

DOI: 10.58490/ctump.2025i92.4222

ĐÁNH GIÁ ĐỘ KHÓ VÀ ĐỘ PHÂN BIỆT CỦA BỘ ĐỀ THI TỐT NGHIỆP SINH VIÊN Y6 – TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH – TIẾP CẬN TỪ CTT ĐẾN MIRT

Vũ Anh Duy, Lê Minh Nguyệt, Nguyễn Văn Trí, Trần Ngọc Sinh, Trịnh Hữu Tùng,
Lê Quang Thanh, Hoàng Văn Quang, Đỗ Trọng Ánh, Lê Thị Kim Nhung*

Trường Đại học Nguyễn Tất Thành

**Email: vaduy@ntt.edu.vn*

Ngày nhận bài: 14/9/2025

Ngày phản biện: 17/10/2025

Ngày duyệt đăng: 25/10/2025

TÓM TẮT

Đặt vấn đề: Bên cạnh lý thuyết kiểm tra cổ điển, lý thuyết đáp ứng câu hỏi, đặc biệt là mô hình hai tham số và mô hình đa chiều giúp đánh giá sâu hơn về độ khó, độ phân biệt và cấu trúc tiềm ẩn của đề thi. **Mục tiêu nghiên cứu:** Đánh giá đặc tính đo lường của bộ đề thi tốt nghiệp bác sĩ đa khoa tại Trường Đại học Nguyễn Tất Thành bằng cách kết hợp phân tích theo lý thuyết kiểm tra cổ điển, lý thuyết đáp ứng câu hỏi và mô hình đa chiều. **Đối tượng và phương pháp nghiên cứu:** Nghiên cứu mô tả cắt ngang thực hiện trên 118 sinh viên Y6, với đề thi trắc nghiệm tốt nghiệp gồm 120 câu hỏi. Sau đó, lý thuyết kiểm tra cổ điển và lý thuyết đáp ứng câu hỏi được áp dụng để ước lượng tham số độ khó và độ phân biệt. Mô hình đa chiều với cấu trúc đơn giản được sử dụng để đánh giá các đặc tính câu hỏi trong từng nhóm chuyên ngành. **Kết quả:** Thang đo đạt độ tin cậy khá (Cronbach's alpha = 0,769). Theo lý thuyết kiểm tra cổ điển, có 15 câu (12,8%) khó, 40 câu (34,2%) dễ. Về độ phân biệt số câu phân biệt chưa tốt chiếm 80/120 câu. Phân tích theo lý thuyết đáp ứng câu hỏi cho thấy 32 câu có hệ số phân biệt trung bình trở lên ($a > 0,65$). Có 57/118 câu dễ ($b < -2$), trong khi chỉ 24 câu được xem là khó ($b > 1$). Trong mô hình đa chiều, nhóm Nội khoa có chất lượng tốt nhất với 12 câu phân biệt tốt, trong khi nhóm Ngoại khoa có đến 16 câu phân biệt kém. **Kết luận:** Việc kết hợp các phương pháp giúp đánh giá toàn diện đặc tính đo lường của đề thi và cải thiện chất lượng đề thi.

Từ khóa: Lý thuyết ứng đáp câu hỏi, Độ khó và độ phân biệt, Chất lượng đề thi trắc nghiệm.

ABSTRACT

EVALUATING THE DIFFICULTY AND DISCRIMINATION OF THE FINAL EXAMINATION FOR 6TH-YEAR MEDICAL STUDENTS - NGUYEN TAT THANH UNIVERSITY FROM CTT TO MIRT APPROACH

Vu Anh Duy, Le Minh Nguyet, Nguyen Van Tri, Tran Ngoc Sinh, Trinh Huu Tung,
Le Quang Thanh, Hoang Van Quang, Do Trong Anh, Le Thi Kim Nhung*

Nguyen Tat Thanh University

Background: In addition to Classical Test Theory, Item Response Theory, particularly the two-parameter logistic model and the multidimensional model—provides deeper insights into item difficulty, discrimination, and the latent structure of examinations. **Objectives:** To evaluate the psychometric properties of the final examination for medical graduates at Nguyen Tat Thanh University by combining analyses from three methods. **Materials and methods:** A cross-sectional descriptive study was conducted on 118 sixth-year medical students, using a 120-item multiple-choice final examination. Classical test theory analysis and the Item Response Theory model were used to estimate item difficulty and discrimination parameters. A simple-structure multidimensional model

was then employed to assess item characteristics within majors. **Results:** The test demonstrated acceptable reliability (Cronbach's alpha = 0.769). Based on classical test theory, 15 items (12.8%) were difficult, and 40 items (34.2%) were easy. In terms of discrimination, 80 out of 120 items showed poor discriminative ability. Under the Item Response Theory analysis, 32 items achieved at least moderate discrimination ($a > 0.65$). A total of 57 items were classified as easy ($b < -2$), while only 24 were considered difficult ($b > 1$). In the simple-structure multidimensional model, the Internal Medicine domain performed best with 12 well-discriminating items, whereas the Surgery domain had 16 poorly discriminating items. **Conclusion:** The integration of those methods enables a comprehensive evaluation of the psychometric properties of examinations, thereby providing guidance for improving test quality.

Keywords: Item Response Theory, Item difficulty and discrimination, Multiple-choice exam quality.

I. ĐẶT VẤN ĐỀ

Lý thuyết kiểm tra cổ điển (Classical Test Theory – CTT) từ lâu đã được sử dụng phổ biến trong phân tích các bộ đề thi y khoa trong nước [1], [2]. Tuy nhiên phương pháp này còn nhiều hạn chế như phụ thuộc vào mẫu khảo sát và không cung cấp đặc tính riêng cho từng câu hỏi [3]. Trong khi đó, lý thuyết ứng đáp câu hỏi (Item Response Theory – IRT), đặc biệt là mô hình hai tham số (2PL) và mô hình đa chiều (MIRT), cho phép ước lượng các tham số đặc trưng của từng câu hỏi một cách chính xác hơn, đồng thời đánh giá cấu trúc tiềm ẩn của bài thi theo từng lĩnh vực kiến thức [4]. Việc kết hợp giữa CTT, IRT 2PL và MIRT trong đánh giá bộ đề không chỉ giúp kiểm định chất lượng đo lường mà còn góp phần cải tiến quy trình xây dựng đề thi, hướng đến một hệ thống đánh giá năng lực sinh viên khách quan và khoa học hơn. Do đó, nghiên cứu được thực hiện với mục tiêu: Đánh giá đặc tính đo lường của bộ đề thi tốt nghiệp bác sĩ đa khoa tại Trường Đại học Nguyễn Tất Thành bằng cách kết hợp lý thuyết kiểm tra cổ điển và lý thuyết ứng đáp câu hỏi. Xác định mối tương quan giữa các nhóm năng lực Nội, Ngoại, Sản, Nhi theo mô hình MIRT.

II. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Đối tượng nghiên cứu

Nghiên cứu được thực hiện trên 118 sinh viên năm 6 ngành Y tại Trường Đại học Nguyễn Tất Thành, với bộ đề thi trắc nghiệm tốt nghiệp gồm 120 câu hỏi.

- **Tiêu chuẩn chọn mẫu:** Sinh viên năm thứ sáu (Y6) ngành Bác sĩ đa khoa của Trường Đại học Nguyễn Tất Thành tham dự kỳ thi tốt nghiệp Y6 năm 2025 và hoàn thành bài thi trắc nghiệm gồm 120 câu hỏi trong thời gian quy định.

- **Tiêu chuẩn loại trừ:** Bài thi không hoàn chỉnh (bỏ trống >10% số câu hỏi) và dữ liệu không hợp lệ trong quá trình nhập liệu.

2.2. Phương pháp nghiên cứu

- **Phương pháp chọn mẫu:** Nghiên cứu được tiến hành trên toàn bộ sinh viên năm thứ sáu (Y6) ngành Bác sĩ đa khoa đủ tiêu chuẩn tham dự kỳ thi tốt nghiệp năm 2025 tại Trường Đại học Nguyễn Tất Thành. Các trường hợp không đáp ứng tiêu chuẩn lựa chọn hoặc thoả tiêu chuẩn loại trừ được loại trừ trước khi phân tích nhằm đảm bảo chất lượng dữ liệu.

- **Phương pháp nghiên cứu:** Áp dụng lý thuyết kiểm tra cổ điển (CTT) để đánh giá độ khó và độ phân biệt của từng câu hỏi. Đối với các câu hỏi tất cả sinh viên đều trả lời đúng hoặc trả lời sai thì loại ra trước khi phân tích. Tiếp theo, mô hình IRT hai tham số (2PL) được sử dụng để ước lượng các tham số độ khó và độ phân biệt, đồng thời kiểm định độ phù hợp mô hình. Bộ đề sau đó được phân chia thành bốn nhóm theo chuyên ngành: Nội

khoa, Ngoại khoa, Sản – Nhi và Tổng hợp. Mô hình IRT đa chiều với cấu trúc đơn giản (MIRT) được áp dụng để phân tích các tham số trong từng nhóm. Chúng tôi phân tích nghiên cứu bằng phần mềm R 4.4.0.

+ **Lý thuyết kiểm tra cổ điển (CTT) [5]**

Độ khó (p): trong CTT, độ khó của một câu hỏi được xác định bằng tỉ lệ phần trăm thí sinh trả lời đúng câu hỏi đó.

$$p = \frac{\text{số thí sinh trả lời đúng}}{\text{tổng số thí sinh}}$$

Giá trị p nằm trong khoảng từ 0 đến 1, với giá trị càng cao thể hiện câu hỏi càng dễ. Giá trị tối ưu thường nằm trong khoảng 0,3 đến 0,8. Trong nghiên cứu của chúng tôi $p < 0,3$ là câu hỏi khó và $p > 0,8$ là câu hỏi dễ.

Độ phân biệt (r.cor): đo lường khả năng phân biệt giữa thí sinh có năng lực cao và thấp. Trong nghiên cứu này, chúng tôi sử dụng hệ số tương quan điểm câu hỏi với điểm tổng (item-total correlation, r.cor) để đánh giá. Giá trị r.cor dao động từ -1 đến 1; r càng cao thể hiện khả năng phân biệt tốt. Trong nghiên cứu này, chúng tôi phân loại: $r < 0,2$ = phân biệt kém; $0,2 \leq r < 0,4$ = phân biệt trung bình; $r \geq 0,4$ = phân biệt tốt.

+ **Mô hình IRT 2PL [6], [7]**

Trong IRT, mỗi câu hỏi được đặc trưng bởi hàm đáp ứng dựa trên năng lực (θ) của thí sinh. Mô hình IRT 2PL đánh giá hai giá trị độ khó (b) và độ phân biệt (a).

Độ khó (b): là giá trị của năng lực θ tại đó xác suất trả lời đúng của thí sinh là 50%. Tham số b càng cao thì câu hỏi càng khó. Giá trị b thường dao động từ -3 đến +3 trong thang đo chuẩn hóa. Câu hỏi được gọi là khó khi $b > 1$; với $b < -2$: câu hỏi rất dễ, $b < -0,5$: câu hỏi dễ.

Độ phân biệt (a): Tham số a đo độ dốc của hàm đáp ứng tại điểm b. Nó phản ánh mức độ của câu hỏi trong việc phân biệt giữa thí sinh có năng lực gần điểm b. Trong nghiên cứu của chúng tôi đánh giá giá trị tham chiếu của $a > 0,65$ được gọi là câu hỏi có mức độ phân biệt trung bình trở lên, $a > 1,70$ là câu hỏi có mức độ phân biệt tốt, $a < 0,65$ là câu hỏi có mức độ phân biệt kém.

+ **Mô hình IRT đa chiều với cấu trúc đơn giản (MIRT) [5,6]**

Trong MIRT, mỗi câu hỏi có thể liên quan đến nhiều chiều năng lực ($\theta_1, \theta_2, \dots, \theta_k$).

Hàm đáp ứng tổng quát cho mô hình logistic đa chiều hai tham số là:

$$P(\theta) = \frac{1}{(1 + \exp[-(a_1\theta_1 + a_2\theta_2 + \dots + a_k\theta_k - d)])}$$

Trong đó:

- $\theta = (\theta_1, \theta_2, \dots, \theta_k)$: năng lực đa chiều
- a_j : hệ số phân biệt trên chiều j
- d: điểm chặn (intercept) – tương tự như một độ khó tổng hợp, không có ý nghĩa trực tiếp như b trong IRT đơn chiều. Với $d < -0,5$ là câu hỏi khó, $d > 0,5$ là câu hỏi dễ.

Trong nghiên cứu của chúng tôi đánh giá năng lực làm đề thi tốt nghiệp ở sinh viên thành 4 nhóm năng lực F1 (Nội khoa), F2 (Ngoại khoa), F3 (Sản – Nhi), F4 (Khác) dựa theo số câu hỏi phân bố theo từng chuyên ngành

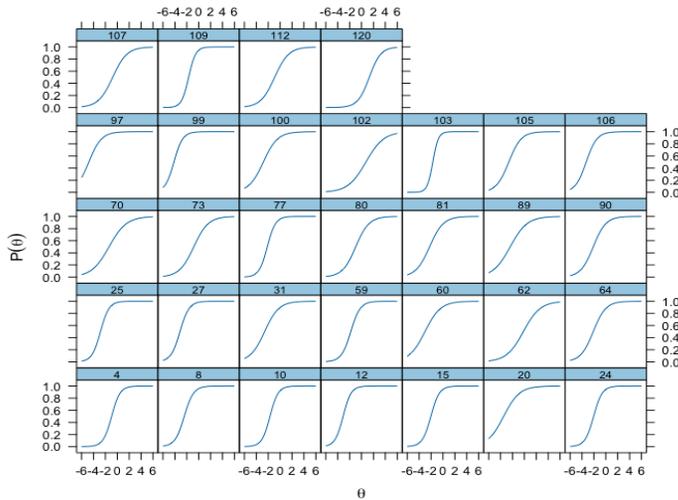
III. KẾT QUẢ NGHIÊN CỨU

3.1. Độ khó và độ phân biệt theo lý thuyết kiểm tra cổ điển

Bộ đề thi gồm 120 câu hỏi trắc nghiệm tốt nghiệp dành cho 118 sinh viên Y6 Trường Đại học Nguyễn Tất Thành được phân tích. Sau khi loại trừ 2 câu hỏi không có phương sai

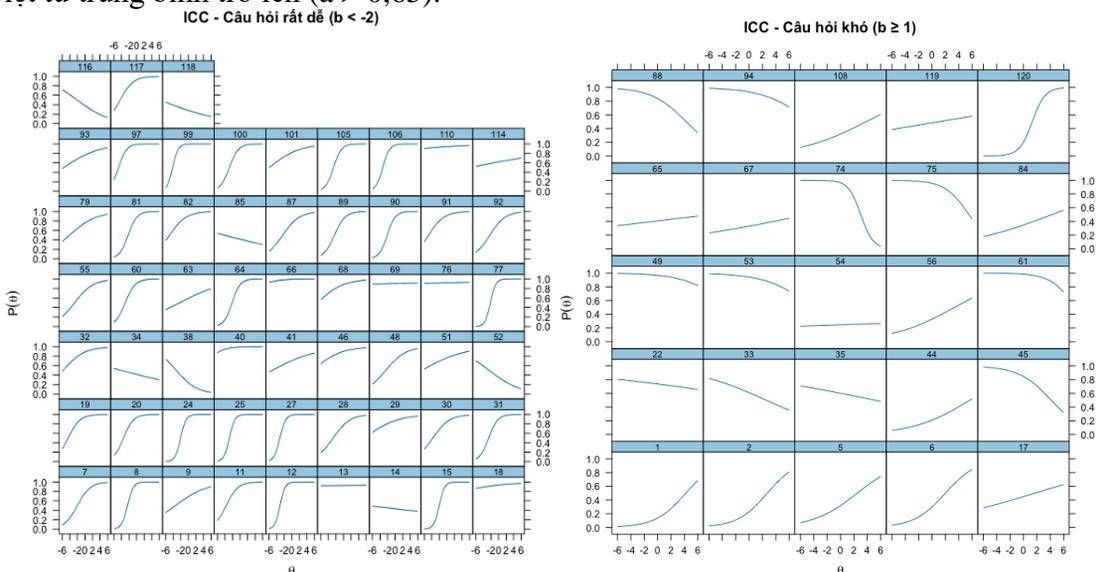
do tất cả sinh viên đều trả lời đúng, số câu còn lại (118 câu) được đưa vào phân tích. Kết quả cho thấy thang đo có độ tin cậy khá tốt với Cronbach's alpha = 0,769, alpha chuẩn hóa = 0,782. Chỉ số Guttman's Lambda 6 đạt 0,996 và chỉ số tín hiệu-nhiều (S/N) là 3,59, cùng cố tính ổn định nội tại. Độ khó của các câu hỏi dao động từ 0,02 đến 0,99, trung bình 0,63. Có 15 câu (12,8%) khó ($p < 0,3$), 63 câu (53%) mức trung bình ($0,3 \leq p \leq 0,8$), và 40 câu (34,2%) quá dễ ($p > 0,8$). Về độ phân biệt, chỉ số tương quan điểm từng câu với điểm tổng (r.cor) dao động từ -0,05 đến 0,43, trung bình 0,15. Trong đó, 80 câu (68,4%) có khả năng phân biệt kém ($r < 0,2$), 34 câu (29,1%) ở mức trung bình ($0,2 \leq r < 0,4$), và chỉ 3 câu (2,6%) phân biệt tốt ($r \geq 0,4$).

3.2. Mô hình IRT 2PL và mô hình IRT đa chiều với cấu trúc đơn giản



Hình 1. Số câu có mức độ phân biệt tốt trong đề thi tốt nghiệp Y6

Nhận xét: Phân tích theo mô hình IRT 2PL cho thấy có 32 câu đạt khả năng phân biệt từ trung bình trở lên ($a > 0,65$).



Hình 2. Hình bên phải nhóm câu hỏi khó ($b \geq 1$), hình bên trái: nhóm câu hỏi rất dễ ($b < -2$). Đường cong thể hiện xác suất trả lời đúng theo mức năng lực (θ), qua đó đánh giá được khả năng phân biệt và độ khó của từng câu.

Nhận xét: Về độ khó, có 24 câu (20,3%) được xếp loại khó ($b \geq 1$), với đường cong đặc trưng lệch phải, phản ánh khả năng phân loại tốt ở nhóm sinh viên có năng lực cao. Ngược lại, có 57 câu (48,3%) rất dễ ($b < -2$), đường cong lệch trái, cho thấy xác suất trả lời đúng cao ngay cả ở nhóm năng lực thấp.

Trong mô hình MIRT với bốn nhóm năng lực (F1: Nội khoa, F2: Ngoại khoa, F3: Sản - Nhi, F4: Khác), tham số phân biệt (a) dao động rộng, từ rất thấp ($a \approx 0,05$ ở một số câu F1) đến cao ($a > 1,0$ ở nhiều câu thuộc F2 và F4). Tham số điểm chặn (d) cho thấy có 16/118 câu khó ($d < -0,5$) và 70/118 câu dễ ($d > 0,5$). Tương quan giữa các nhóm năng lực cho thấy F1 (Nội khoa) có liên hệ mức trung bình với F2 (Ngoại khoa, $r = 0,331$) và F3 (Sản - Nhi, $r = 0,526$), và tương quan cao với F4 (Khác, $r = 0,776$). Ngược lại, F2 (Ngoại khoa) và F3 (Sản - Nhi) gần như không tương quan ($r = -0,039$).

Bảng 1. Phân bố câu hỏi theo độ khó ở các nhóm năng lực

Nhóm năng lực	Khó ($d < -0,5$)	Trung bình ($-0,5 \leq d \leq 0,5$)	Dễ ($d > 0,5$)	Tổng số câu
F1 (Nội khoa)	4	6	17	27
F2 (Ngoại khoa)	5	8	14	27
F3 (Sản - Nhi)	3	10	19	32
F4 (Khác)	4	8	20	32
Tổng số câu	16	32	70	118

d: “Phân loại theo tham số d trong mô hình MIRT”.

Nhận xét: Khi phân tích độ khó, phần lớn câu hỏi ở cả bốn nhóm đều có độ khó ở mức dễ, đặc biệt ở nhóm F3 (Sản - Nhi) và F4 (Khác), với tỉ lệ câu hỏi dễ lần lượt là 61,29% và 62,50%. Ngược lại, số lượng câu hỏi ở mức trung bình còn khá hạn chế.

Bảng 2. Phân bố câu hỏi theo độ phân biệt ở các nhóm năng lực

Nhóm năng lực	Phân biệt kém	Phân biệt trung bình	Phân biệt tốt	Tổng số câu
F1 (Nội khoa)	6	9	12	27
F2 (Ngoại khoa)	16	9	2	27
F3 (Sản - Nhi)	13	12	7	32
F4 (Khác)	13	8	11	32
Tổng số câu	48	38	32	118

Nhận xét: Khi đánh giá độ phân biệt cho thấy nhóm F1 (Nội khoa) có chất lượng câu hỏi cao nhất với 12/27 câu hỏi có độ phân biệt tốt và chỉ 6/27 câu có độ phân biệt kém. Trong khi đó, nhóm F2 (Ngoại khoa) có tới 16/27 câu hỏi (59,3%) thuộc loại phân biệt kém, chỉ có 2 câu được xếp loại tốt. Nhóm F4 có phân bố 11 câu tốt và 13 câu kém.

IV. BÀN LUẬN

4.1. Độ khó và độ phân biệt theo lý thuyết kiểm tra cổ điển (CTT)

Kết quả cho thấy thang đo có độ tin cậy khá tốt với hệ số Cronbach’s alpha = 0,769 (alpha chuẩn hóa = 0,782), nằm trong khoảng 0,7–0,8, mức giá trị được xem là chấp nhận được khi đánh giá độ nhất quán nội bộ của công cụ đo lường. Điều này phù hợp với các nghiên cứu trong lĩnh vực giáo dục y khoa và đánh giá bài thi, tác giả Mohsen cho rằng Cronbach’s alpha $\geq 0,70$ và $\leq 0,9$ thường được coi là đạt yêu cầu về độ tin cậy để sử dụng trong phân tích tiếp theo [8]. Chỉ số Guttman’s Lambda 6 và tín hiệu-nhiều (S/N) cũng có tính ổn định của thang đo.

Tuy nhiên, phân bố độ khó chưa hợp lý. Tỷ lệ câu hỏi dễ và rất dễ chiếm đến 34,2%, trong khi chỉ có 12,8% câu khó. Điều này có nguy cơ gây ra hiệu ứng trần (ceiling effect), khiến đề thi chưa đủ khả năng phân loại sinh viên khá – giỏi. Về độ phân biệt, đa số câu hỏi có khả năng phân biệt kém (68,4%), chỉ 2,6% đạt mức tốt. Kết quả này phản ánh rằng nhiều câu hỏi chưa thực sự đo lường được sự khác biệt về năng lực giữa các sinh viên. Đặc biệt, một số câu hỏi có chỉ số tương quan âm, cho thấy nội dung hoặc cách diễn đạt chưa phù hợp, cần được điều chỉnh hoặc loại bỏ. Nhìn chung, phân tích theo CTT chỉ ra rằng ngân hàng câu hỏi tuy có độ tin cậy tổng thể, nhưng phân bố độ khó và độ phân biệt chưa cân bằng, cần tiếp tục chuẩn hóa để đảm bảo tính công bằng và hiệu quả trong đánh giá.

4.2. Mô hình IRT 2PL và mô hình IRT đa chiều (MIRT)

Khi áp dụng IRT 2PL, chỉ có 32 câu đạt khả năng phân biệt trung bình trở lên ($a > 0,65$). Về độ khó, có 24 câu khó ($b \geq 1$) và đến 57 câu rất dễ ($b < -2$). Điều này phù hợp với kết quả từ CTT, cho thấy bộ đề vẫn thiên về nhiều câu dễ, hạn chế trong việc phân loại sinh viên có năng lực cao. Đường cong đặc trưng của các câu hỏi cũng minh họa rõ hiện tượng này, khi nhiều câu có xác suất trả lời đúng cao ngay cả ở mức năng lực thấp.

Trong mô hình MIRT với bốn nhóm năng lực (Nội khoa, Ngoại khoa, Sản – Nhi, Khác), sự khác biệt giữa các nhóm được thể hiện rõ. Nhóm Nội khoa có chất lượng tốt nhất với 12/27 câu phân biệt tốt, trong khi nhóm Ngoại khoa có đến 59,3% câu phân biệt kém và chỉ 2 câu đạt mức tốt. Đây là điểm yếu nổi bật, gợi ý rằng bộ câu hỏi Ngoại khoa cần được rà soát và bổ sung những câu có khả năng phân biệt cao hơn.

Ngoài ra, phân tích tương quan cho thấy năng lực Nội khoa có mối liên hệ mạnh với các lĩnh vực khác, đặc biệt với nhóm Khác ($r = 0,776$). Ngược lại, Ngoại khoa và Sản – Nhi gần như độc lập ($r = -0,039$), phản ánh tính đặc thù của từng chuyên ngành. Kết quả này cho thấy MIRT không chỉ đánh giá đặc điểm câu hỏi mà còn cung cấp thông tin giá trị về mối liên hệ giữa các lĩnh vực kiến thức, giúp định hướng phát triển đề thi đa chiều và toàn diện.

V. KẾT LUẬN

Phân tích đề thi tốt nghiệp Y6 bằng cả hai phương pháp CTT và IRT cho thấy thang đo có độ tin cậy tương đối tốt, nhưng chất lượng từng mục câu hỏi vẫn còn nhiều điểm cần cải thiện. Dưới góc nhìn CTT, phần lớn câu hỏi có độ khó nằm trong khoảng trung bình, tuy nhiên tỷ lệ câu hỏi phân biệt tốt còn rất thấp, với đa số chỉ đạt mức kém hoặc trung bình. Về mô hình IRT 2PL và mô hình đa chiều MIRT, phân tích sâu hơn về các tham số độ khó và độ phân biệt đã chỉ ra sự không đồng đều giữa các nhóm nội dung. Do đó có thể cần cải thiện bộ đề thi Y6 để đánh giá và cải thiện độ phân biệt tốt hơn.

TÀI LIỆU THAM KHẢO

1. Bùi Anh Tú, Võ Đăng Khoa, Nguyễn Anh Vũ. Đánh giá đề thi MCQ đầu vào và tương quan độ khó, độ phân cách các câu. *Tạp chí Y học Thành phố Hồ Chí Minh*. 2024. 27(5), 38-47, doi: 10.32895/hcjm.m.2024.05.05.
2. Phạm Dương Uyển Bình, Trần Thị Diệu. Mối quan hệ giữa độ khó năng lực và độ phân cách của câu hỏi trắc nghiệm sinh lý học trong các đề thi tuyển sinh sau đại học dành cho đối tượng chuyên khoa cấp I từ năm 2018-2022 tại Đại học Y Dược Thành Phố Hồ Chí Minh. *Tạp chí Y học Thành phố Hồ Chí Minh*. 2024. 27 (1), 170-176, doi: 10.32895/hcjm.m.2024.01.24.
3. T. Rusch, P. Lowry, P. Mair, and H. Treiblmaier. Breaking Free from the Limitations of Classical Test Theory: Developing and Measuring Information Systems Scales Using Item Response Theory. *Information & Management*. 2017. 54 (2), 189-203, doi: 10.1016/j.im.2016.06.005.

4. Nathan Thompson. Classical Test Theory vs Item Response Theory. 2023. <https://assess.com/classical-test-theory-vs-item-response-theory>.
 5. Pimentel J.L., Villaruz M.L.A. Comparison of item difficulty estimates in a basic statistics test using *ltm* and CTT software packages in R. *International Journal of Advanced Computer Science and Applications*. 2020. 11(3), 367-372, doi: 10.14569/IJACSA.2020.0110346.
 6. De Ayala R.J. *The Theory and Practice of Item Response Theory*. Guilford Press. 2022.
 7. Paek I., Cole K. Using R for Item Response Theory Model Applications. Routledge. 2020.
 8. Tavakol M, Dennick R. Making sense of Cronbach's alpha. *Int J Med Educ*. 2011. 2, 53-55, doi: 10.5116/ijme.4dfb.8dfd.
-