu hỏi đạt được độ phân cách tốt theo chỉ số D27% ($D \geq 0,3$), và 94,5% khi sử dụng chỉ số r_{pbis} . Đáng chú ý, trong số 19 câu hỏi rất dễ ($p \geq 92\%$) có chỉ số D27% $< 0,3$ cần được chỉnh sửa hoặc loại bỏ. Tuy nhiên, có 15/19 câu hỏi nêu trên có độ phân cách tốt với chỉ số r_{pbis} trong khoảng từ 0,313 đến 0,531 do các câu này có số câu sai tập trung vào nhóm điểm số thấp dù số câu đúng hai nhóm cao và thấp gần như nhau. Đa số các câu dễ đều có từ 2 NPD trở lên nhưng có độ phân cách tốt từng mỗi nhử với r_{pbis} .

Kết luận: Hơn 80% các câu hỏi MCQ rất dễ $p > 0,8$ cũng như có trên 2 NPD trong bài kiểm tra nhưng có độ phân cách tốt với r_{pbis} . Điều này làm nổi bật mối quan hệ phức tạp giữa độ khó và độ phân cách câu. Đối với ngân hàng đề cấp trường hay quốc gia để chọn lựa câu hỏi nên sử dụng r_{pbis} .

Ngày nhận bài: 20-09-2024 / Ngày chấp nhận đăng bài: 18-10-2024 / Ngày đăng bài: 25-10-2024

*Tác giả liên hệ: Bùi Anh Tú. Bộ môn Toán, Khoa Khoa học cơ bản, Đại học Y Dược Thành phố Hồ Chí Minh, Thành phố Hồ Chí Minh, Việt Nam. E-mail: buianhtu@ump.edu.vn

© 2024 Bản quyền thuộc về Tạp chí Y học Thành phố Hồ Chí Minh.

Từ khóa: độ khó; độ phân cách tối đa; độ phân cách; tương quan điểm nhị phân rpbis; mỗi như không hoạt động; lý thuyết trắc nghiệm cổ điển

Abstract

EVALUATION OF ENTRANCE MCQ EXAMS AND THE CORRELATION BETWEEN QUESTION DIFFICULTY AND DISCRIMINATION

Bui Anh Tu, Vo Dang Khoa, Nguyen Anh Vu, Tran Thị Diệu, Pham Duong Uyen Binh, Vinh Son, Pham Le An, Mai Phuong Thao

Objective: The aim is to create multiple-choice questions (MCQs) that align with the curriculum and effectively evaluate the cognitive abilities of medical students. The questions should be complete and have a discrimination index (D) of 27% and suitable r_{pbis} . The question has a moderate difficulty level, denoted as p , falling within the range of 0.6 to 0.7. However, for simple words with a probability greater than 0.8, r_{pbis} demonstrates a favorable value, while $D_{27\%}$ has a poorer value. This study aims to analyze the relationship between difficulty and discrimination ($D_{27\%}$, $D_{max27\%}$, r_{pbis}) of the easy and very easy multiple-choice questions in the head anatomy graduate entrance exam for the master's training program at University of Medicine and Pharmacy at Ho Chi Minh City.

Methods: A cross-sectional research design was utilized to assess question 1 of the anatomy graduate entrance exam, which consisted of 120 questions and was administered in 2022. The study gathered 325 test results from a diverse set of applicants who were taking the postgraduate exam. The difficulty of the questions was determined using question analysis based on Classical Test Theory (CTT), along with the discrimination index $D_{27\%}$ and the point-biserial correlation (r_{pbis}) using the Basicstat software. The maximum discrimination ($D_{max27\%}$) was calculated in Excel, as well as the frequency of non-functional distractors (NPD) with less than 5% selection.

Results: The discrimination index D , which was 27%, exhibited a weak negative connection with difficulty ($r = -0.34$, $p = 0.00012 < 0.01$). The questions that were classified as easy/moderately difficult ($P = 40\% - 80\%$) showed the highest level of discrimination compatibility, as indicated by a $D_{27\%}$ value of 0.6 – 0.7 and r_{pbis} . However, within the category of easy questions (where the probability of success is greater than 70%), only 63% of the questions showed a significant distinction based on the D index, which requires a discrimination of at least 0.3. On the other hand, when utilizing the r_{pbis} index, 94.5% of the questions achieved good discrimination. Significantly, out of the 19 questions that are considered very easy (with a success rate of 92% or more), the D index of 27% is below the threshold of 0.3, indicating that these questions should be revised or removed. Out of the 19 questions listed, 15 of them show a significant distinction with the r_{pbis} index ranging from 0.313 to 0.531. This is because these questions have a concentrated amount of incorrect responses in the low score group, despite having nearly the same both high and low numbers of correct answers in both groups. The easiest questions have 2 or more NPDs but have good discrimination between each distractor with their r_{pbis} .

Conclusion: More than 80% of MCQ questions are very easy $p > 0.8$ as well as there are more than 2 NPDs in the test, but they show good discrimination based on point-biserial correlations (r_{pbis}). This underscores the intricate correlation between the level of difficulty and the discrimination power of the questions. When it comes to school or national question choice banks, it is advisable to utilize r_{pbis} .

Keywords: difficulty; maximum discrimination; discrimination index; point-biserial correlations; non-functional distractors; classical test theory

1. ĐẶT VẤN ĐỀ

Các chuyên gia giáo dục y khoa thừa nhận sự tương tác giữa đánh giá và học tập, và ở một mức độ lớn, đánh giá thúc đẩy việc học tập. Vì thế, phát triển một chiến lược đánh giá phù hợp là một phần quan trọng phát triển chương trình giảng dạy bền vững hiệu quả. Các kỳ thi cùng bài kiểm tra câu hỏi trắc nghiệm có nhiều lựa chọn (MCQs) là một phần quan trọng của quá trình giảng dạy Y sinh, nó giúp hiệu chỉnh kịp thời việc dạy và học trong khi chúng đang diễn ra để nâng cao kiến thức học viên cần có để hoàn thành công việc. Từ năm 1960 đã sử dụng MCQ để đánh giá kiến thức của sinh viên và trong lĩnh vực khoa học sức khỏe từ năm 1999, MCQ đã được đa dạng hóa sát hợp cho các kỳ thi tuyển, kiểm tra với các bậc học khác nhau. MCQ giúp hiểu được điểm mạnh, điểm yếu, lỗ hổng kiến thức học viên và cung cấp phản hồi cho giáo viên về các hoạt động giáo dục của họ (Sadler, 1998; Nicol, 2006; Hubbard, 1961).

Bài kiểm tra bằng câu hỏi trắc nghiệm (MCQ) là hình thức đánh giá bằng văn bản hiệu quả nhất vừa đáng tin cậy vừa có giá trị nhờ phạm vi nội dung rộng rãi. Đây là công cụ khách quan thích hợp hơn để đánh giá lĩnh vực kiến thức được sử dụng để đánh giá sinh viên y khoa chưa tốt nghiệp và sau đại học ở hầu hết các ngành y khoa [1]. Chúng có lợi thế là lấy mẫu các lĩnh vực kiến thức rộng một cách hiệu quả và đáng tin cậy với test blue print để đánh giá hoàn thành mục tiêu học tập và đủ kiến thức đáp ứng được công việc được huấn luyện. Nếu được xây dựng cẩn thận, MCQ (đặc biệt là loại câu trả lời một câu đúng tốt nhất) kiểm tra được kỹ năng tư duy bậc cao (Norman, 1995; Peitzman, 1990). Để đảm bảo tuyển chọn được người học sau đại học có đủ năng lực tham gia vào quá trình đào tạo cần sử dụng đề thi tuyển sinh có độ tin cậy và tính giá trị cao bao gồm các MCQs trong ngân hàng câu hỏi có độ khó phù hợp và có khả năng phân biệt tốt năng lực của thí sinh. Để có được điều đó, bước quan trọng đầu tiên trong việc phát triển các MCQ có giá trị là viết câu MCQ với các chuyên gia có kỹ năng cần thiết am tường công việc. Tuy nhiên, giá trị của MCQ không thể được xác định hoàn toàn chỉ bằng cách viết thành thạo các câu hỏi mà còn liên quan đến phân tích thống kê thông số câu MCQ như độ khó, chỉ số phân cách $D_{27\%}$, r_{pbis} giúp tìm ra những câu kém chất lượng cần cải thiện hoặc loại ra hay cải thiện chất lượng của các câu có thể được sử dụng lại trong các thử nghiệm tiếp theo. Nó cũng cung cấp thông tin phản hồi cho giáo viên để xác định những thay đổi trong tiêu

chuẩn giảng dạy, nội dung khóa học cần nhấn mạnh hơn hoặc rõ ràng hơn cũng như hình thành thói quen cải thiện kỹ năng xây dựng các bài kiểm tra của giảng viên (Ary, 2002). Sau phân tích câu MCQ thô, phân tích câu MCQ dựa trên bài làm của thí sinh là một quá trình kiểm tra đáp ứng của học sinh đối với các câu kiểm tra riêng lẻ để đánh giá chất lượng của các câu đó và chất lượng của toàn bộ bài kiểm tra.

Một thách thức lớn là các câu hỏi dễ ($p > 80\%$) thường có chỉ số phân cách thấp khi đánh giá bằng $D_{27\%}$, nhưng lại có thể đạt phân cách tốt với r_{pbis} . Để khắc phục điều này, khái niệm độ phân cách tối đa $D_{max27\%}$ được đề xuất nhằm hỗ trợ $D_{27\%}$ trong việc đánh giá các câu hỏi dễ. Nghiên cứu này sẽ phân tích mối liên hệ giữa độ khó và các chỉ số phân cách ($D_{27\%}$, $D_{max27\%}$, r_{pbis}) của các câu hỏi dễ và rất dễ trong kỳ thi tuyển sinh sau đại học môn Giải phẫu tại Đại học Y Dược TP. Hồ Chí Minh, đồng thời đánh giá hạn chế của $D_{27\%}$ và đề xuất $D_{max27\%}$ như một giải pháp bổ sung.

2. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Đối tượng nghiên cứu

Nghiên cứu được thực hiện trên bài thi tuyển sinh đầu vào sau đại học môn Giải phẫu học tại Đại học Y Dược Thành phố Hồ Chí Minh năm 2022, bao gồm 120 câu hỏi trắc nghiệm 4 lựa chọn. Nghiên cứu sử dụng dữ liệu phản hồi của 325 thí sinh tham gia tuyển sinh, là các sinh viên tốt nghiệp từ nhiều trường đại học y khoa ở Việt Nam. Các bài kiểm tra của các thí sinh đã được thu thập và phân tích để đánh giá các chỉ số độ khó, độ phân cách và số môi như không hoạt động của từng câu hỏi theo CTT.

2.1.1. Tiêu chuẩn chọn mẫu

Tất cả các bài thi có đủ dữ liệu liên quan đến tổng điểm và các đáp án đã chọn của thí sinh đều được chọn vào nghiên cứu.

2.1.2. Tiêu chuẩn loại trừ

Các bài thi không đầy đủ dữ liệu hoặc có lỗi trong việc ghi nhận kết quả (ví dụ như vi phạm quy chế thi cử hoặc sai sót trong quá trình chấm điểm).

2.2. Phương pháp nghiên cứu

2.2.1. Thiết kế nghiên cứu

Nghiên cứu được thực hiện theo thiết kế cắt ngang nhằm

đánh giá các câu hỏi trắc nghiệm (MCQs) và kết quả của các thí sinh trong kỳ thi tuyển sinh đầu vào sau đại học môn Giải phẫu học tại Đại học Y Dược TP. Hồ Chí Minh năm 2022.

2.2.2. Các tham số câu MCQ theo lý thuyết khảo thí cổ điển CTT

Độ khó của câu hỏi (giá trị P):

Là tỷ lệ thí sinh trả lời đúng câu hỏi đó trong tổng số thí sinh trả lời bài kiểm tra. Nó phản ánh câu dễ hay khó đối với thí sinh. Phạm vi p từ 0% - 100%: (< 30% = quá khó); (30% - 70% = được khuyến nghị, tốt hoặc chấp nhận được) và (>70% = quá dễ dàng). P dao động từ 50 - 60% được coi là lý tưởng và được chuyên gia đánh giá khuyến dùng [2, 3].

Độ phân cách D27% (Upper – Lower difference):

Phương pháp tính độ phân cách D27% dựa vào tổng điểm thô của từng thí sinh được mô tả lần đầu bởi Jonhson (1951). Đầu tiên, sắp xếp kết quả bài làm thí sinh theo tổng điểm thô giảm dần. Nhóm cao gồm 27% thí sinh đạt điểm cao tính từ trên xuống dưới và nhóm thấp gồm 27% thí sinh đạt điểm thấp. Khi đó:

D27% = Tỷ lệ trả lời đúng ở nhóm cao – Tỷ lệ trả lời đúng ở nhóm thấp.

Giá trị D27% càng cao thì câu càng có khả năng phân biệt giữa thí sinh giỏi và kém. Ưu điểm của D27% là việc tính toán và giải thích sẽ đơn giản hơn so với các chỉ số phân biệt khác như rpbis hay rbis. Tuy nhiên không có câu hỏi nào quá dễ hoặc quá khó mà có độ phân cách tốt nếu sử dụng chỉ số phân cách D27% để đánh giá Ebel RL (1991) [4].

Hệ số tương quan điểm nhị phân rpbis (point biserial correlation):

Là hệ số tương quan giữa tổng điểm thô trên toàn bài (X) và cách thí sinh trả lời câu hỏi (Y) (Y = 1 nếu trả lời đúng, Y = 0 nếu trả lời sai). Hệ số này được tính và giải thích tương tự như hệ số tương quan Pearson và được sử dụng để đánh giá khả năng phân cách của câu hỏi trắc nghiệm.

$$r_{pbis} = \frac{M_1 - M_0}{S_x} \sqrt{\frac{n_1 \times n_0}{n^2}}, \text{ với}$$

M1 và n1: Điểm trung bình và số lượng thí sinh trả lời đúng câu hỏi.

M0 và n0: Điểm trung bình và số lượng thí sinh trả lời sai câu hỏi.

Sx và n: Độ lệch chuẩn của tổng điểm bài thi và tổng số thí sinh.

Câu hỏi góp phần đánh giá năng lực thí sinh, câu hỏi có độ phân cách rpbis tốt khi câu hỏi và cả đề trắc nghiệm đều đo lường cùng một thứ. Giá trị rpbis cao cho thấy rằng những thí sinh có tổng điểm thi cao đã trả lời đúng câu hỏi, và ngược lại những thí sinh có tổng điểm thi thấp đã làm sai câu hỏi. Ưu điểm của việc sử dụng hệ số rpbis đối với D27% là trong rpbis mỗi người tham gia bài kiểm tra được sử dụng để tính rpbis trong khi chỉ 54% (27% trên + 27% dưới) được sử dụng để tính D27%. Hệ số tương quan điểm nhị phân rpbis có giá trị nằm trong khoảng từ -1 đến +1. Các giá trị dương và càng cao càng có khả năng phân cách, tức là cho biết người đạt tổng điểm bài thi cao đã trả lời đúng câu này nhiều hơn người có tổng điểm bài thi thấp. Nó phải là tích cực đối với câu trả lời đúng và tiêu cực đối với những câu trả lời sai tức là người có tổng điểm cao trả lời sai câu đó ít hơn người có tổng điểm thấp. rpbis bằng 1 là lý tưởng vì câu phân biệt hoàn hảo giữa học sinh có khả năng thấp và cao.

Theo nguyên tắc kinh nghiệm và được khuyến nghị bởi Ebel RL (1991), về các chỉ số phân cách D27%, rpbis (0,40 trở lên) được coi là câu rất tốt, xuất sắc và cần được giữ lại; (0,30 đến 0,39) là khá tốt nhưng có thể cần cải thiện; (0,20 đến 0,29) là những câu có lỗi không đáng kể nhưng có thể chấp nhận được và cần được sửa đổi; (< 0,2) được coi là câu kém và cần sửa đổi lớn hoặc cần loại bỏ; và (< 0) là loại tồi tệ nhất của các câu và chắc chắn phải bị loại bỏ. Các câu cần được xem xét cẩn thận vì có những nguyên nhân phổ biến dẫn đến sự phân cách kém như cách diễn đạt mơ hồ, từ khóa sai và các lãnh vực gây tranh cãi, rpbis tính được cho từng câu cũng như từng mỗi như và có ý nghĩa với phép kiểm T.test [5].

Độ phân cách tối đa Dmax 27% cho các câu hỏi dễ (p > 70%):

Xuất phát từ vấn đề khi phân tích câu hỏi dễ (p > 0,7) thường có khuynh hướng phân cách kém theo chỉ số D27% do số người làm đúng ở hai nhóm điểm cao và thấp gần giống nhau. Chúng tôi chỉ ra rằng khi độ khó p > 91,9% thì độ phân cách D27% < 0,3 và khi p > 94,6% thì D27% < 0,2. Vậy phải chăng tất cả các câu hỏi có độ khó lớn hơn 91,9% thì đều không có phân cách tốt và cần bị loại bỏ? Nhưng trên thực tế số khá lớn các câu này, khi xét độ phân cách toàn bài rpbis thì > 0,3 đều phân cách tốt nên chúng tôi đề nghị khái niệm độ phân cách tối đa Dmax27% quan tâm đến số thí sinh làm sai trong cả hai nhóm điểm số cao và thấp.

Gọi N là tổng số thí sinh dự thi. Do ta đang xét độ phân cách của câu hỏi dễ, là câu hỏi mà đa số thí sinh đều làm đúng, nên ta sẽ quan tâm đến những thí sinh làm sai và ta mong muốn câu hỏi này sẽ giúp ta phân biệt được số ít những thí sinh làm sai câu hỏi so với phần còn lại. Đặt ts là tổng số thí sinh làm sai câu hỏi, sc là tổng số thí sinh làm sai câu hỏi trong nhóm cao và st là tổng số thí sinh làm sai câu hỏi trong nhóm thấp. Khi đó, độ phân cách D27% của câu hỏi được tính lại là: $D = \frac{st - ts}{27\%N}$ hay D = Tỷ lệ trả lời sai ở nhóm thấp – Tỷ lệ trả lời sai ở nhóm cao [6].

Với câu hỏi dễ, độ phân cách D27% có thể đạt được giá trị lớn nhất khi tất cả thí sinh nhóm cao đều làm đúng đồng thời tất cả những thí sinh làm sai đều thuộc nhóm thấp. Vì là câu hỏi dễ nên chỉ có số ít thí sinh làm sai và khi câu hỏi dễ mà đạt được độ phân cách tối đa $D = D_{max}$, nghĩa là nó đã phân biệt được tất cả các thí sinh có năng lực thấp vào nhóm dưới. Trong trường hợp này, chúng tôi nói đây là câu hỏi có khả năng phân cách tốt với Dmax.

Mối liên hệ giữa độ phân cách tối đa của một câu hỏi với độ khó của nó: $D_{max} = \min(\frac{1-P}{27\%}, 1)$ [7].

Từ kết quả trên, Dmax rất hữu ích và có thể khắc phục hạn chế của chỉ số D27% khi đánh giá các câu hỏi rất dễ ($p \geq 92\%$), mà những câu như vậy thì tỷ lệ chọn cho phương án đúng cũng $\geq 92\%$. Do đó, tổng sự chọn lựa cho 3 phương án sai còn lại sẽ dưới $< 8\%$. Như vậy, chắc chắn sẽ có ít nhất 1 phương án sai có tỷ lệ chọn dưới 5%.

Mỗi như không hoạt động (non-functioning distractor; NFD):

Một phương án sai được chọn bởi ít hơn 5% thí sinh dự thi được coi là một mỗi như không hoạt động (NFD). Một phương án sai được coi là một mỗi như tốt khi số lượng thí sinh kém chọn nó vượt trội so với số lượng thí sinh giỏi, và sự chênh lệch này có ý nghĩa. Bằng cách phân tích các phương án lựa chọn sai, giúp chúng ta nhận ra các lỗi của chúng để có thể sửa đổi, thay thế hoặc loại bỏ.

Các nghiên cứu trước đây cho thấy một bộ MCQ có thể bao gồm một hoặc nhiều NFD. Rahma NA (2017) nhận thấy rằng 37,5% câu trong bài kiểm tra có ba NFD, 25% có hai NFD, 32,5% có một NFD và chỉ 5% câu có các phương án sai hoạt động tốt [8]. Deepak KK (2015) phát hiện rằng sự

không hoạt động của phương án sai ảnh hưởng đảo ngược đến độ tin cậy và chất lượng của bài kiểm tra [9]. Các phương án sai không hoạt động làm cho các câu hỏi dễ hơn và làm giảm đáng kể chỉ số phân biệt. Ba phương án sai không hoạt động trong một câu hỏi nhiều lựa chọn 5 phương án ảnh hưởng đáng kể đến tất cả các thuộc tính đo lường tâm lý. Tuy nhiên, nghiên cứu của Deepak KK đã sử dụng MCQ 5 lựa chọn thay vì MCQ 4 lựa chọn. Cũng trên các MCQ 5 lựa chọn, Hingorjo MR & Jaleel F (2012) kết luận rằng các câu hỏi có 3 phương án sai hoạt động thực hiện tốt nhất trong việc phân biệt giữa các sinh viên. Các câu hỏi có hai phương án sai không hoạt động, mặc dù dễ hơn, nhưng có khả năng phân biệt tốt hơn so với các câu hỏi không có phương án sai không hoạt động [10].

2.2.3. Phân tích thống kê

Các tham số của câu hỏi như độ khó, độ phân cách, số mỗi như không hoạt động được phân tích theo mô hình lý thuyết trắc nghiệm cổ điển CTT. Dữ liệu được phân tích bằng phần mềm BasicStat [11] và Excel. Các thống kê mô tả như tần số, trung bình và độ lệch chuẩn được sử dụng để tóm tắt dữ liệu. Hệ số tương quan Pearson được áp dụng để đánh giá mối quan hệ giữa độ khó và các chỉ số phân cách.

3. KẾT QUẢ

Đầu tiên, phân tích tổng quan đề thi, chúng tôi thu được kết quả sau: Độ tin cậy toàn bài. Hệ số KR - 20 = 0,9632.

Đây là một đề thi dễ (Độ khó trung bình là 78,1%), phần lớn câu hỏi trong đề thi ở mức dễ (76,7%). Nhưng đề có độ tin cậy KR - 20 = 0,9632, độ phân cách tốt, trung bình độ phân cách theo D27% hay rpbis đều lớn hơn 0,3 (Bảng 1, Bảng 2).

Trên toàn đề thi, xét theo chỉ số D27% thì có 67,5% câu hỏi có độ phân cách từ tốt đến rất tốt ($D \geq 0,3$), tương ứng với rpbis tỷ lệ này là 85% ($r_{pbis} \geq 0,3$) (Bảng 3). Tuy nhiên, khi xét từng câu hỏi một, chúng tôi nhận thấy có sự khác biệt đáng kể giữa chỉ số D27% và chỉ số rpbis. Đặc biệt, khi độ khó của câu hỏi giảm (tức câu hỏi dễ hơn), giá trị chỉ số rpbis có xu hướng cao hơn chỉ số D (Hình 1).

Về tương quan, chỉ số phân cách D27% thể hiện tương quan nghịch nhẹ với độ khó ($r = -0,34$, $p = 0,00012 < 0,01$) (Hình 2). Phân cách tốt nhất theo D27% ($D27\% = 0,6 - 0,7$)

được quan sát với các câu dễ/khó vừa phải ($p = 40\% - 80\%$).

Trên toàn bài thi, có 23 câu không có mỗi như không hoạt động chiếm tỷ lệ 19,2%, còn lại là các câu có ít nhất một mỗi như không hoạt động. Khi độ khó câu hỏi giảm, tức là câu hỏi càng dễ thì số lượng mỗi như không hoạt động tăng lên (Bảng 4).

Có 18/19 câu có 3 NFD (chiếm tỷ lệ 94,7%) và 1 câu có 2 NFD (chiếm tỷ lệ 5,3%). Tuy nhiên, không phải tất cả các câu hỏi này đều có phân cách không tốt hay mỗi như kém. Có đến 15 trong số 19 câu trên được xem là có phân cách tốt và mỗi như tốt lên (Bảng 5).

Do độ khó $p \geq 92\%$ câu rất dễ nên tất cả các câu trên đều có độ phân cách $D27\% < 0,3$, thậm chí có 3 câu (28, 30 và 48) có độ khó $p > 94,6\%$ nên độ phân cách $D27\% < 0,2$. Nếu chỉ dựa vào chỉ số $D27\%$ tính theo số thí sinh làm đúng, chúng ta có thể xem xét việc loại bỏ hầu hết các câu này. Tuy nhiên, nếu xem xét độ phân cách $D_{\max} 27\%$ theo số thí sinh làm sai, ta thấy sự chênh lệch giữa $D27\%$ và $D_{\max} 27\%$ là rất nhỏ, hầu hết những thí sinh làm sai đều thuộc nhóm điểm số thấp. Nghĩa là các câu này đã phân biệt được số ít những thí sinh có năng lực thấp so với phần còn lại, đây là dấu hiệu của câu có phân cách tốt. Tương tự với chỉ số phân cách r_{pbis} , ta thấy

tất cả các câu này đều có $r_{pbis} \geq 0,3$, thậm chí có đến 10/15 câu có $r_{pbis} > 0,4$ với p value $< 0,0001$ (Bảng 6).

Sử dụng phần mềm của Vinh Sơn để đánh giá chi tiết 15 câu trên, chúng tôi thấy có 14 câu tốt, có thể đưa vào ngân hàng câu hỏi và 1 câu (câu 30) cần chỉnh sửa lại mỗi như (Bảng 6).

Nhìn vào đáp án chi tiết của câu 28, đây là câu hỏi dễ, độ phân cách thực tế $D27\% = 0,18$ rất gần với độ phân cách tối đa mà nó có thể đạt được là $D_{\max} 27\% = 0,194$. Đáp án đúng (D) có $r_{pbis} = 0,493$ với p value $< 0,05$ và các phương án sai (lựa chọn A, B và C) đều có $r_{pbis} < 0$ với p value $< 0,05$. Rõ ràng đây là một câu hỏi tốt (Bảng 7).

Tương tự ở câu 30, đây là câu hỏi dễ, độ phân cách thực tế $D27\% = 0,16$ rất gần với độ phân cách tối đa mà nó có thể đạt được là $D_{\max} 27\% = 0,171$. Đáp án đúng (B) có $r_{pbis} = 0,394$ với p value $< 0,05$ và các phương án sai (lựa chọn C và D) đều có $r_{pbis} < 0$ với p value $< 0,05$, trừ lựa chọn A có $r_{pbis} = -0,076$ với p value $= 0,171 > 0,05$ chưa được tốt lắm cần cải thiện. Tuy nhiên, về cơ bản chúng ta thấy đây là một câu hỏi có độ phân cách tốt (Bảng 7).

Bảng 1. Phân phối độ khó và độ phân cách của đề thi Cao học Giải phẫu (n = 120 câu hỏi)

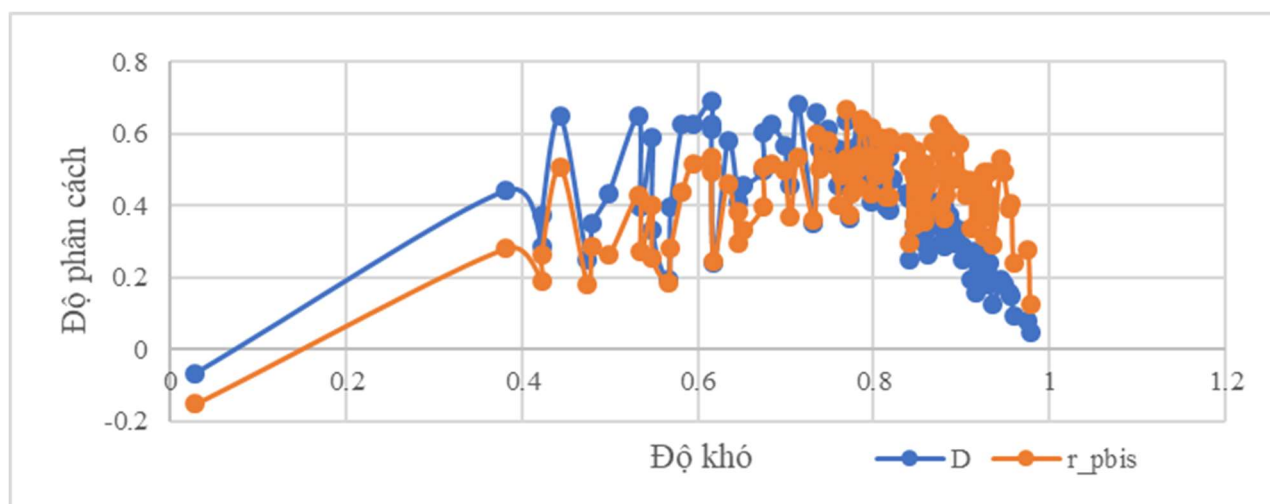
Độ khó (P) (%)		Độ phân cách (D27%)		Độ phân cách (r_{pbis})	
Mean $\pm$ SD	Range	Mean $\pm$ SD	Range	Mean $\pm$ SD	Range
78,1 $\pm$ 15,5	2,7 đến 97,8	0,387 $\pm$ 0,155	-0,068 đến 0,693	0,44 $\pm$ 0,12	-0,153 đến 0,668

Bảng 2. Phân phối độ khó của đề thi Cao học Giải phẫu (n = 120 câu hỏi)

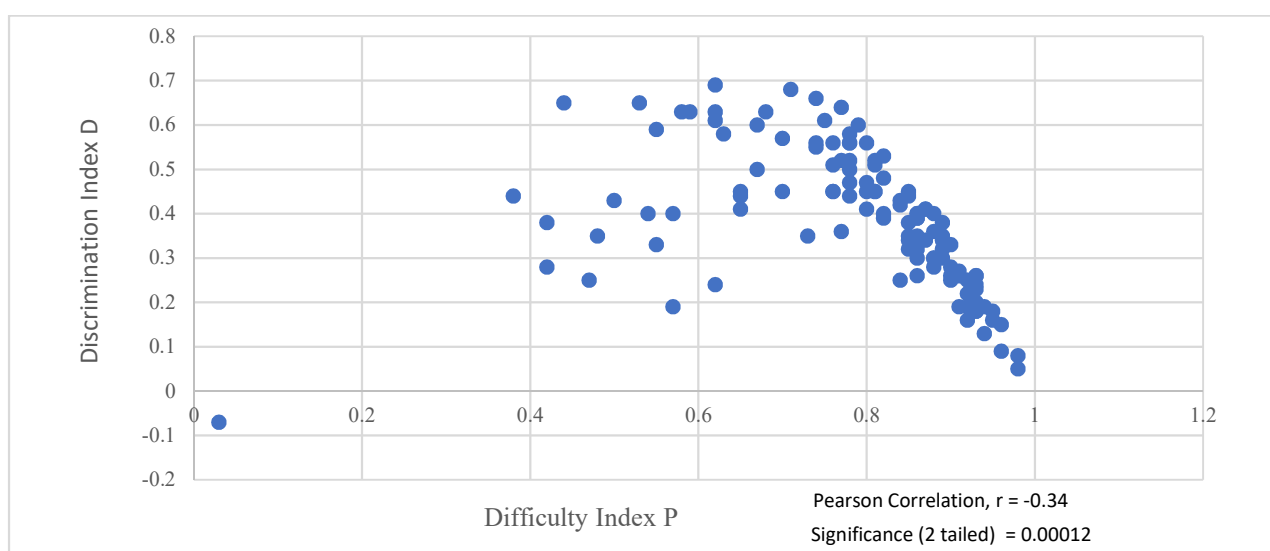
Độ khó	Khó ($p < 30\%$)	Trung bình ($30\% \leq p \leq 70\%$)	Dễ ($p > 70\%$)
% (Số câu)	0,8 (1)	22,5 (27)	76,7 (92)
Mean $\pm$ SD (%)	2,8	56,9 $\pm$ 8,85	85,2 $\pm$ 6,8

Bảng 3. Phân phối độ phân cách của đề thi Cao học Giải phẫu (n = 120 câu hỏi)

Độ phân cách (ID)		Rất tốt ($ID \geq 0,4$)	Tốt ($0,4 > ID \geq 0,3$)	Chấp nhận ($0,3 > ID \geq 0,2$)	Kém ($ID < 0,2$)
D	% (số câu)	45 (54)	22,5 (27)	19,2 (23)	13,3 (16)
	TB $\pm$ SD	0,526 $\pm$ 0,082	0,363 $\pm$ 0,028	0,261 $\pm$ 0,025	0,141 $\pm$ 0,072
r_{pbis}	% (số câu)	69,2 (83)	15,8 (19)	10,8 (13)	4,2 (5)
	TB $\pm$ SD	0,504 $\pm$ 0,06	0,369 $\pm$ 0,024	0,273 $\pm$ 0,018	0,106 $\pm$ 0,147



Hình 1. Phân phối độ phân cách theo độ khó của đề thi Cao học Giải phẫu (n =120 câu hỏi)



Hình 2. Tương quan độ khó và độ phân cách (D27%) của đề Cao học Giải phẫu (n = 120 câu hỏi)

Bảng 4. Phân phối mỗi như không hoạt động của đề Cao học Giải phẫu (n = 120 câu hỏi)

	Trên toàn bài thi (120 câu)			
Số phương án có dưới 5% TS chọn	0	1	2	3
Số câu hỏi	23	31	40	26
Tỷ lệ	19,2%	25,8%	33,3%	21,7%
	Trung bình (p ≤70%) (28 câu)			
Số phương án có dưới 5% TS chọn	0	1	2	3
Số câu hỏi	16	9	3	0
Tỷ lệ	57,14%	32,14%	10,71%	0
	Dễ (p >70%) (92 câu)			
Số phương án có dưới 5% TS chọn	0	1	2	3
Số câu hỏi	7	22	37	26
Tỷ lệ	7,61%	23,91%	40,22%	28,26%

Bảng 5. Kết quả thống kê trên 19 câu hỏi dễ ($p \geq 92\%$)

Câu	Số PA sai có tỷ lệ chọn dưới 5%	Tổng sai	Sai nhóm cao	Sai nhóm thấp	P	D _{max} 27%	D27%	r _{pbis}	p
1	3	8	0	7	0,975	0,011	0,080	0,277	0
5	3	24	0	23	0,926	0,274	0,261	0,491	0
6	3	22	0	21	0,932	0,251	0,239	0,493	0
28	3	17	0	16	0,948	0,194	0,182	0,493	0
29	3	21	1	12	0,935	0,239	0,125	0,292	0
30	3	15	0	14	0,954	0,171	0,159	0,394	0
48	3	14	0	13	0,957	0,16	0,148	0,407	0
52	3	13	0	8	0,960	0,148	0,091	0,241	0
64	3	22	0	16	0,932	0,251	0,182	0,37	0
67	3	24	0	23	0,926	0,274	0,261	0,488	0
76	3	24	0	16	0,926	0,274	0,182	0,313	0
83	3	22	0	18	0,932	0,251	0,205	0,427	0
85	3	24	1	21	0,926	0,274	0,227	0,479	0
90	3	18	0	17	0,945	0,205	0,193	0,531	0
98	3	22	0	17	0,932	0,251	0,193	0,398	0
106	3	25	0	22	0,923	0,285	0,25	0,455	0
114	2	25	0	17	0,923	0,285	0,193	0,389	0
117	3	26	2	21	0,92	0,296	0,216	0,405	0
118	3	7	0	4	0,978	0,080	0,045	0,124	0

Bảng 6. Kết quả thống kê 15 câu rất dễ ($P \geq 92\%$) nhưng có độ phân cách tốt dựa trên lựa chọn sai

Câu	Số PA sai có tỷ lệ chọn dưới 5%	Tổng sai	Sai nhóm cao	Sai nhóm thấp	P	D _{max} 27%	D27%	r _{pbis}
5	3	24	0	23	0,926	0,274	0,261	0,491
6	3	22	0	21	0,932	0,251	0,239	0,493
28	3	17	0	16	0,948	0,194	0,182	0,493
30	3	15	0	14	0,954	0,171	0,159	0,394
48	3	14	0	13	0,957	0,16	0,148	0,407
64	3	22	0	16	0,932	0,251	0,182	0,37
67	3	24	0	23	0,926	0,274	0,261	0,488
76	3	24	0	16	0,926	0,274	0,182	0,313
83	3	22	0	18	0,932	0,251	0,205	0,427
85	3	24	1	21	0,926	0,274	0,227	0,479
90	3	18	0	17	0,945	0,205	0,193	0,531
98	3	22	0	17	0,932	0,251	0,193	0,398
106	3	25	0	22	0,923	0,285	0,25	0,455
114	2	25	0	17	0,923	0,285	0,193	0,389
117	3	26	2	21	0,92	0,296	0,216	0,405

Bảng 7. Phân tích chi tiết một số câu hỏi trắc nghiệm theo CTT

Câu 28:

Độ khó: 0,95
Độ phân cách: 0,18
Quá dễ. Phân cách kém.
Độ phân cách tối đa tính theo độ khó: 0,194
Hệ số tương quan câu-bài:
 r_{pbis} : 0,493, $p=3,946E-19^*$

Lựa chọn	A	B	C	D*
Nhóm cao	0	0	0	88
Nhóm thấp	5	9	2	72
Tỉ lệ	3%	5%	1%	91%
r-pbis	-0,335	-0,329	-0,153	0,493
p value	0,000	0,000	0,006	0,000

Câu 30:

Độ khó: 0,95
Độ phân cách: 0,16
Quá dễ. Phân cách kém.
Độ phân cách tối đa tính theo độ khó: 0,171
Hệ số tương quan câu-bài:
 r_{pbis} : 0,394, $p=5,142E-13^*$

Lựa chọn	A	B*	C	D
Nhóm cao	0	88	0	0
Nhóm thấp	3	74	3	8
Tỉ lệ	2%	92%	2%	5%
r-pbis	-0,076	0,394	-0,197	-0,357
p value	0,171	0,000	0,000	0,000

4. BÀN LUẬN

Bài thi 120 câu trong nghiên cứu có tính giá trị do được soạn thảo bởi các chuyên gia Giải phẫu cũng như tin cậy hệ số KR-20 toàn bài là 0,96. Bài ở mức độ dễ với độ khó trung bình $78,1 \pm 15,5$, phân cách tốt với trung bình D27% $0,387 \pm 0,155$, trung bình r_{pbis} $0,44 \pm 0,12$.

Có 25,8% số câu trong bài thi có 1 NFD, 33,3% câu có 2 NFD và 21,7% số câu có 3 NFD là các môi như không hoạt động theo tiêu chí đếm thô khi số người chọn ít hơn 5%, chính điều này làm cho các câu trở nên dễ hơn. Tuy nhiên trong nghiên cứu chúng tôi bài thi 120 câu vẫn có độ tin cậy cũng như khả năng phân cách tốt.

Kết quả đạt được từ bài thi cho thấy đề thi có mức độ dễ và phần lớn câu hỏi nằm ở mức độ dễ. Tuy nhiên, đề thi có độ phân cách tốt, với tỷ lệ câu hỏi có độ phân cách tốt theo chỉ số D27% hay r_{pbis} là khá cao. Điều này cho thấy đề thi có khả năng phân biệt tốt giữa những thí sinh có kiến thức khác nhau.

Đối với câu hỏi dễ, chỉ số D27% thường không lớn. Đặc biệt khi độ khó của câu hỏi vượt ngưỡng 0,946 thì D27% sẽ giảm dưới 0,2. Trong nghiên cứu này, nhóm câu hỏi dễ ($p > 70\%$), chỉ có 63% câu hỏi đạt được độ phân cách tốt theo chỉ số D27% ($D \geq 0,3$), trong khi tỷ lệ này là 94,5% khi sử dụng chỉ số r_{pbis} . Đáng chú ý, trong số 19 câu hỏi rất dễ ($p \geq 92\%$), theo kết quả đã trình bày ở phần 1.4, tất cả các câu hỏi này sẽ cần được chỉnh sửa hoặc loại bỏ vì chỉ số D27% $< 0,3$. Tuy nhiên, trong số này, có 15 câu hỏi đã được xác định có độ phân cách tốt với chỉ số

r_{pbis} trong khoảng từ 0,313 đến 0,531, đây là các câu lượng giá mức kiến thức cơ bản cần giữ lại. Điều này làm rõ hạn chế của chỉ số D27% khi đánh giá độ phân cách của các câu hỏi dễ. Do đó, cần thận trọng khi dùng D27% để đánh giá loại bỏ câu hỏi dễ bằng cách xem xét độ phân cách tính theo r_{pbis} hoặc sử dụng kết hợp D27% và D_{max} 27%.

5. KẾT LUẬN

Đề thi được đánh giá là dễ và có độ phân cách tốt. Điều này cho thấy đề thi có thể đáp ứng mục tiêu đánh giá kiến thức cơ bản và phân biệt khả năng của thí sinh.

Không phải tất cả các câu hỏi dễ ($p > 91,9\%$) đều không có phân cách tốt. Việc chỉ sử dụng chỉ số D27% truyền thống có thể dẫn đến kết luận sai lầm và loại bỏ các câu hỏi khỏi ngân hàng một cách không chính xác. Nghiên cứu đã chỉ ra rằng đối với các câu hỏi dễ, chỉ số D27% dựa trên lựa chọn sai hay D_{max} 27% có thể hỗ trợ thầy cô đứng lớp trong việc đánh giá độ phân cách của câu hỏi tại lớp, độ phân cách r_{pbis} phù hợp hơn do đánh giá trên toàn bài nên dùng lựa chọn câu cấp độ khảo thí trường hay quốc gia.

Lời cảm ơn

Chúng tôi chân thành cảm ơn Phòng Đào tạo sau đại học, Phòng Khoa học Công nghệ, Phòng Đảm bảo Chất lượng Giáo dục và Khảo thí đã ủng hộ và cung cấp số liệu để hoàn thành nghiên cứu này.

Xung đột lợi ích

Không có xung đột lợi ích tiềm ẩn nào liên quan đến bài viết này được báo cáo.

ORCID

Bùi Anh Tú

<https://orcid.org/0000-0002-5050-5707>

Võ Đăng Khoa

<https://orcid.org/0000-0002-5798-6355>

Nguyễn Anh Vũ

<https://orcid.org/0009-0003-7148-8840>

Trần Thị Diệu

<https://orcid.org/0000-0002-2605-9003>

Phạm Dương Uyển Bình

<https://orcid.org/0000-0003-3398-3210>

Vĩnh Sơn

<https://orcid.org/0009-00054864-8601>

Phạm Lê An

<https://orcid.org/0000-0003-1186-0543>

Mai Phương Thảo

<https://orcid.org/0000-0002-6045-099X>

Đóng góp của các tác giả

Ý tưởng nghiên cứu: Phạm Lê An

Đề cương và phương pháp nghiên cứu: Vĩnh Sơn, Nguyễn Anh Vũ, Bùi Anh Tú, Phạm Dương Uyển Bình, Võ Đăng Khoa

Thu thập dữ liệu: Trần Thị Diệu

Giám sát nghiên cứu: Phạm Lê An

Nhập dữ liệu: Trần Thị Diệu

Quản lý dữ liệu: Vĩnh Sơn

Phân tích dữ liệu: Vĩnh Sơn

Viết bản thảo đầu tiên: Bùi Anh Tú

Góp ý bản thảo và đồng ý cho đăng bài: Phạm Lê An, Vĩnh Sơn, Nguyễn Anh Vũ, Mai Phương Thảo

Cung cấp dữ liệu và thông tin nghiên cứu

Tác giả liên hệ sẽ cung cấp dữ liệu nếu có yêu cầu từ Ban biên tập.

TÀI LIỆU THAM KHẢO

1. Afraa M, et al. Item difficulty & item discrimination as quality indicators of physiology MCQ examinations at the Faculty of Medicine Khartoum University. Khartoum Medical Journal. 2018;11:2.
2. Kelley TL. The selection of upper and lower groups for validation of test items. J Educ Psychol. 1939;30:17–24.
3. Shakurnia A, Ghafourian M, Khodadadi A, Ghadiri A, Amari A, Shariffat M. Evaluating functional and non-functional distractors and their relationship with difficulty and discrimination indices in four-option multiple-choice questions. Education in Medicine Journal. 2022;14(4):55–62.
4. Ebel RL and Frisbie DA. Essentials of educational measurement, fifth edition. Prentice Hall of India. NewDelhi-110001. 1991.
5. LeBlanc V & Cox MAA. Interpretation of the point-biserial correlation coefficient in the context of a school examination. The Quantitative Methods for Psychology. 2017;13(1):46–56.
6. Lâm Quang Thiệp. Đo lường trong giáo dục lý thuyết và ứng dụng. NXB Đại học quốc gia Hà Nội. 2010.
7. Dương Thiệu Tống (2000). Thống kê ứng dụng trong Nghiên cứu khoa học giáo dục, Phần I: Thống kê mô tả. NXB Đại học quốc gia Hà Nội.
8. Rahma NA, Shamad MM, Idris ME, et al. Comparison in the quality of distractors in three and four options type of multiple choice questions. Advances in Medical Education and Practice. 2017;8:287.
9. Deepak KK, Al-Umrán KU, Al-Sheikh MH, Dkoli BV & Al-Rubaish A. Psychometrics of Multiple Choice Questions with Non-Functioning Distractors: Implications to Medical Education. Indian Journal of Physiology and Pharmacology. 2015;59(4):428.
10. Hingorjo MR & Jaleel F. Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. JPMA-Journal of the Pakistan Medical Association. 2012;62(2):142.
11. Vĩnh Sơn. Phân tích câu hỏi trắc nghiệm theo lý thuyết trắc nghiệm cổ điển MCQ item analysis in classical test theory (CTT). 2010. http://basicstat.net/mcq2022/mcq_ett.php.