ỨNG DỤNG MÔ HÌNH LÝ THUYẾT TRẮC NGHIỆM CỔ ĐIỂN (CTT) VÀ LÝ THUYẾT ỨNG ĐÁP CÂU HỎI (IRT) TRONG PHÂN TÍCH ĐỀ THI TRẮC NGHIỆM TẠI ĐẠI HỌC Y DƯỢC TPHCM

Tan Nguyen, Doan Thi Thu Hoa, Tran Quang Nam, Nguyen Thi Mai Lan, Nguyen Hoang Tam, Ly Huu Tuan, Tran Thanh Hung, Pham Thi Minh Hong, Vuong Thi Ngoc Lan

Đại học Y Dược Thành phố Hồ Chí Minh

1. MỞ ĐẦU

Phân tích đề thi là quá trình sử dụng những phương pháp thống kê để xác định chất lượng của đề thi thông qua xem xét từng câu hỏi trắc nghiệm riêng lẻ và đánh giá mức độ phù hợp của chúng, từ đó giúp xác định xem liệu có nên loại bỏ, giữ lại hoặc sửa đổi câu hỏi hay không. Phân tích đề thi là một quá trình hậu kiểm, được tiến hành sau khi đề thi đã được cho sinh viên thi, dùng để đảm bảo tất cả các câu hỏi thi đều công bằng. Tuy nhiên để đảm bảo đề thi chính xác và có khả năng đo lường được đúng năng lực của người học, cần thiết phải có công cụ để đánh giá chất lượng đề thi, làm cơ sở để điều chỉnh và cải tiến chất lượng, giúp giảng viên và các nhà quản lý đổi mới về phương pháp giảng dạy, phương pháp quản lý để hỗ trợ người học đạt được các mục tiêu trong học tập. Những năm gần đây, tại Đại học Y Dược TPHCM, bên cạnh việc đổi mới chương trình và phương pháp giảng dạy, hoạt động đổi mới phương pháp lượng giá cũng được quan tâm, chú trọng bằng việc thay đổi quan điểm tiếp cận về lý luận lượng giá, thay đổi phương pháp lượng giá phù hợp với yêu cầu của hoạt động giảng dạy, chuẩn bị thành lập ngân hàng câu hỏi thi trắc nghiệm cho các kỳ thi quan trọng. IRT dần được phổ biên trong thời gian gần đây vì đặc tính kết quả phân tích câu hỏi thi không phụ thuộc vào năng lực của từng nhóm học viên khác nhau, có thể hỗ trợ cho việc thiết kế câu hỏi thi và xây dựng ngân hàng câu hỏi thi. Vì vậy chúng tôi tiến hành phân tích 46 câu hỏi thi trắc nghiệm của học phần Nhi trong đề thi tốt nghiệp 2022 – 2023 theo IRT để đánh giá độ khó câu hỏi cũng như năng lực sinh viên và so sánh kết quả của IRT với CTT hiện tại đang sử dụng tại Đại học Y Dược TPHCM.

2. Các phương pháp phân tích câu hỏi thi, đề thi trắc nghiệm[1-3]

Có 2 thuyết chính được sử dụng để phân tích câu hỏi, đề thi trắc nghiệm là: lý thuyết trắc nghiệm cổ điển (Classical Test Theory - CTT) và lý thuyết ứng đáp câu hỏi (Item Response Theory - IRT).

Lý thuyết trắc nghiệm cổ điển (Classical Test Theory - CTT)[4]

Lý thuyết trắc nghiệm cổ điển (CTT) (Novick, 1966; Lord & Novick, 1968) là một cách tiếp cận định lượng truyền thống để kiểm tra độ tin cậy và tính hợp lệ của một thang đo dựa trên các hạng mục của nó (trắc nghiệm là một thang đo). Các chỉ số của CTT sử dụng để đánh giá đề thi, câu hỏi thi bao gồm:

Độ khó (DIFF I) của một câu hỏi trắc nghiệm là tỉ lệ phần trăm sinh viên trả lời đúng câu hỏi đó trong tổng số sinh viên làm bài thi. Giá trị của độ khó nằm trong khoảng [0 – 1], < 0,3 là khó, 0,3 – 0,7 là chấp nhận được, > 0,7 là dễ. Độ khó càng cao thì câu hỏi thi càng dễ.

Độ phân cách (DI) của một câu hỏi trắc nghiệm nói lên khả năng phân biệt sinh viên giỏi và không giỏi khi trả lời câu hỏi đó. Độ phân cách của câu hỏi liên quan đến độ khó của câu hỏi. Nếu một câu hỏi quá khó hay quá dễ thì phản ứng của sinh viên có năng lực khác nhau là giống nhau: hoặc sai hết hoặc đúng hết, do đó không phân biệt được năng lực của sinh viên. Vì vậy, một câu hỏi có khả năng phân cách tốt cần có độ khó ở mức trung bình và một đề thi trắc nghiệm tốt cần có nhiều câu hỏi có mức độ trung bình. Khi đó, điểm số của sinh viên có phổ trải rộng.

Lý thuyết ứng đáp câu hỏi (Item Response Theory - IRT)[5,6]

Lý thuyết ứng đáp câu hỏi (IRT), còn gọi là lý thuyết trắc nghiệm hiện đại, được ra đời vào thế kỷ XX và phát triển mạnh mẽ cho đến nay. IRT là một phương pháp tiếp cận xác suất và thống kê để khắc phục một số hạn chế của phương pháp lý thuyết trắc nghiệm cổ điển (CTT), đó là không tách biệt được các đặc trưng của sinh viên độc lập (**năng lực**) với đặc trưng của đề trắc nghiệm; CTT coi sai số tiêu chuẩn của phép đo năng lực giữa các sinh viên là như nhau, quan tâm mức độ đáp ứng của sinh viên với đề thi mà không chú trọng mức độ đáp ứng của sinh viên với các câu hỏi riêng biệt. IRT là mô hình hóa mối quan hệ giữa biến không thể quan sát là năng lực của sinh viên (được ký hiệu là θ) và xác suất mà tại đó sinh viên trả lời đúng một câu hỏi. Hiểu đơn giản hơn, IRT sử dụng mô hình toán học để dự đoán xác suất trả lời đúng một câu hỏi, dựa trên chỉ số về năng lực của người trả lời và độ khó của câu hỏi. Câu hỏi trắc nghiệm được đặc trưng bởi 3 tham số là: **độ khó** (response probability: b), **độ phân cách** (discriminator: a) và **độ dự đoán** (guessing: c). Tương ứng các tham số đó, các mô hình ứng đáp được đưa ra bao gồm: Mô hình ứng đáp một tham số (**1PL - mô hình Rasch**): chỉ sử dụng một tham số là độ khó của câu hỏi; mô hình hai tham số (**2PL**): sử dụng cả 2 biến là độ khó và độ phân cách của câu hỏi; mô hình ba tham số (**3PL**): sử dụng cả 3 tham số là độ khó, độ phân cách và độ dự đoán.

So với CTT, IRT có những ưu điểm nổi bật là các mô hình tính toán mang lại là các tham số đặc trưng của câu hỏi (độ khó (b), độ phân cách (a), độ dự đoán (c)) không phụ thuộc vào mẫu thử để định cỡ CH và năng lực (θ) đo được của TS không phụ thuộc vào ĐTN cụ thể được lấy từ ngân hàng câu hỏi đã được định chuẩn. Như vậy theo IRT, mỗi câu hỏi có các thuộc tính đặc trưng của nó, mỗi TS ở một trình độ nào đó có một năng lực xác định, các thuộc tính đặc trưng này không phụ thuộc vào phép đo, hay nói cách khác chúng là bất biến (invariance). Việc ứng dụng IRT sẽ góp phần gia tăng độ chính xác của phép đo lường trong giáo dục. Từ đó, chúng ta có thể đề xuất quy trình xây dựng NHCH, phân tích kết quả trả lời các câu hỏi trắc nghiệm để xác định chất lượng câu hỏi, chủ động trong việc thiết kế, xây dựng đề kiểm tra trắc nghiệm đáp ứng tốt các mục đích đã đề ra.

3. PHƯƠNG PHÁP NGHIÊN CỨU

Chúng tôi tiến hành nghiên cứu cắt ngang tại khoa Y, Đại Học Y Dược Thành Phố Hồ Chí Minh. Kết quả thi của 46 câu hỏi trong học phần Nhi của đề thi tốt nghiệp chương trình bác sĩ y khoa năm 2022 của 364 sinh viên được thu thập và xóa định danh. Phần mềm phân tích trắc nghiệm YDS và phần mềm R được sử dụng để phân tích độ khó của 46 câu hỏi thi theo mô hình CTT và IRT. Trước khi sử dụng IRT để phân tích câu hỏi thi, chúng tôi kiểm tra tính phụ thuộc lẫn nhau của các câu hỏi thi dựa vào chỉ số Yen's Q3, trong đó Q3 > 0.2 có thể là gợi ý các câu có phụ thuộc vào nhau. Để kiểm tra mức độ phù hợp của mô hình Rasch để ước đoán các đặc tính của đề thi, các chỉ số M2 được sử dụng. Item infit và Item outfit được dùng để xác định mức độ phù hợp của mô hình trong việc ước đoán đặc điểm của từng câu hỏi thi, với giá trị trong khoảng 0.5 đến 1.5 thể hiện mức độ phù hợp của mô hình. Nếu có câu hỏi nào có trung bình bình phương nằm ngoài khoảng này, điều đó có nghĩa là câu hỏi đó có rất ít hoặc không có giá trị đo lường (ví dụ cho đề sai, câu hỏi cũ học viên đã biết trước…) cần cân nhắc loại bỏ ra khi phân tích đề thi. Năng lực ước đoán của từng sinh viên được tính theo % trả lời đúng 46 câu hỏi của sinh viên theo CTT và tính theo giá trị theta trong thang năng lực ước đoán theo 46 câu hỏi thi theo IRT. Độ khó của câu hỏi thi và năng lực ước đoán của từng sinh viên được tính từ hai phương pháp và được chuẩn hóa về Z-score trước khi tiến hành so sánh mức độ tương quan của hai phương pháp CTT và IRT. Chúng tôi sử dụng kiểm định t-test bắt cặp để thực hiện phép kiểm về sự khác biệt (nếu có) từ hai phương pháp trên. Hệ số tương quan Pearson's correlation được dùng để đánh giá mức độ tương quan trong việc ước đoán năng lực của sinh viên dựa trên hai phương pháp CTT và IRT.

4. KẾT QUẢ VÀ BÀN LUẬN

Chúng tôi tiến hành kiểm tra sự phụ thuộc lẫn nhau của 46 câu hỏi thi dựa trên chỉ số Q3, với điểm chặn là 0.2. Câu 2 và câu 3 trong 46 câu hỏi thi có chỉ số tương quan tồn dư là 0.241. Tuy nhiên, khi phân tích hai câu hỏi thi, chúng tôi xác định 2 câu hỏi thi số 2 và số 3 không có liên quan hay phụ thuộc và nhau. Từ đó, chúng tôi quyết định giữ nguyên 46 câu hỏi thi để tiếp tục sử dụng IRT với mô hình Rasch để phân tích các đặc tính của câu hỏi thi.

Chúng tôi kiểm tra sự phù hợp khi sử dụng mô hình Rasch để đánh giá các đặc tính của 46 câu hỏi thi và đánh giá năng lực ước đoán của sinh viên. Các chỉ số thống kê của M2 thể hiện sự phù hợp của mô hình Rasch khi đánh giá bộ đề thi 46 câu trong học phần Nhi của đề thi tốt nghiệp ($p<0.001$, RMSEA = 0.03). Kết quả cho thấy các câu hỏi trong bộ đề phù hợp với mô hình tiên đoán Rasch đang sử dụng và có hiệu quả trong đo lường trung bình bình phương của các câu hỏi (Item infit và outfit nằm trong khoảng 0.5-1.5).

4.1. Các đặc tính của câu hỏi thi theo phương pháp CTT và IRT

Bảng 1 trình bày Độ khó của 46 câu hỏi thi theo hai phương pháp và độ khó sau khi được chuẩn hóa. Theo mô hình CTT, câu 1 là câu khó nhất trong 46 câu hỏi thi trong học phần Nhi, với 6% sinh viên trả lời đúng được câu hỏi số 1. Trong khi đó, đối với câu 9, câu 10, câu 29, câu 31, câu 32, câu 36 và câu 41 có hơn 90% trong tổng số sinh viên có thể trả lời đúng các câu trên. Tổng số câu dễ chiếm 30% trong các câu hỏi của học phần Nhi và tổng số câu khó chiếm 17%.

Bảng 1: Độ khó của 46 câu hỏi thi học phần Nhi trong đề thi tốt nghiệp bác sĩ y khoa năm 2022 (file đính kèm)

Khi sử dụng mô hình Rasch để đánh giá mức độ khó, các câu hỏi trong học phần Nhi được dùng để đánh giá năng lực của học viên từ mức theta thấp nhất là -3.78 đến mức theta cao nhất là 3.32. Phần lớn các câu hỏi tập trung trong mức theta từ -1 đến 0.5 (Biểu đồ 1).

Biểu đồ 1: Phân bố câu hỏi thi và số lượng sinh viên theo khung năng lực ước đoán theta (file đính kèm)

Theo biểu đồ trên, chúng ta thấy phân bố năng lực của SV từ -1 đến +1, và phần lớn câu hỏi tập trung đánh giá ở Theta -1 đến 0.5. Tuy nhiên, số câu hỏi dễ hơn (đánh giá Theta < -1) cũng chiếm tỉ trọng khá nhiều trong đề thi. Với quy mô một đề thi tốt nghiệp dành cho SV Y khoa năm thứ 6, ta có thể kết luận đề thi này là dễ so với năng lực trung bình của học viên.

Biểu đồ 2: Các câu hỏi thi dễ trong 46 câu hỏi thi (file đính kèm)

Biểu đồ 2 là các đường cong ICC của các câu hỏi trong nhóm dễ của đề thi. Các câu hỏi ở biểu đồ 2 cho thấy một thí sinh có mức năng lực trung bình (Theta= 0) có khả năng trả lời đúng đến trên 90%, và thí sinh có năng lực thấp nhất (Theta= -1) cũng có khả năng trả lời đúng trên 80%. Nhóm ra đề thi cần xem xét lại các câu hỏi này để chỉnh sửa hoặc loại bỏ khỏi ngân hàng đề thi tốt nghiệp do khả năng không giúp đo lường được năng lực sinh viên năm thứ 6. Tuy nhiên, đối với nhóm sinh viên có năng lực thấp hơn (ví dụ năm 4), các câu hỏi có thể có giá trị đo lường.

Biểu đồ 3: Các câu hỏi thi khó trong 46 câu hỏi thi (file đính kèm)

Biểu đồ 3 là các đường cong ICC của các câu hỏi trong nhóm khó của đề thi. Các câu hỏi ở biểu đồ 3 cho thấy một thí sinh có mức năng lực trung bình (Theta= 0) chỉ có khả năng trả lời đúng tối đa khoảng 25%, và thí sinh có năng lực cao nhất (Theta= 1) cũng chỉ có khả năng trả lời đúng từ 10-50%. Xét trên toàn bộ đề thi, số lượng câu hỏi khó là vừa phải (8/46 câu, 17%). Các câu hỏi thuộc nhóm trên có thể được sử dụng cho các đối tượng học viên có năng lực cao hơn (ví dụ bác sĩ nội trú).

Phép kiểm bắt cặp t-test trên độ khó đã được chuẩn hóa theo Z-score từ mô hình CTT và mô hình Rasch cho thấy có không có sự khác biệt có ý nghĩa thống kê giữa hai phương pháp, với mức độ tương quan cao giữa độ khó câu hỏi thi dựa trên hai phương pháp ($r = -0.983$). Cách tiếp cận xác định độ khó của CTT dựa vào % học viên trả lời được đúng câu hỏi thi, trong khi độ khó của câu hỏi thi khi dùng IRT thì dựa vào thang năng lực ước đoán mà học viên có 50% trả lời được đúng câu hỏi thi. Từ đó có thể lý giải vì sao độ khó khi xác định bằng hai phương pháp có mối tương quan nghịch với nhau. Kết quả này cũng tương tự với kết quả trong nghiên cứu của các tác giả Malaysia: sự khác biệt trong độ khó của câu hỏi tính theo CTT và IRT không có ý nghĩa thống kê ($p>0.05$)[7].

4.2. Năng lực ước đoán của sinh viên theo mô hình CTT và Rasch

Khi dùng mô hình CTT để ước đoán năng lực của sinh viên, chúng tôi dựa vào % số câu trả lời đúng của từng bạn sinh viên trong tổng số 46 câu hỏi thi trong học phần Nhi. Sinh viên có năng lực thấp nhất trả lời đúng được khoảng 28% tổng số câu, trong khi sinh viên có năng lực cao nhất trả lời đúng được khoảng 85% tổng số câu. Khi dùng mô hình Rasch, năng lực ước đoán của sinh viên được ước tính dựa vào hàm fscore của mô hình fitRasch. Năng lực theta thấp nhất trong 364 bạn sinh viên là -1.12 và năng lực theta cao nhất là 1.06. Trong khi các câu hỏi thi được trải dài trong khoảng theta từ -3.78 đến 3.22 (Biểu đồ 1). Hệ số tương quan Pearson's correlation thể hiện mức độ tương quan rất mạnh khi dùng mô hình CTT và mô hình Rasch để ước đoán năng lực của 364 bạn sinh viên tham gia trong kỳ thi tốt nghiệp bác sĩ y khoa năm 2022 ($r = 0.999$, $p < 0.001$).

Về mặt lý thuyết, khi phân tích trên một đoàn hệ thí sinh, kết quả số câu đúng (Raw score) hay kết quả thi theo CTT và kết quả thi theo IRT mô hình 1-PL gần như tương đương với nhau. Đó là vì ở mô hình 1-PL, chúng ta sử dụng một giả định rằng tất cả các câu hỏi đều có độ phân biệt (discrimination) như nhau. Nếu sử dụng mô hình 2-PL, tham số độ phân cách sẽ được tính đến. Khi đó, hai học viên cùng làm đúng 50/100 câu hỏi (raw score) có thể sẽ có kết quả đánh giá năng lực khác nhau (do 50 câu hỏi làm đúng của học viên thứ nhất khác với 50 câu hỏi làm đúng của học viên thứ hai).

KẾT LUẬN

Mô hình lý thuyết trắc nghiệm cổ điển (CTT) và lý thuyết đáp ứng câu hỏi (IRT) đều có giá trị trong việc phân tích câu hỏi và đề thi. Kết quả nghiên cứu cho thấy nếu tính ở cùng một cỡ mẫu học viên, những thông tin về độ khó của câu hỏi thi và năng lực ước đoán của học viên là tương đồng ở 2 mô hình. Tuy nhiên, mô hình IRT sẽ ưu thế hơn trong việc phân tích độ khó của câu hỏi thi trên khung năng lực ước đoán của thí sinh, tạo điều kiện cho việc xây dựng ngân hàng câu hỏi và xây dựng bộ đề thi đáp ứng trên máy tính (CAT). Việc sử dụng phần mềm R để chạy mô hình Rasch trong phân tích các câu hỏi và đề thi là hoàn toàn khả thi và có thể ứng dụng để phân tích các điểm tối ưu của IRT trong những nghiên cứu tiếp theo.

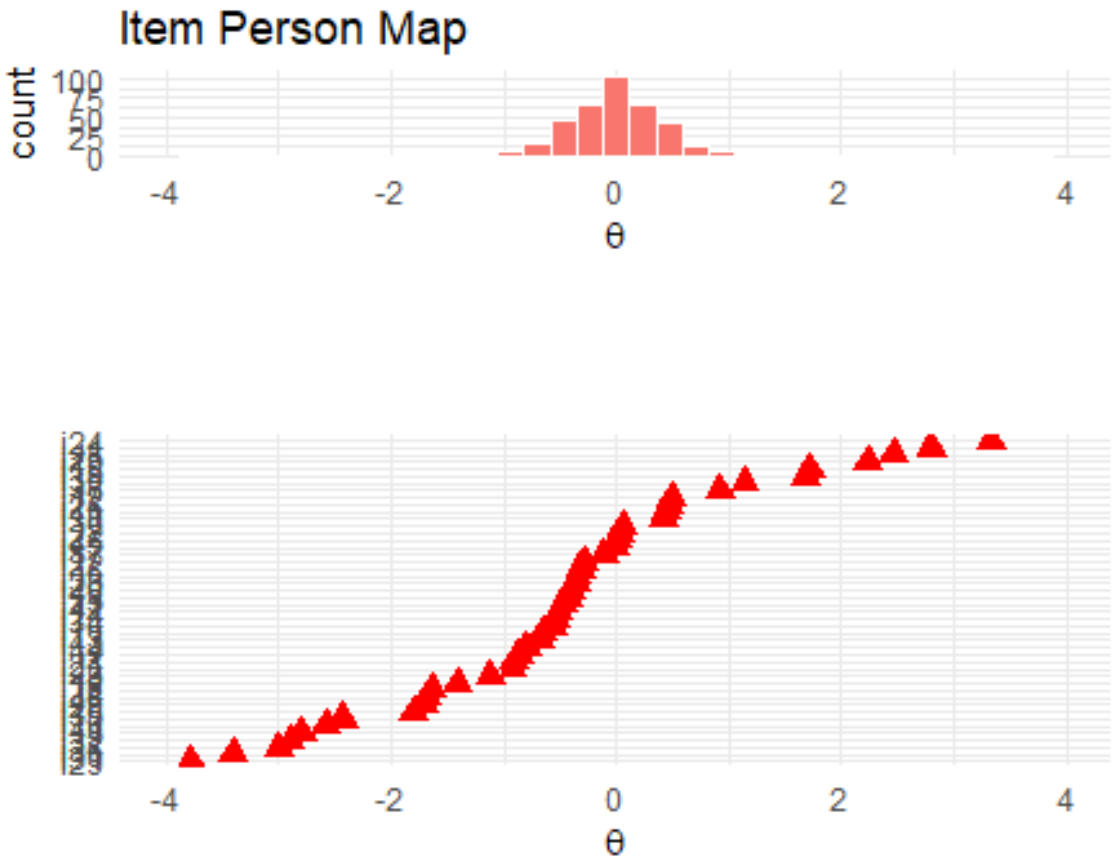
TÀI LIỆU THAM KHẢO

1. Thiệp LQ. *Đo lường và đánh giá hoạt động học tập trong nhà trường*. NXB Đại học sư phạm, Việt Nam; 2012.
2. Frank B. Baker, Kim. S-H. *Item Response Theory: Parameter Estimation Techniques, Second Edition*. 2nd ed. Boca Raton; 2004.
3. Baker FB. *The basics of item response theory*. College Park, MD: University of Maryland, ERIC Clearinghouse on Assessment and Evaluation; 2001.
4. Brennan LR. *Educational Measurement*. 4th ed. American Council on Education 2006.
5. Hambleton RK, Swaminathan H. *Item response theory: Principles and applications*. Springer Science & Business Medias; 2013.
6. Rasch G. *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Studies in mathematical psychology: I Probabilistic models for some intelligence and attainment tests. Nielsen & Lydiche; 1960:xiii, 184-xiii, 184.
7. abdul latif A, Yusof I, Amin N, Libunao W, Yusri S. Multiple-choice items analysis using classical test theory and rasch measurement model. *Man in India*. 01/01 2016;96:173-181.

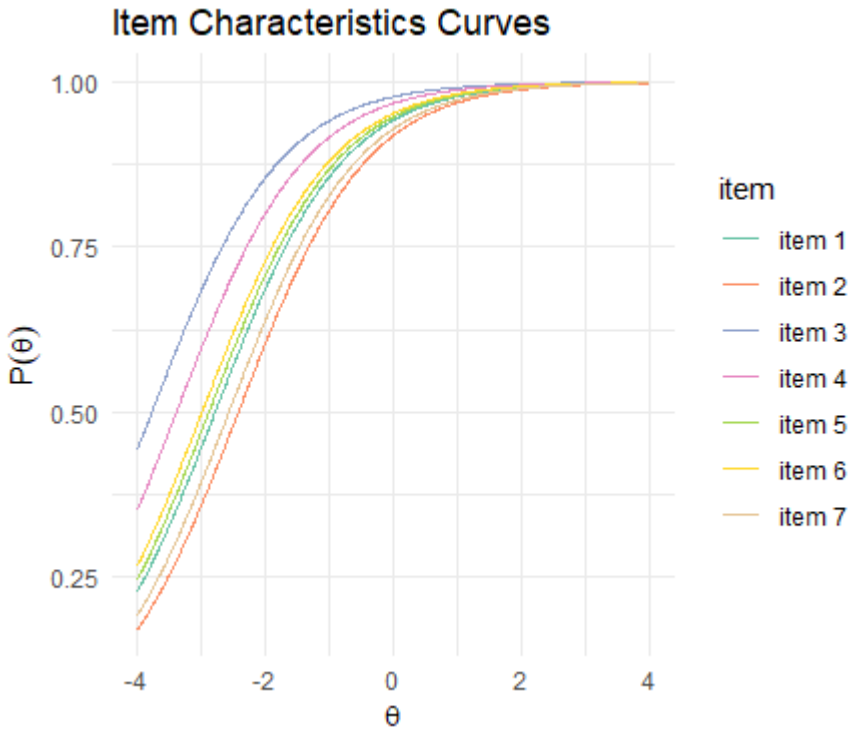
Bảng 1: *Độ khó của 46 câu hỏi thi học phần Nhi trong đề thi tốt nghiệp bác sĩ y khoa năm 2022*

STT câu hỏi	Độ khó theo CTT	Độ khó theo CTT đã chuẩn hóa (Z-score)	Độ khó theo Rasch	Độ khó theo IRT đã chuẩn hóa (Z-score)
1	**0.06**	-2.17	**2.79**	3.08
2	0.70	-1.53	-0.87	-0.58
3	0.60	-1.63	-0.42	-0.13
4	0.83	-1.40	-1.64	-1.35
5	0.49	-1.74	0.06	0.35
6	0.57	-1.66	-0.32	-0.03
7	**0.29**	-1.94	0.92	1.21
8	0.74	-1.49	-1.12	-0.83
9	**0.94**	-1.29	**-2.8**	-2.51
10	**0.91**	-1.32	**-2.43**	-2.14
11	0.40	-1.83	0.44	0.73
12	0.38	-1.85	0.50	0.79
13	0.65	-1.58	-0.66	-0.37
14	0.62	-1.61	-0.51	-0.22
15	**0.16**	-2.07	1.72	2.01
16	0.63	-1.60	-0.57	-0.28
17	0.53	-1.70	-0.12	0.17
18	0.79	-1.44	-1.41	-1.12
19	**0.10**	-2.13	**2.24**	2.53
20	0.59	-1.64	-0.36	-0.07
21	0.70	-1.53	-0.91	-0.62
22	0.57	-1.66	-0.29	0.00
23	0.71	-1.52	-0.92	-0.63
24	**0.04**	-2.19	**3.32**	3.61
25	**0.09**	-2.14	**2.47**	2.76
26	0.39	-1.84	0.47	0.76
27	0.49	-1.74	0.04	0.33
28	0.84	-1.39	-1.72	-1.43
29	**0.98**	-1.25	**-3.78**	-3.49
30	0.40	-1.83	0.42	0.71
31	**0.96**	-1.27	**-3.40**	-3.11
32	**0.94**	-1.29	**-2.89**	-2.60
33	0.58	-1.65	-0.35	-0.06
34	0.63	-1.60	-0.54	-0.25

STT câu hỏi	Độ khó theo CTT	Độ khó theo CTT đã chuẩn hóa (Z-score)	Độ khó theo Rasch	Độ khó theo IRT đã chuẩn hóa (Z-score)
35	0.85	-1.38	-1.8	-1.51
36	**0.95**	-1.28	**-3.00**	-2.71
37	0.50	-1.73	-0.01	0.28
38	**0.25**	-1.98	1.14	1.43
39	**0.17**	-2.06	1.68	1.97
40	0.60	-1.63	-0.41	-0.12
41	**0.92**	-1.31	**-2.58**	-2.29
42	0.61	-1.62	-0.48	-0.19
43	0.66	-1.67	-0.68	-0.39
44	0.68	-1.55	-0.80	-0.51
45	0.50	-1.73	0.00	0.29
46	0.83	-1.40	-1.68	-1.38

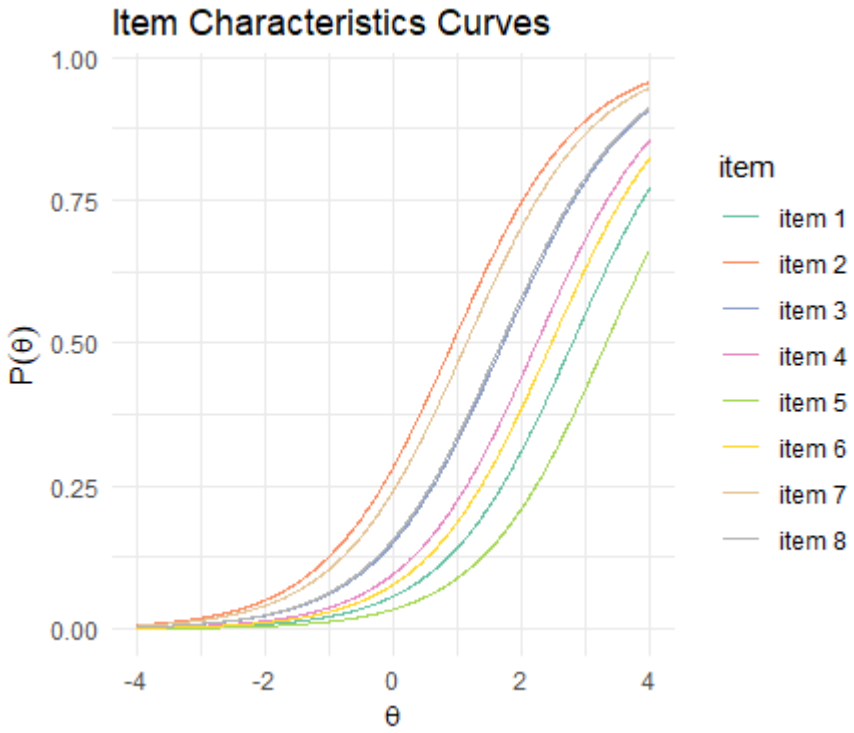


Biểu đồ 1: *Phân bố câu hỏi thi và số lượng sinh viên theo khung năng lực ước đoán theta*



Biểu đồ 2: *Các câu hỏi thi dễ trong 46 câu hỏi thi*

Chú thích: Item 1: câu 9, Item 2: câu 10, Item 3: câu 29, Item 4: câu 31, Item 5: câu 32, Item 6: câu 36, Item 7: câu 41



Biểu đồ 3: *Các câu hỏi thi khó trong 46 câu hỏi thi*

Chú thích: Item 1: câu 1, Item 2: câu 7, Item 3: câu 15, Item 4: câu 19, Item 5: câu 24, Item 6: câu 25, Item 7: câu 38, Item 8: câu 39