ĐÁNH GIÁ ĐỀ THI MCQ ĐẦU VÀO VÀ TƯƠNG QUAN ĐỘ KHÓ, ĐỘ PHÂN CÁCH CÁC CÂU

Bùi Anh Tú[1], Võ Đăng Khoa, Trần Thị Diệu, Vĩnh Sơn, Nguyễn Anh Vũ,
Phạm Dương Uyển Bình, Phạm Lê An

Đại học Y Dược Thành phố Hồ Chí Minh

TÓM TẮT

Tiêu chuẩn để đưa câu hỏi trắc nghiệm có nhiều lựa chọn (MCQs) vào ngân hàng câu hỏi là câu có độ phân cách tốt hoặc rất tốt (chỉ số phân cách lớn hơn hoặc bằng 0.3). Điều chỉnh câu hỏi có độ phân cách chấp nhận được (giá trị từ 0.2 đến dưới 0.3) và loại ra các câu có độ phân cách kém (giá trị bé hơn 0.2) (Robert L.Ebel, 1991). Nghiên cứu cắt ngang này nhằm mô tả đặc trưng của MCQs trong đề thi cao học đầu vào môn Giải phẫu và xác định mối tương quan giữa độ khó và độ phân cách D, đồng thời chỉ ra mặt hạn chế của chỉ số phân cách D trong việc đánh giá độ phân cách của câu hỏi.

ABSTRACT

EVALUATION MCQ ENTRY TEST AND CORRELATION OF DIFFICULTY, DISCRIMINANT INDEX OF ITEMS

*Bui Anh Tu, Vo Dang Khoa, Tran Thi Dieu, Vinh Son, Nguyen Anh Vu,
Pham Duong Uyen Binh, Pham Le An*

University of Medicine and Pharmacy at Ho Chi Minh City

The criteria for including multiple-choice questions (MCQs) in the question bank are questions with a good or excellent discrimination index (DI) (DI greater than or equal to 0.3). Questions with an acceptable discrimination index (DI between 0.2 and below 0.3) are adjusted, while questions with poor discrimination index (DI less than 0.2) are excluded (Robert L. Ebel, 1991). This cross-sectional study aims to describe the characteristics of MCQs in the entrance exam for Anatomy and determine the correlation between difficulty and the discrimination index D, while also highlighting the limitations of the discrimination index D in assessing question discriminability.

1. Đặt vấn đề

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 giúp hiệu chỉnh kịp thời việc dạy và học trong khi chúng đang diễn ra để nâng cao kiến thức học viên. Việc sử dụng các MCQ để đánh giá kiến thức của sinh viên đã có từ năm 1960 và trong lĩnh vực khoa học sức khoẻ từ năm 1999, MCQ đã được

[1] *Tác giả liên lạc*: **Bùi Anh Tú**, Đại học Y Dược Thành phố Hồ Chí Minh
Điện thoại: 0937501106 *Email:* buianhtu@ump.edu.vn

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).

Thiết kế MCQ tốt đảm bảo tiêu chuẩn và chất lượng là một quá trình phức tạp, đầy thách thức và yêu cầu đầu tư thời gian. MCQ loại "đáp ứng tốt nhất duy nhất" được thiết kế một cách rõ ràng để đánh giá kiến thức (Skakun, 1979). 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 đơn lẻ tốt nhất) kiểm tra được kỹ năng tư duy bậc cao (Norman, 1995; Peitzman, 1990). Do đó, MCQ vẫn là một công cụ đánh giá hữu ích, mặc dù nó có một số hạn chế và phản đối.

Phân tích câu MCQ 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. Thống kê thông số câu MCQ 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.

Mặc dù một số hình thức phân tích câu cơ bản của các bài kiểm tra MCQ có thể đã được thực hiện thường xuyên nhưng ít có bằng chứng nào cho thấy dữ liệu được tạo ra đã được sử dụng để giúp phát triển hoặc lựa chọn các câu MCQ tiếp theo (Si-Mui Sim, 2006; Zubairi, 2006).

1.1. Độ phân cách D (Upper – Lower difference) trong lý thuyết khảo thí cổ điển

Được giới thiệu lần đầu bởi Johnson (1951). Đầu tiên, sắp xếp bài làm thí sinh theo số câu đúng giảm dần. Chọn 27% thí sinh đầu danh sách làm nhóm cao và 27% cuối danh sách làm nhóm thấp. Khi đó, độ phân cách D của một câu hỏi bằng hiệu số giữa tỷ lệ trả lời đúng trong nhóm cao và tỷ lệ trả lời đúng trong nhóm thấp.

1.2. Độ phân cách tối đa D_{max} cho các câu hỏi dễ (P > 70%)

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 trả lời sai câu hỏi, sc là tổng số thí sinh làm sai câu hỏi trong nhóm cao và st là tổng số thí sinh trả lời sai câu hỏi trong nhóm thấp. Khi đó, độ phân cách D được tính lại là:

$$D = \frac{st - sc}{27\%N} \quad (1)$$

Độ phân cách tối đa có thể đạt được khi tất cả thí sinh nhóm cao đều làm đúng đồng thời hầu hết những thí sinh làm sai đều thuộc nhóm thấp. Ta được

$$D_{max} = \min(\frac{ts}{27\%N}, 1) \quad (2)$$

Vì là câu 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.

Mối liên hệ giữa độ phân cách tối đa của một câu hỏi với độ khó của nó

$$D_{max} = \min(\frac{1-P}{27\%}, 1) \quad (3)$$

Từ kết quả trên, khi độ khó P > 91.9% thì độ phân cách D < 0.3 và khi P > 94.6% thì D < 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?

1.3. Hệ số tương quan điểm nhị phân r_{pbis} (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**.**

MCQ của chúng ta "tốt" đến mức nào? Chúng có thực sự có thể phân biệt thành tích của học sinh trong các kỳ thi không? Chúng tôi đã cố gắng trả lời những câu hỏi này trong nghiên cứu này. Chúng tôi cũng đã cố gắng tìm ra mối quan hệ tồn tại giữa độ khó (P) và chỉ số phân cách (D, r_{pbis}) của các MCQ trong nghiên cứu này, đồng thời chỉ ra mặt hạn chế của chỉ số phân cách D.

2. Đối tượng và phương pháp nghiên cứu

Nghiên cứu cắt ngang mô tả với đề thi cao học đầu vào môn Giải phẫu có 120 câu trắc nghiệm 4 lựa chọn và 325 thí sinh tham gia, được phân tích về mức độ khó (P) và chỉ số phân cách (D, r_{pbis}). Số liệu được xử lý bằng phần mềm Excel, R và phần mềm của Ths. Vĩnh Sơn theo đường dẫn http://basicstat.net/mcq2022/.

3. Kết quả nghiên cứu

Đầu tiên, phân tích tổng quan đề thi, chúng tôi thu được kết quả sau:

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 (D)**		**Độ phân cách (r_{pbis})**	
Mean ± SD	Range	Mean ± SD	Range	Mean ± SD	Range
78.1 ± 15.5	2.7 đến 97.8	0.387 ± 0.155	-0.068 đến 0.693	0.44 ± 0.12	-0.153 đến 0.668

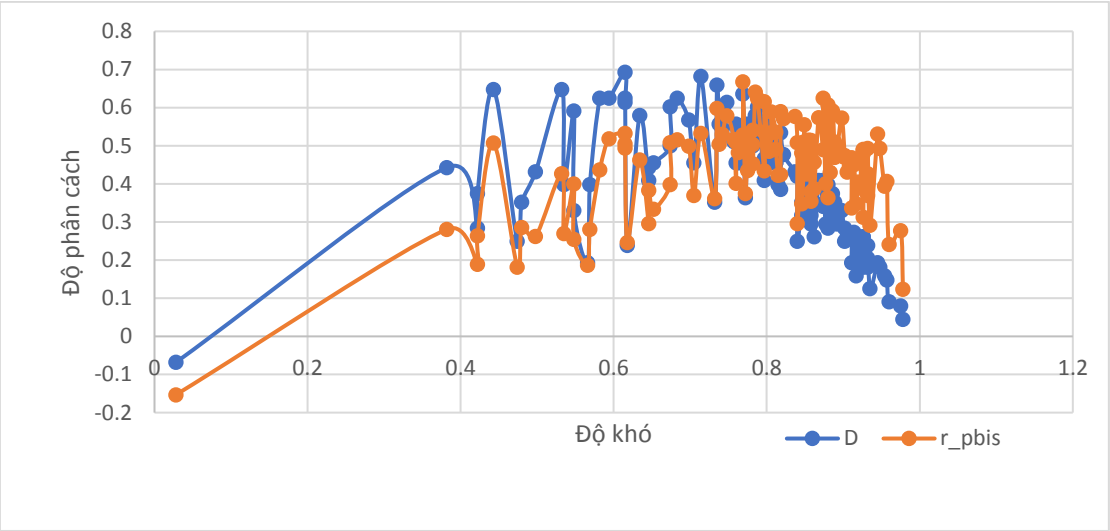
***Bảng 2**. Phân phối các mức độ khó của đề thi Cao học Giải phẫu (n = 120 câu hỏi)*

Mức độ khó	Khó (P < 30%)	Trung bình (30% ≤ P ≤ 70%)	Dễ (P > 70%)
% (Số câu)	0.8 (1)	22.5 (27)	76.7 (92)
Mean ± SD (%)	2.8	56.9 ± 8.85	85.2 ± 6.8

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

Mức độ phân biệt		Rất tốt (DI ≥ 0.4)	Tốt (0.4 > DI ≥ 0.3)	Chấp nhận (0.3 > DI ≥ 0.2)	Kém (DI < 0.2)
D	% (no.)	45 (54)	22.5 (27)	19.2 (23)	13.3 (16)
	Mean ± SD	0.526 ± 0.082	0.363 ± 0.028	0.261 ± 0.025	0.141 ± 0.072
r_{pbis}	% (no.)	69.2 (83)	15.8 (19)	10.8 (13)	4.2 (5)
	Mean ± SD	0.504 ± 0.06	0.369 ± 0.024	0.273 ± 0.018	0.106 ± 0.147

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 phân cách tốt, trung bình độ phân cách theo D hay r_{pbis} đều lớn hơn 0.3 (Bảng 1, Bảng 2).



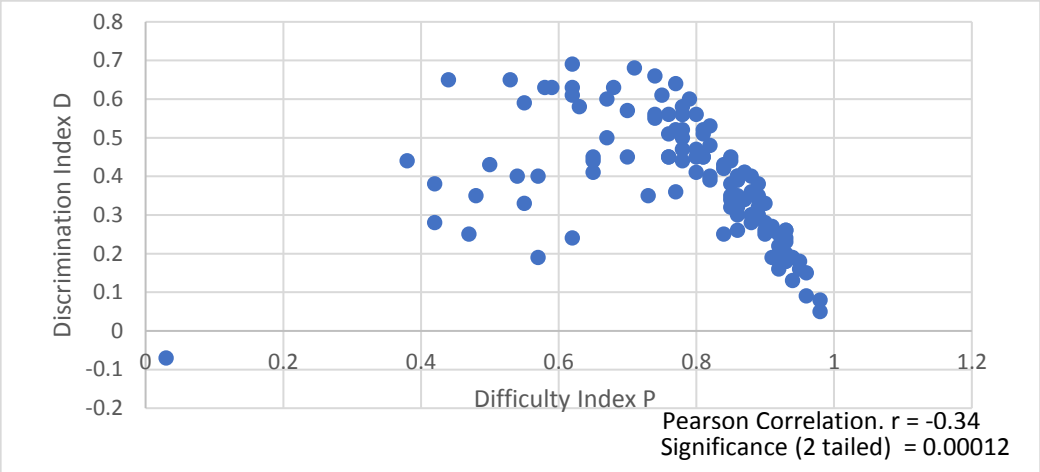
***Hình 1.** Phân phối độ phân cách theo độ khó của đề thi Cao học Giải phẫu (n =120 câu hỏi)*

Trên toàn đề thi, xét theo chỉ số D thì có 67.5% câu hỏi có độ phân cách từ tốt đến rất tốt (D ≥ 0.3), tương ứng với r_{pbis} tỷ lệ này là 85% (r_{pbis} ≥ 0.3) (Xem Bảng 3). Tuy nhiên, khi xét từng câu hỏi một, chúng tôi nhận thấy có sự khác biệt đáng kể giữa chỉ số D và chỉ số r_{pbis}. Đặc biệt, khi độ khó của câu hỏi giảm (tức câu hỏi dễ hơn), giá trị chỉ số r_{pbis} có xu hướng cao hơn chỉ số D (Xem Hình 1).

Trong nhóm câu hỏi dễ (P > 70%), chỉ có 63% câu hỏi đạt được độ phân cách tốt theo chỉ số D (D ≥ 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 ≥ 92%), theo kết quả đã trình bày ở phần 4.2, tất cả các câu hỏi này sẽ cần

được chỉnh sửa hoặc loại bỏ vì chỉ số D < 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 cơ bản cần giữ lại. Điều này làm rõ hạn chế của chỉ số D khi đánh giá độ phân cách của các câu hỏi dễ.

Về tương quan, chỉ số phân cách D thể hiện tương quan nghịch nhẹ với độ khó (r = -0.34, p = 0,00012 < 0.01) (Hình 2). Phân cách tốt nhất theo D (D = 0.6 – 0.7) được quan sát với các câu dễ/khó vừa phải (P = 40% - 80%).



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

Tiếp theo, phân tích 15/19 câu rất dễ (P ≥ 92%), chúng tôi thu được kết quả sau đây.

Bảng 4. *Các câu rất dễ (P ≥ 92%) nhưng có độ phân cách tốt (15 câu) dựa trên lựa chọn sai*

Câu	Tổng sai	Sai nhóm cao	Sai nhóm thấp	P	D_{max}	D	r_{pbis}	p
5	24	0	23	0.926	0.274	0.261	0.491	0
6	22	0	21	0.932	0.251	0.239	0.493	0
28	17	0	16	0.948	0.194	0.182	0.493	0
30	15	0	14	0.954	0.171	0.159	0.394	0
48	14	0	13	0.957	0.16	0.148	0.407	0
64	22	0	16	0.932	0.251	0.182	0.370	0
67	24	0	23	0.926	0.274	0.261	0.488	0
76	24	0	16	0.926	0.274	0.182	0.313	0
83	22	0	18	0.932	0.251	0.205	0.427	0
85	24	1	21	0.926	0.274	0.227	0.479	0
90	18	0	17	0.945	0.205	0.193	0.531	0
98	22	0	17	0.932	0.251	0.193	0.398	0
106	25	0	22	0.923	0.285	0.250	0.455	0
114	25	0	17	0.923	0.285	0.193	0.389	0
117	26	2	21	0.920	0.296	0.216	0.405	0

Do độ khó P ≥ 92% nên tất cả các câu trên đều có độ phân cách D < 0.3, thậm chí có 3 câu (28, 30 và 48) có độ khó P > 94.6% nên độ phân cách D < 0.2. Nếu chỉ dựa vào chỉ số D

tính theo số câu đúng, chúng ta có thể xem xét việc loại bỏ hầu hết các câu này. Tuy nhiên, nếu xem xét độ phân cách D tính theo số câu sai, ta thấy 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. Nếu dựa vào 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.

Sử dụng phần mềm của Ths. Vĩnh 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 mồi nhử.

Bảng 5. *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:
rpbis: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:
rpbis: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

Nhìn vào đáp án chi tiết của câu 28, đây là câu hỏi dễ, độ phân cách thực tế D = 0.18 rất gần với độ phân cách tối đa mà nó có thể đạt được là $D_{max} = 0.194$. 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.

Tương tự ở câu 30, đây là câu hỏi dễ, độ phân cách thực tế D = 0.16 rất gần với độ phân cách tối đa mà nó có thể đạt được là $D_{max} = 0.171$. 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.

4. Bàn luận

Kết quả đạt được từ bài thi cho thấy đề thi có mức độ khó là 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 D và 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 D thường không lớn. Đặc biệt khi độ khó của câu hỏi vượt ngưỡng 0.946 thì D sẽ giảm dưới 0.2. Do đó, cần thận trọng khi dùng D để đánh giá câu hỏi dễ bằng cách xem xét chỉ số D dựa trên lựa chọn sai kết hợp với $D_{max.}$ hoặc r_{pbis}.

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 D 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 D dựa trên lựa chọn sai và D_{max} có thể hỗ trợ thầy cô đứng lớp trong việc đánh giá độ phân cách của câu hỏi, r_{pbis} phù hợp hơn do đánh giá trên toàn bài nên dùng lựa chọn câu cấp độ khảo thí trường.

6. Tài liệu tham khảo

[1] GS.TSKH. LÂM QUANG THIỆP (2010). Đo lường trong giáo dục lý thuyết và ứng dụng. NXB Đại học quốc gia Hà Nội.

[2] TS. 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.

[3] Vĩnh Sơn (2010), **URL:** http://basicstat.net/mcq2022/mcq_ctt.php.

[4] Robert L.Ebel and David A. Frisbie (1991). Essentials of educational measurement, fitth edition. Prentice Hall of India. NewDelhi-110001.