

ỨNG DỤNG LÝ THUYẾT ỨNG ĐÁP CÂU HỎI TRONG PHÂN TÍCH ĐỀ THI ĐÁNH GIÁ NĂNG LỰC KIỂM ĐỊNH VIÊN KIỂM ĐỊNH CHẤT LƯỢNG GIÁO DỤC ĐẠI HỌC

APPLYING THE ITEM RESPONSE THEORY IN THE ANALYSIS OF COMPETENCY ASSESSMENT
EXAMINATIONS FOR HIGHER EDUCATION ACCREDITORS

Vũ Minh Hiếu^{1,2,+},
Tạ Thị Thu Hiền^{1,2}

¹Trung tâm Kiểm định chất lượng giáo dục - Đại học Quốc gia Hà Nội;

²Trường Đại học Giáo dục - Đại học Quốc gia Hà Nội

+Tác giả liên hệ • Email: hieuvvm@vnu.edu.vn

Article history

Received: 15/6/2026

Accepted: 22/7/2026

Published: 05/10/2026

Keywords

Item Response Theory,
Rasch model, accreditor,
competency assessment,
multiple-choice test,
professional training in
educational accreditation

ABSTRACT

Objective multiple-choice tests are among the most used assessment formats for evaluating learners' academic achievement and competencies upon completion of professional training programmes. However, to ensure the scientific rigor and reliability of assessment results, such tests need to be analysed using appropriate psychometric methods. This study employs Item Response Theory, specifically the Rasch model, to analyse a competency assessment test administered to learners after completing the accreditor's training program. The analysis yielded a Separation Reliability coefficient of 0,985. Candidates' scores ranged from 32,5 to 87,5 on a 100-point scale; the mean score was 66,03, with a standard deviation of 11,15; and 63,93% of candidates scored above the sample mean. Based on item difficulty, fit estimates and other indices based on the Rasch model, the 40 items were classified into four groups corresponding to recommendations on whether to retain, revise, or remove them. Based on the research findings, the article recommends further improvement of the process of developing, reviewing, piloting, and analysing competency assessment tests for learners after professional training in educational accreditation. These improvements aim to enhance test quality, refine the competency assessment instrument for learners, and contribute to strengthening the objectivity, reliability, and validity of assessment practices in professional training for educational accreditation.

1. Mở đầu

Trong hoạt động đào tạo và bồi dưỡng, đo lường là một trong những thành tố quan trọng của quá trình dạy và học (Amua-Sekyi, 2016). Việc đo lường được thiết kế và triển khai một cách khách quan, nghiêm túc, phù hợp về phương pháp và định hướng sẽ góp phần phản ánh trung thực mức độ đạt được về năng lực của người học (NH), đồng thời tạo ra những tác động tích cực đối với hoạt động giáo dục. Kiểm tra, đánh giá trong đào tạo có thể được thực hiện bằng nhiều hình thức khác nhau; trong đó, xu hướng lượng hóa ngày càng chú trọng các phương pháp đánh giá khách quan. Đặc biệt, bài kiểm tra sử dụng câu hỏi (CH) trắc nghiệm nhiều lựa chọn (Multiple Choice Questions - MCQ) được xem là một trong những hình thức phổ biến và được sử dụng rộng rãi trong bối cảnh đào tạo hiện nay (Butler, 2018; Haladyna, 2004). Theo đó, MCQ được sử dụng phổ biến trong đánh giá giáo dục do có ưu điểm về tính thuận tiện, khả năng chấm điểm nhanh, chính xác, đồng thời cho phép bao quát phạm vi nội dung tương đối rộng trong khoảng thời gian hạn chế (Gierl và cộng sự, 2017). So với CH tự luận hay trả lời ngắn, MCQ có lợi thế trong việc đo lường kiến thức trên nhiều nội dung, hạn chế ảnh hưởng của yếu tố chủ quan trong chấm điểm (Clauser và cộng sự, 2006). Tuy nhiên, việc xây dựng các CH MCQ có chất lượng không đơn giản, bởi quá trình này đòi hỏi phải tuân thủ nhiều nguyên tắc nhằm bảo đảm độ tin cậy và độ giá trị của CH (Dell và Wantuch, 2017; Frey và cộng sự, 2005). Nhiều nghiên cứu về chất lượng MCQ trong giáo dục đã chỉ ra rằng các sai sót trong thiết kế CH có thể dẫn đến việc diễn giải không chính xác kết quả của NH, từ đó ảnh hưởng đến các quyết định đánh giá (Downing, 2005; Haladyna, 2004). Đặc biệt, trong lĩnh vực khảo thí giáo dục, với mục đích khắc phục hạn chế của lý thuyết khảo thí cổ điển (chỉ cho phép biểu diễn độ khó của CH và năng lực của thí sinh (TS) trên cùng một thang đo logit), lý thuyết ứng đáp CH (Item Response Theory - IRT) đã được sử dụng để ước lượng đặc trưng của CH và năng lực tiềm ẩn của TS, đồng

thời đánh giá mức độ phù hợp của CH với mô hình và xem xét sự tương thích giữa độ khó của đề thi và năng lực của TS (Bond và Fox, 2001; Clauser và cộng sự, 2006).

Trong lĩnh vực kiểm định chất lượng (KĐCL) giáo dục đại học, chương trình bồi dưỡng nghiệp vụ kiểm định viên (KĐV) do các tổ chức KĐCL giáo dục thực hiện được triển khai theo quy định tại Thông tư số 14/2022/TT-BGDĐT của Bộ GD-ĐT (2022). Sau khi hoàn thành bồi dưỡng, việc tổ chức đánh giá năng lực cho NH không chỉ có ý nghĩa trong việc xác định mức độ đạt được các chuẩn đầu ra, mà còn giúp NH chuẩn bị cho kì thi sát hạch cấp thẻ KĐV do Bộ GD-ĐT tổ chức. Trong đó, bài thi lí thuyết gồm 40 CH trắc nghiệm khách quan, tập trung vào các kiến thức cốt lõi mà một KĐV cần có khi tham gia hoạt động KĐCL giáo dục. Tuy nhiên thực tế cho thấy, việc ứng dụng IRT/mô hình Rasch để phân tích thực nghiệm đề thi đánh giá năng lực KĐV, đặc biệt là sau chương trình bồi dưỡng nghiệp vụ KĐV còn chưa được nghiên cứu đầy đủ. Khoảng trống này đặt ra nhu cầu cung cấp bằng chứng định lượng về mức độ phù hợp của CH, sự tương thích giữa độ khó đề thi và năng lực TS, cũng như căn cứ khoa học để giữ lại, hiệu chỉnh, thử nghiệm lại hoặc thay thế các CH. Trên cơ sở đó, bài báo được thực hiện nhằm đánh giá đề thi đánh giá năng lực KĐV với đối tượng là NH sau khi hoàn thành chương trình bồi dưỡng nghiệp vụ KĐV. Thông qua việc phân tích mức độ phù hợp của dữ liệu với mô hình Rasch, ước lượng độ khó, sự tương thích giữa độ khó của đề thi và phân bố năng lực của TS và kết hợp các chỉ số khảo thí cổ điển, bài báo sẽ tập trung trả lời 03 CH: (1) Dữ liệu trả lời của TS và các CH trong đề thi phù hợp với mô hình Rasch ở mức độ nào?; (2) Phân bố độ khó của các CH tương thích như thế nào với phân bố năng lực của TS?; (3) Những phương án xử lí các CH sau khi được phân tích là gì?. Qua đó, bài báo bổ sung bằng chứng thực nghiệm và đề xuất các khuyến nghị nhằm hỗ trợ cải tiến đề thi đánh giá năng lực KĐV.

2. Phương pháp nghiên cứu

Chúng tôi sử dụng lí thuyết ứng đáp CH (Item Response Theory - IRT), cụ thể là mô hình Rasch nhị phân một tham số để phân tích đề thi đánh giá năng lực KĐV. Nghiên cứu lựa chọn mô hình Rasch nhị phân một tham số, với mục tiêu cung cấp một cấu trúc giản lược, có khả năng diễn giải rõ và phù hợp với mục đích hiệu chỉnh đề thi. Nghiên cứu sử dụng các chỉ số theo lí thuyết khảo thí cổ điển (tỉ lệ trả lời đúng, tương quan point-biserial, hiệu quả của phương án nhiều) như những bằng chứng hỗ trợ cho kết quả phân tích Rasch (Bond và Fox, 2001).

Trong phạm vi nghiên cứu, đối tượng được tiếp cận là 760 học viên thuộc các khóa 40-49 đã tham gia các lớp bồi dưỡng nghiệp vụ KĐV do Trung tâm KĐCL giáo dục - Đại học Quốc gia Hà Nội tổ chức tính đến tháng 12/2025. Theo Linacre (1994), quy mô mẫu cần thiết trong phân tích Rasch phụ thuộc vào mục tiêu và mức độ chính xác của ước lượng, mặc dù khả năng khái quát hóa kết quả và độ chính xác của các ước lượng ở hai đầu thang năng lực vẫn hạn chế. Bài kiểm tra được tiến hành vào tháng 10/2025, theo hình thức trực tuyến, có giám sát qua Zoom Meeting, TS làm bài trên Google Forms trong thời gian quy định. Các bài làm được đưa vào phân tích khi TS thuộc nhóm đối tượng nghiên cứu, hoàn thành bài thi trong thời gian quy định và trả lời đầy đủ 40 CH. Sau khi rà soát, nghiên cứu thu được 122 bài làm hợp lệ. Do mẫu được hình thành trên cơ sở tự nguyện và khả năng tiếp cận đối tượng, kết quả nghiên cứu chỉ được diễn giải trong phạm vi khảo sát, không đại diện đầy đủ cho toàn bộ học viên.

Đề thi gồm 40 CH, mỗi câu có 04 phương án trả lời A, B, C, D, trong đó chỉ có một phương án đúng. Căn cứ ma trận đề thi, các CH được xây dựng theo các mức độ đánh giá nhớ/biết, hiểu, vận dụng, phân tích, tổng hợp. Thời gian làm bài của TS là 60 phút. Sau khi được thu thập, dữ liệu được xử lí thô chứa các biến: TS (mã TS); C1 - C40 là biến ghi lại câu trả lời của mỗi TS ứng với câu 1 đến câu 40 trong đề thi, biến này nhận các giá trị A, B, C, D; điểm mà NH đạt được sẽ được xử lí bằng phần mềm ConQuest và JASP. Theo đó, ConQuest được sử dụng để ước lượng tham số theo mô hình Rasch bằng phương pháp hợp lí cực đại biên; các chỉ số thuộc lí thuyết khảo thí cổ điển để hỗ trợ việc chẩn đoán chất lượng CH gồm: độ khó (chỉ số này khác với ước lượng độ khó trên thang logit và không phải là p-value của một phép kiểm định giả thuyết); khả năng phân biệt được đánh giá thông qua tương quan point-biserial (bổ trợ theo khảo thí cổ điển, không phải tham số do mô hình Rasch ước lượng); các phương án trả lời được phân tích theo số lượng, tỉ lệ lựa chọn và tương quan với điểm toàn bài. Ngoài ra, JASP được sử dụng để tính các thống kê mô tả liên quan đến dữ liệu. Nghiên cứu chưa thực hiện kiểm tra riêng giả định đơn hướng và độc lập cục bộ; vì vậy, các kết quả Rasch được diễn giải theo hướng thăm dò trong phạm vi dữ liệu của lần thi này.

3. Kết quả nghiên cứu

3.1. Phân tích sự phù hợp của dữ liệu với mô hình

Kết quả xử lí dữ liệu cho thấy, trong 122 bài làm của TS, không có TS nào để bài thi trắng, không có TS nào lựa chọn duy nhất một phương án cho tất cả các câu trong đề trắc nghiệm; không có TS bỏ trống bất cứ CH nào và không có CH nào nhận giá trị trống từ bất kì TS nào tham dự.

Theo Bond và Fox (2001), giá trị kì vọng của chỉ số Infit và Outfit Mean Square (MNSQ) là 1,0; giá trị càng gần 1,0 càng cho thấy dữ liệu trả lời quan sát được phù hợp với kì vọng của mô hình Rasch. Trong nghiên cứu này, CH có MNSQ nằm trong khoảng từ 0,75 đến 1,30 và trị số chuẩn hóa tương ứng nằm trong khoảng từ -2,0 đến 2,0 được xem là có mức độ phù hợp có thể chấp nhận, giá trị Infit/Weighted fit dao động tùy từng CH. Kết quả ước lượng cho thấy 37/40 CH có chỉ số Outfit MNSQ nằm trong khoảng chấp nhận và trị số chuẩn hóa tương ứng tương đối gần giá trị kì vọng 0,0.

Về ước lượng độ khó (Estimate) của 40 CH cho thấy, 12/40 CH là các CH rất khó gồm C1, C4, C7, C15, C20, C24, C26, C28, C32, C33, C37 và C40 (giá trị Estimate > 0,5); 16/40 CH có độ khó vừa phải, là các câu C2, C6, C8, C10, C16, C18, C19, C23, C25, C27, C29, C30, C31, C34, C36 và C38 (giá trị Estimate nằm trong khoảng -0,5 đến 0,5); còn lại 12 câu là các CH dễ, chủ yếu đánh giá mức độ nhớ/biết, cụ thể là các câu C3, C5, C9, C11, C12, C13, C14, C17, C21, C22, C35 và C39 (giá trị Estimate < -0,5). Hệ số tin cậy Separation Reliability ở mức cao là 0,985, cho thấy thứ bậc độ khó của các CH được ước lượng với mức độ ổn định cao trong phạm vi mẫu nghiên cứu.

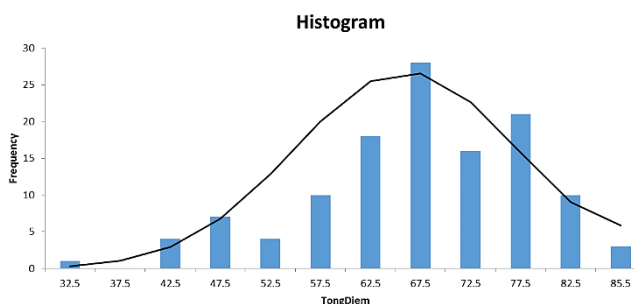
3.2. Phân tích thống kê mô tả kết quả bài thi

Kết quả thống kê mô tả với dữ liệu đã được mã hóa, thu được các chỉ số đặc trưng và biểu đồ phân bố điểm bài thi của 122 TS ở bảng 1.

Bảng 1. Thống kê mô tả kết quả đề thi đánh giá năng lực

N (Cỡ mẫu)	Giá trị	Số câu đúng	Điểm	Giá trị	Số câu đúng	Điểm
	Valid (Số giá trị)	122	122	Phương sai (Variance)	19,9	124,3
	Missing (Trống)	0	0	Độ lệch (Skewness)	-0,59	-0,59
	Trung bình (Mean)	26,41	66,03	Độ nhọn (Kurtosis)	0,17	0,17
	Trung vị (Median)	27	67,5	Giá trị thấp nhất (Minimum)	13	32,5
	Giá trị trội (Mode)	26	65	Giá trị cao nhất (Maximum)	35	87,5
	Độ lệch chuẩn (Std. Deviation)	4,46	11,15			

Bảng 1 cho thấy điểm trung bình bài làm của 122 TS là 66,03/100 với độ lệch chuẩn xấp xỉ 11,15, có 63,93% TS đạt trên điểm trung bình. Điểm cao nhất là 87,5/100 tương ứng trả lời đúng 35/40 câu và có 03 TS đạt được mức điểm này, chiếm 2,46%. Điểm thấp nhất là 32,5/100, tức TS chỉ đúng 13/40 câu, có 01 TS đạt mức điểm này, chiếm 0,82%. Với kết quả này, có thể nhận định rằng, đề thi có những CH dễ nhằm kiểm tra kiến thức cơ bản của TS, tạo điều kiện để TS không bị điểm liệt, đồng thời đề thi cũng không có nhiều CH khó, mang tính đánh đố TS. Từ biểu đồ phân bố điểm cho thấy, điểm bài thi của 122 TS chủ yếu tập trung trong khoảng từ 62,5 đến 77,5 (biểu đồ 1).



Biểu đồ 1. Biểu đồ phân bố điểm bài thi đánh giá năng lực

3.3. Phân tích chất lượng các câu hỏi

Trên cơ sở kết hợp chỉ số độ khó cổ điển, được xác định bằng tỉ lệ TS trả lời đúng, tương quan point-biserial và hiệu quả của các phương án nhiễu, đồng thời đối chiếu với kết quả phù hợp mô hình Rasch và nội dung chuyên môn, 40 CH được phân thành 04 nhóm CH: Nhóm I - Nhóm CH chất lượng tốt; Nhóm II - Nhóm CH chất lượng tốt, chỉ hiệu chỉnh nhỏ; Nhóm III - Nhóm CH quá dễ và quá khó và Nhóm IV - Nhóm CH cần chỉnh sửa nhiều hoặc cần nhắc loại bỏ. Các ngưỡng được sử dụng trong nghiên cứu được coi như một tiêu chí chẩn đoán, không phải căn cứ duy nhất để giữ lại hoặc loại bỏ CH.

a. Nhóm CH chất lượng tốt (nhóm I): Nhóm CH I gồm các CH đáp ứng được đồng thời các điều kiện: độ khó từ 0,40 đến 0,85, độ PB $\geq 0,30$ và chất lượng của các phương án nhiễu tốt.

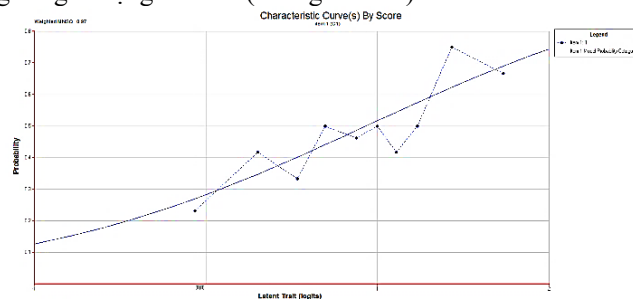
Bảng 2. Nhóm CH chất lượng tốt (nhóm I)

Câu	Đ/A	Độ khó	Độ PB	A		B		C		D	
				SL	%	SL	%	SL	%	SL	%
1	D	0,48	0,41	55	45,08	2	1,64	7	5,74	58	47,54
2	B	0,64	0,32	27	22,13	78	63,93	3	2,46	14	11,48
4	A	0,56	0,31	68	55,74	45	36,89	5	4,10	4	3,28
5	B	0,82	0,33	15	12,30	100	81,97	5	4,10	2	1,64
9	D	0,81	0,30	9	7,38	3	2,46	11	9,02	99	81,15

10	A	0,70	0,36	86	70,49	14	11,48	22	18,03	0	0,00
13	D	0,80	0,45	12	9,84	6	4,92	6	4,92	98	80,33
19	D	0,71	0,36	3	2,46	7	5,74	25	20,49	87	71,31
23	A	0,63	0,37	77	63,11	11	9,02	16	13,11	18	14,75
25	D	0,60	0,37	14	11,48	22	18,03	13	10,66	73	59,84
34	C	0,70	0,36	7	5,74	13	10,66	86	70,49	16	13,11
36	D	0,80	0,38	5	4,10	5	4,10	15	12,30	97	79,51
40	C	0,48	0,32	22	18,03	20	16,39	59	48,36	21	17,21

**Chú thích: Đ/A: Đáp án; SL: Số lượt lựa chọn; p: tỉ lệ TS trả lời đúng; PB: hệ số tương quan point-biserial giữa kết quả trả lời CH và điểm toàn bài.*

Nhóm CH chất lượng tốt (Nhóm I) gồm 13/40 CH, bao gồm các câu C1, C2, C4, C5, C9, C10, C13, C19, C23, C25, C34, C36 và C40. Cụ thể, các CH trong nhóm này có độ khó dao động từ 0,48 đến 0,82, độ PB dao động từ 0,30 đến 0,45, đạt mức tốt đến rất tốt. Phân tích cụ thể đối với CH C1, độ khó của câu này là 0,48 đạt mức phù hợp (có 47,54% TS trả lời đúng CH này), độ PB = 0,41 - đạt mức rất tốt. Xem xét từng phương án cho thấy, phương án D là đáp án đúng, có Pt Bis = 0,41 ($> 0,3$), trong khi các phương án A, B, C là phương án nhiễu và đều có Pt Bis âm, lần lượt là -0,35, -0,14 và -0,05. Chỉ số Pt Bis cho biết mức độ tương quan giữa điểm của từng phương án với điểm toàn bài của TS; theo đó, Pt Bis dương đối với phương án đúng, đặc biệt từ 0,3 trở lên, cho thấy CH có chất lượng tốt; ngược lại, Pt Bis âm đối với phương án nhiễu cho thấy các phương án này có xu hướng được lựa chọn nhiều hơn bởi nhóm TS có năng lực thấp. Như vậy, các giá trị Pt Bis trong CH C1 đều thỏa mãn yêu cầu của một CH trắc nghiệm có chất lượng. Đối với các phương án nhiễu, phương án A có tỉ lệ lựa chọn khá cao, chiếm 45,08%, gần tương đương với tỉ lệ lựa chọn đáp án đúng D, cho thấy đây là một phương án nhiễu có sức hấp dẫn, đặc biệt là với nhóm TS có năng lực thấp, thể hiện qua Pt Bis âm (-0,35); tuy nhiên cũng cần rà soát thêm nội dung của phương án này để bảo đảm không gây hiểu nhầm. Phương án B và C có tỉ lệ lựa chọn thấp, lần lượt là 1,64% và 5,74%, đồng thời Pt Bis âm nhưng có giá trị nhỏ. Điều này cho thấy hai phương án này chưa thực sự hấp dẫn TS và khả năng gây nhiễu còn hạn chế. Vì vậy, mặc dù C1 là một CH có chất lượng tốt nhưng vẫn cần điều chỉnh thêm các phương án nhiễu để có tính gây nhiễu cao hơn. Thông qua đường cong đặc trưng CH, ta thấy đường cong thực tế của C1 (đường nét đứt) bám khá sát đường cong kì vọng của CH (đường nét liền).



Biểu đồ 2. Đường cong đặc trưng của CH 1 (ICC Item 1)

b. *Nhóm CH chất lượng tốt, chỉ hiệu chỉnh nhỏ (nhóm II):* Nhóm CH II gồm các CH có độ khó phù hợp nhưng độ PB chỉ ở mức chấp nhận được, chỉ cần hiệu chỉnh nhỏ, đáp ứng các điều kiện: độ khó từ 0,40 đến 0,85 và độ PB từ 0,20 đến dưới 0,30.

Bảng 3. Nhóm CH chất lượng tốt, chỉ hiệu chỉnh nhỏ (nhóm II)

Câu	Đ/A	Độ khó	Độ PB	A		B		C		D	
				SL	%	SL	%	SL	%	SL	%
8	A	0,65	0,23	79	64,75	13	10,66	16	13,11	14	11,48
16	B	0,75	0,20	24	19,67	91	74,59	1	0,82	6	4,92
18	A	0,61	0,24	74	60,66	15	12,30	10	8,20	23	18,85
28	B	0,43	0,24	16	13,11	53	43,44	9	7,38	44	36,07
30	B	0,66	0,24	13	10,66	81	66,39	19	15,57	9	7,38
32	B	0,57	0,23	23	18,85	69	56,56	2	1,64	28	22,95
38	A	0,72	0,21	88	72,13	14	11,48	12	9,84	8	6,56
39	B	0,83	0,25	6	4,92	101	82,79	6	4,92	9	7,38

**Chú thích: Đ/A: Đáp án; SL: Số lượt lựa chọn; p: tỉ lệ TS trả lời đúng; PB: hệ số tương quan point-biserial giữa kết quả trả lời CH và điểm toàn bài.*

Nhóm CH chất lượng tốt, chỉ hiệu chỉnh nhỏ (nhóm II) gồm 08/40 CH, bao gồm các câu C8, C16, C18, C28, C30, C32, C38 và C39. Các CH trong nhóm này có độ khó dao động từ 0,40 đến 0,85 (đạt mức phù hợp), nhưng độ PB chỉ ở mức chấp nhận được (đạt từ 0,20 đến dưới 0,30), cho thấy các CH không quá khó hoặc quá dễ đến mức làm mất ý nghĩa đo lường; phần lớn TS có thể tiếp cận được nội dung CH, nhưng vẫn có sự phân hóa nhất định giữa các nhóm TS. Tuy các CH nhóm II vẫn có khả năng PB giữa TS có năng lực cao và năng lực thấp, nhưng mức độ PB chưa thật sự rõ nét. Ví dụ như C28 với độ khó là 0,43 và độ PB là 0,24 - đều ở mức có thể chấp nhận được, cho thấy CH này có khả năng PB nhất định giữa 02 nhóm TS, tuy nhiên mức độ PB chưa cao. Xem xét các phương án nhiều: phương án A và C lần lượt có 16 và 9 TS lựa chọn, phương án D có 44 TS lựa chọn. Việc có tới 36,07% TS lựa chọn phương án D cho thấy đây là phương án có khả năng gây nhiễu mạnh. Tuy nhiên, tỉ lệ lựa chọn D quá gần với đáp án đúng B cũng cần được xem xét lại - có thể nội dung phương án D chưa đủ khác biệt so với đáp án đúng, hoặc CH có khả năng gây hiểu nhầm, hoặc có nhiều hơn một cách diễn giải khiến TS phân vân giữa B và D.

c. *Nhóm CH quá dễ và quá khó (nhóm III): Nhóm CH III gồm các CH quá dễ - chỉ dùng để kiểm tra kiến thức nền tảng (độ khó > 0,85) và quá khó - có khả năng phân hóa nhóm năng lực cao (độ khó < 0,40), đều có độ PB $\geq 0,20$*

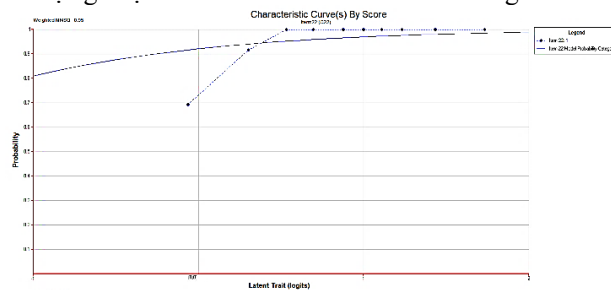
Bảng 4. Nhóm CH quá dễ và quá khó (nhóm III)

Câu	Đ/A	Độ khó	Độ PB	A		B		C		D	
				SL	%	SL	%	SL	%	SL	%
CH quá dễ											
3	A	0,91	0,28	111	90,98	2	1,64	0	0,00	9	7,38
11	C	0,89	0,39	7	5,74	3	2,46	108	88,52	4	3,28
14	D	0,93	0,28	2	1,64	3	2,46	3	2,46	114	93,44
17	A	0,86	0,26	105	86,07	5	4,10	5	4,10	7	5,74
21	A	0,93	0,41	113	92,62	5	4,10	1	0,82	3	2,46
22	A	0,96	0,44	117	95,90	3	2,46	1	0,82	1	0,82
35	D	0,88	0,33	6	4,92	2	1,64	7	5,74	107	87,70
CH quá khó											
20	B	0,25	0,31	5	4,10	31	25,41	86	70,49	0	0,00
33	C	0,12	0,24	21	17,21	76	62,30	15	12,30	10	8,20

**Chú thích: Đ/A: Đáp án; SL: Số lượt lựa chọn; p: tỉ lệ TS trả lời đúng; PB: hệ số tương quan point-biserial giữa kết quả trả lời CH và điểm toàn bài.*

Nhóm III gồm các CH quá dễ và quá khó gồm 09/40 CH, trong đó các câu C3, C11, C14, C17, C21, C22 và C35 thuộc nhóm quá dễ; và C20 và C33 thuộc nhóm quá khó. Mặc dù các CH trong nhóm này đều có độ PB từ 0,20 trở lên, tuy nhiên độ khó của các CH nằm ngoài khoảng phù hợp (> 0,85 và < 0,40).

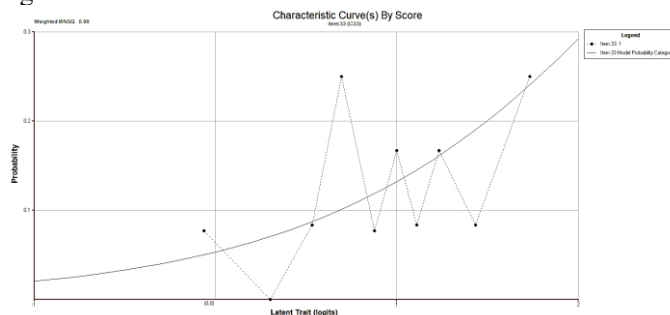
Các CH quá dễ có vai trò đánh giá những kiến thức nền tảng và các nội dung tối thiểu mà NH cần nắm được. Tuy nhiên, nếu tỉ lệ CH quá dễ trong đề thi quá cao, khả năng phân hóa năng lực của đề thi sẽ bị hạn chế. Trong nhóm này, một số CH như C21 và C22 tuy rất dễ nhưng vẫn có độ PB cao, cho thấy các CH này vẫn có giá trị đo lường và có thể được giữ lại nếu nội dung CH thuộc nhóm kiến thức cốt lõi. Phân tích C22 cho thấy, mặc dù CH vẫn có khả năng PB giữa nhóm TS có kết quả cao và kết quả thấp, nhưng các phương án nhiễu chưa đủ hấp dẫn. Thực tế, C22 có mục đích để đo lường kiến thức về xây dựng và phát triển chương trình đào tạo trong giáo dục đại học ở mức độ Hiểu, nên C22 có thể được giữ lại nếu để kiểm tra kiến thức nền tảng của KĐV.



Biểu đồ 3. Đường cong đặc trưng của CH 22 (ICC Item 22)

Ngược lại, các CH quá khó được sử dụng với mục đích phân hóa các TS có năng lực cao, đặc biệt nếu CH đo lường các kiến thức chuyên sâu hoặc đòi hỏi xử lý tình huống phức tạp. Tuy nhiên, nếu một CH có tỉ lệ trả lời đúng quá thấp thì cần phải xác định rõ nguyên nhân để chỉnh sửa cho phù hợp: có thể do nội dung thực sự khó, hoặc CH diễn đạt chưa rõ, hoặc các phương án nhiễu quá mạnh, hoặc đáp án đúng chưa đủ thuyết phục. C33 là CH có tỉ lệ trả

lời đúng rất thấp nhưng vẫn có tương quan điểm - câu ở mức chấp nhận được. Thực tế cho thấy, theo ma trận đề thi, C33 được thiết kế để phản ánh một số yêu cầu liên quan đến kiến thức về hệ thống KĐCL giáo dục và khả năng xử lý thông tin trong tình huống.



Biểu đồ 4. Đường cong đặc trưng của CH 33 (ICC Item 33)

d. Nhóm CH cần thẩm định, hiệu chỉnh và thử nghiệm lại (nhóm IV): Nhóm CH IV là các CH có độ PB < 0,20, bao gồm cả câu có độ PB âm.

Bảng 5. Nhóm CH cần chỉnh sửa nhiều hoặc cân nhắc loại bỏ (nhóm IV)

Câu	Đ/A	Độ khó	Độ PB	A		B		C		D	
				SL	%	SL	%	SL	%	SL	%
6	C	0,75	0,12	5	4,10	25	20,49	92	75,41	0	0,00
7	C	0,10	-0,11	11	9,02	58	47,54	12	9,84	41	33,61
12	D	0,84	0,16	6	4,92	5	4,10	8	6,56	103	84,43
15	B	0,53	0,18	9	7,38	65	53,28	26	21,31	22	18,03
24	B	0,54	0,17	8	6,56	66	54,10	10	8,20	38	31,15
26	C	0,35	0,18	38	31,15	20	16,39	43	35,25	21	17,21
27	A	0,69	0,19	84	68,85	13	10,66	22	18,03	3	2,46
29	B	0,67	0,12	11	9,02	82	67,21	28	22,95	1	0,82
31	B	0,74	0,15	2	1,64	90	73,77	1	0,82	29	23,77
37	B	0,51	0,15	17	13,93	62	50,82	23	18,85	20	16,39

**Chủ thích: Đ/A: Đáp án; SL: Số lượt lựa chọn; p: tỉ lệ TS trả lời đúng; PB: hệ số tương quan point-biserial giữa kết quả trả lời CH và điểm toàn bài.*

Nhóm CH cần chỉnh sửa nhiều hoặc cân nhắc loại bỏ (nhóm IV) gồm 10/40 CH, bao gồm: C6, C7, C12, C15, C24, C26, C27, C29, C31 và C37 (độ PB < 0,20). Mặc dù một số câu có độ khó nằm trong khoảng phù hợp nhưng lại có độ PB thấp, cho thấy các CH chưa phân hóa tốt giữa nhóm TS có năng lực cao và năng lực thấp. Theo tiếp cận IRT, point-biserial thấp cho thấy kết quả trả lời CH có tương quan yếu với điểm toàn bài; point-biserial âm là dấu hiệu bất thường cần ưu tiên kiểm tra đáp án, nội dung và cách diễn đạt. Việc xác định mức độ cung cấp thông tin của CH theo IRT cần dựa trên tham số và hàm thông tin của mô hình tương ứng. Câu C7 là trường hợp cần đặc biệt lưu ý trong nhóm IV. CH dùng để đo lường kiến thức về đảm bảo chất lượng và KĐCL giáo dục quốc tế, với nội dung hỏi về số lượng tổ chức KĐCL giáo dục của Việt Nam là thành viên chính thức của INQAAHE. Độ khó của câu chỉ đạt 0,10, tức chỉ có 9,84% TS trả lời đúng, trong khi 47,54% TS chọn phương án B và 33,61% chọn phương án D. Đặc biệt, độ PB của CH là -0,11 cho thấy, CH không PB được TS có năng lực cao và thấp, thậm chí có thể khiến nhóm TS có kết quả cao không lựa chọn đáp án đúng. Nguyên nhân có thể liên quan đến đáp án, cách diễn đạt, tính cập nhật của thông tin hoặc mức độ tương đồng giữa các phương án. Bên cạnh đó, C24 - với mục đích nhằm đo lường kỹ năng tổ chức, phối hợp và thực hiện hoạt động KĐCL giáo dục, đồng thời đánh giá nhận thức về đạo đức nghề nghiệp - là một trường hợp khác của nhóm IV vì có độ khó phù hợp (0,54) nhưng độ PB thấp (0,17). Mặc dù, CH có giá trị thực tiễn vì đặt TS vào tình huống cụ thể, nhưng vẫn cần rà soát lại cách diễn đạt các phương án nhằm làm rõ quy trình xử lý, thẩm quyền của Trường đoàn và trách nhiệm báo cáo với tổ chức KĐCL giáo dục. Do vậy, C24 chưa cần phải loại bỏ, nhưng cần thẩm định lại về nội dung, đáp án và các phương án nhiễu.

Kết quả phân loại cho thấy, 21/40 CH thuộc nhóm có chỉ số độ khó và point-biserial ở mức phù hợp hoặc chấp nhận được; 09 câu có tỉ lệ trả lời đúng nằm ngoài khoảng lựa chọn nhưng vẫn có khả năng sử dụng cho những mục đích cụ thể; và 10 câu có point-biserial thấp hoặc âm cần được ưu tiên thẩm định và thử nghiệm lại. Điều này cho thấy, đề thi đã có một bộ phận CH có khả năng phân hóa tương đối tốt, nhưng vẫn cần điều chỉnh tỉ trọng CH theo các mức độ khó, cải thiện phương án nhiễu và kiểm soát chất lượng trước khi sử dụng lâu dài.

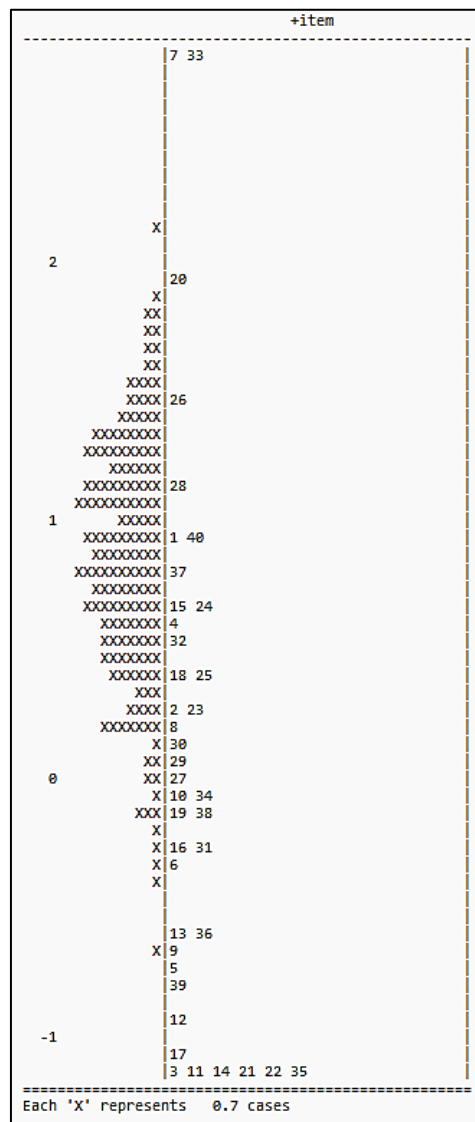
3.4. Mức độ tương thích giữa độ khó câu hỏi và năng lực thí sinh

Nghiên cứu xem xét sự tương thích giữa phân bố độ khó của CH và năng lực của TS thông qua Bản đồ phân bố năng lực TS và độ khó CH (hình 1). Kết quả cho thấy, năng lực của TS phân bố trong khoảng xấp xỉ từ -0,6 đến 2,4 logit, trong đó tập trung chủ yếu ở vùng từ 0,4 đến 1,3 logit. Mốc 0 logit được hiểu là mức độ khó trung bình của tập CH sau khi thang đo được chuẩn hóa, không phải điểm trung bình quan sát của mẫu. Phần lớn TS có năng lực nằm trên mốc này, cho thấy mức năng lực của nhóm tham gia cao hơn độ khó trung bình của đề thi. Các CH trong đề chủ yếu tập trung ở vùng độ khó trung bình và tương đối dễ, có khả năng đo lường tương đối phù hợp đối với TS ở vùng năng lực trung bình và trên trung bình; tuy nhiên, mức độ bao phủ chưa đồng đều trên toàn bộ dải năng lực. Cụ thể, các câu C3, C11, C14, C17, C21, C22 và C35 là những CH dễ, nằm ở vùng thấp hơn so với năng lực của phần lớn TS. Các CH này vẫn có thể dùng để kiểm tra kiến thức nền tảng, nhưng đóng góp chưa nhiều vào việc phân hóa năng lực. Ngược lại, các câu C7 và C33 nằm ở vùng độ khó cao hơn năng lực của đa số TS, nên khả năng cung cấp thông tin cho vùng năng lực trung bình còn hạn chế. Tuy nhiên, C33 có thể vẫn hữu ích trong việc phân hóa nhóm TS có năng lực cao nếu bảo đảm mức độ phù hợp với mô hình và chất lượng nội dung, trong khi C7 cần được ưu tiên thẩm định hơn do đồng thời có tỉ lệ trả lời đúng thấp và tương quan point-biserial âm. Ngoài ra, các CH còn lại chủ yếu đo lường được năng lực ở vùng trung bình đến trên trung bình của các TS; theo thứ tự này, các CH có mức độ khó tăng dần trên thang logit. Kết quả này cho thấy cần phải hiệu chỉnh đề thi, giảm tỉ trọng các CH quá dễ, duy trì số lượng các CH kiểm tra kiến thức nền tảng và bổ sung thêm các CH có độ khó cao, qua đó cải thiện khả năng bao phủ và độ chính xác đo lường trên toàn bộ thang năng lực.

4. Kết luận và bình luận

Bài báo đã sử dụng mô hình Rasch kết hợp với các chỉ số của lý thuyết khảo thí cổ điển để phân tích đề thi đánh giá năng lực gồm 40 CH trắc nghiệm của 122 TS sau bồi dưỡng nghiệp vụ KĐV. Kết quả phân tích cho thấy, phần lớn CH có chỉ số Outfit MNSQ nằm trong khoảng chấp nhận được; hệ số Item Separation Reliability đạt 0,985, phản ánh thứ bậc độ khó của các CH được ước lượng tương đối ổn định trong phạm vi mẫu nghiên cứu. Điểm bài thi dao động từ 32,5 đến 87,5 trên thang 100, với điểm trung bình là 66,03 và độ lệch chuẩn là 11,15; có 63,93% TS đạt điểm cao hơn điểm trung bình. Đề thi tương đối phù hợp với nhóm TS có năng lực trung bình và trên trung bình, với phần lớn CH tập trung ở vùng độ khó trung bình và tương đối dễ. Đề thi còn thiếu các CH có chất lượng tốt ở vùng năng lực cao, trong khi một số câu khó có dấu hiệu bất thường cần tiếp tục được thẩm định. Do đó, phân bố độ khó của đề thi chưa bao phủ đồng đều toàn bộ thang năng lực và cần được điều chỉnh để nâng cao độ chính xác khi đánh giá những TS có năng lực cao. Trên cơ sở kết hợp ước lượng Rasch, tỉ lệ trả lời đúng, tương quan point-biserial và các phương án nhiễu, 40 CH được phân thành bốn nhóm để đề xuất giữ lại, hiệu chỉnh, thử nghiệm lại hoặc thay thế. Đóng góp của nghiên cứu không chỉ nằm ở việc nhận diện các CH, mà còn ở việc minh họa cách kết hợp mô hình Rasch với phân tích CH theo lý thuyết khảo thí cổ điển để chẩn đoán chất lượng đề thi đánh giá năng lực KĐV. Tuy nhiên, kết luận của nghiên cứu chỉ giới hạn trong mẫu 122 TS và dữ liệu của một lần tổ chức thi, chưa được xem là bằng chứng đầy đủ về độ giá trị của toàn bộ công cụ.

Đánh giá năng lực KĐV, đặc biệt là với đối tượng là NH sau bồi dưỡng nghiệp vụ KĐV không chỉ là hoạt động có ý nghĩa quan trọng trong việc xác định mức độ đáp ứng yêu cầu của chuẩn đầu ra chương trình bồi dưỡng nghiệp



Hình 1. Bản đồ phân bố năng lực TS và độ khó CH

vụ, mà còn có tác động đến quá trình chuẩn bị của NH cho kì thi sát hạch cấp thẻ KĐV do Bộ GD-ĐT tổ chức. Vì vậy, đề thi cần được xây dựng, thẩm định và phân tích khoa học, phản ánh đúng các nhóm năng lực cần có của KĐV. Việc cải tiến đề thi theo đó cũng cần được thực hiện theo từng nhóm CH và dựa trên sự kết hợp giữa bằng chứng thống kê với đánh giá chuyên môn. Những CH có chỉ số phù hợp có thể tiếp tục được sử dụng và theo dõi sự ổn định qua những lần tổ chức thi tiếp theo, đồng thời rà soát lại đối với nhóm cần hiệu chỉnh nhỏ. Các CH có tương quan point-biserial thấp cần được kiểm tra lại về nội dung, đáp án và mức độ phù hợp với năng lực; đặc biệt, các câu có hệ số âm phải được ưu tiên thẩm định về tính chính xác, tính cập nhật và khả năng tồn tại nhiều cách hiểu; sau khi hiệu chỉnh, các CH này cần được thử nghiệm lại trước khi tiếp tục sử dụng chính thức. Bên cạnh đó, đối với các CH quá dễ, có thể duy trì một tỉ lệ hợp lí nếu nội dung phản ánh kiến thức cốt lõi hoặc các yêu cầu tối thiểu. Các CH khó cũng không nên bị loại bỏ, bởi những câu này có thể có giá trị trong việc đánh giá nhóm TS có năng lực cao. Kết quả phân tích có thể hỗ trợ điều chỉnh chương trình bồi dưỡng, xác định độ giá trị nội dung, độ giá trị bề mặt và độ giá trị cấu trúc của đề thi thông qua phương pháp chuyên gia, nhằm giảm thiểu các yếu tố có thể ảnh hưởng đến độ giá trị của đề thi.

Bài báo còn một số hạn chế do mẫu được lựa chọn theo phương pháp thuận tiện với bối cảnh dữ liệu mới được thu thập từ một đề thi trong một lần tổ chức; chưa kiểm tra riêng giả định đơn hướng và độc lập cục bộ, chưa phân tích sự khác biệt chức năng CH giữa các nhóm TS và chưa kết hợp đầy đủ kết quả thống kê với thẩm định chuyên gia hoặc phỏng vấn. Các nghiên cứu tiếp theo cần hiệu chỉnh và thử nghiệm lại những CH có dấu hiệu bất thường trên mẫu lớn, đa dạng hơn; kiểm tra các giả định của mô hình Rasch; thực hiện phân tích DIF theo đặc điểm đối tượng, lấy ý kiến phản biện từ các đối tượng không tham gia thiết kế CH; đồng thời so sánh kết quả qua nhiều lần tổ chức thi. Khi có cơ mẫu phù hợp, nghiên cứu có thể áp dụng thêm các kĩ thuật khác như cân bằng độ khó, neo đề, sử dụng các mô hình IRT hai hoặc ba tham số,... Kết quả phân tích đề thi theo IRT cần được sử dụng như một căn cứ thường xuyên để cập nhật ngân hàng CH, cải tiến chương trình bồi dưỡng và nâng cao chất lượng KĐV.

Tuyên bố về vai trò của các tác giả: Vũ Minh Hiếu: Viết bản thảo ban đầu, Thiết kế phương pháp nghiên cứu, Phân tích dữ liệu, Thực hiện nghiên cứu và thu thập dữ liệu, Trực quan hoá dữ liệu, Chỉnh sửa và phản biện bản thảo, Chịu trách nhiệm liên hệ và nội dung; Tạ Thị Thu Hiền: Xây dựng ý tưởng nghiên cứu, Trực quan hoá dữ liệu, Giám sát, chỉ đạo quá trình nghiên cứu, Chỉnh sửa và phản biện bản thảo.

Tuyên bố về GenAI và Quyền tác giả: Trong quá trình chuẩn bị bản thảo này, các tác giả đã sử dụng ChatGPT (OpenAI, Hoa Kỳ) nhằm hỗ trợ chỉnh sửa ngôn ngữ, cải thiện tính rõ ràng và mạch lạc của bản thảo.

Tuyên bố về xung đột lợi ích: Các tác giả tuyên bố không có xung đột lợi ích.

Tài liệu tham khảo

- Amua-Sekyi, E. T. (2016). Assessment, Student Learning and Classroom Practice: A Review. *Journal of Education and Practice*, 7(21), 1-6.
- Bộ GD-ĐT (2022). *Thông tư số 14/2022/TT- BGDĐT ngày 10/10/2022 quy định về Kiểm định viên giáo dục đại học và cao đẳng sư phạm*.
- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Lawrence Erlbaum Associates.
- Butler, A. C. (2018). Multiple-choice testing in education: Are the best practices for assessment also good for learning?. *Journal of Applied Research in Memory and Cognition*, 7(3), 323-331. <https://doi.org/10.1016/j.jarmac.2018.07.002>
- Clauser, B. E., Margolis, M. J., & Case, S. M. (2006). Testing for licensure and certification in the professions. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 701-731). Praeger Publishers.
- Dell, K. A., & Wantuch, G. A. (2017). How-to-guide for writing multiple choice questions for the pharmacy instructor. *Currents in Pharmacy Teaching and Learning*, 9(1), 137-144. <https://doi.org/10.1016/j.cptl.2016.08.036>
- Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, 10, 133-143. <https://doi.org/10.1007/s10459-004-4019-5>
- Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education*, 21(4), 357-364. <https://doi.org/10.1016/j.tate.2005.01.008>
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87(6), 1082-1116. <https://doi.org/10.3102/0034654317726529>
- Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Routledge.