THIẾT KẾ ỨNG DỤNG WEB CHO TRẮC NGHIỆM TRÊN MÁY TÍNH THÍCH NGHI NĂNG LỰC THÍ SINH TRONG GIÁO DỤC KHOA HỌC CƠ BẢN TẠI ĐẠI HỌC Y DƯỢC THÀNH PHỐ HỒ CHÍ MINH

Vĩnh Sơn[1], Trần Thị Diệu[2], Hoàng Đạo Bảo Trâm[1], Nguyễn Anh Vũ[2], Phạm Dương Uyển Bình[1], Phạm Lê An[3]

TÓM TẮT

Mục đích: Nghiên cứu này nhằm thiết kế và phát triển một phần mềm web trắc nghiệm thích nghi năng lực thí sinh (CAT) phục vụ cho sinh viên năm thứ nhất tại Đại học Y Dược Thành phố Hồ Chí Minh. Nội dung chính tập trung vào việc phát triển thuật toán hệ thống CAT.

Phương pháp: Nguyên tắc thiết kế thuật toán hệ thống dựa trên cơ sở tích hợp mô hình đo lường và phân tích Rasch vào trắc nghiệm thích nghi năng lực thí sinh trên máy tính, và sử dụng ngôn ngữ lập trình PHP và cơ sở dữ liệu MySQL. Trong nghiên cứu này, chúng tôi mô tả những nguyên tắc chủ chốt và một số mô phỏng được dùng để xây dựng và đánh giá thuật toán, đồng thời phát hiện các hạn chế của thuật toán và xác định cách giải quyết.

Kết quả: Bài kiểm tra trắc nghiệm thích nghi năng lực thí sinh đã được chứng minh là nhỏ gọn và chính xác hơn so với phiên bản không thích nghi. Thông qua mô phỏng, các tham số có thể được điều chỉnh dựa trên mục tiêu và ngân hàng câu hỏi thi. Bên cạnh đó, một số kết quả của việc triển khai CAT cho sinh viên y khoa năm thứ nhất cũng được cung cấp.

Kết luận: Kết quả nghiên cứu cho thấy ứng dụng CAT hoạt động hiệu quả theo đúng thiết kế. Tuy nhiên, thuật toán vẫn còn một số hạn chế như phương pháp ước lượng có độ chính xác chưa cao, ngân hàng câu hỏi cỡ nhỏ và tỷ lệ bộc lộ cao cho một số câu hỏi cụ thể. Vì vậy, cần nâng cao thuật toán để cải thiện việc bảo mật ngân hàng câu hỏi.

Từ khóa: trắc nghiệm thích nghi năng lực thí sinh, mô hình đo lường Rasch, mô phỏng, phần mềm mạng.

ABSTRACT

DESIGNING A WEB APPLICATION FOR COMPUTERIZED ADAPTIVE TESTING IN BASIC SCIENCES EDUCATION IN UNIVERSITY OF MEDICINE AND PHARMACY AT HO CHI MINH UNIVERSITY

Purpose: This study is to create and execute a web-based tool for computerized adaptive testing (CAT) specifically tailored for first-year students at University of Medicine and Pharmacy at Ho Chi Minh City, focus on developing the CAT system algorithm.

Methods: The designing principle lies in incorporating Rasch measurement and analysis into an adaptive testing software that employs the programming language PHP with MySQL database. In this study, we describe the core principles of the algorithm and the simulations

[1] Phòng Đảm bảo Chất lượng Giáo dục và Khảo thí, Đại học Y Dược TP Hồ Chí Minh
[2] Khoa Khoa học Cơ bản, Đại học Y Dược TP Hồ Chí Minh
[3] Khoa Y, Đại học Y Dược TP Hồ Chí Minh
Tác giả liên hệ: Nguyễn Anh Vũ, *ĐT:* 0909090838 *Email:* nguyenanhvu@ump.edu.vn

conducted to construct and evaluate it, while also identify any limitations of the CAT algorithm and to suggest ways to improve it.

Results: The computerized adaptive test was shown to be leaner and more accurate than the non-adaptive one. Through simulations, we were able to fine-tune certain parameters based on our goals and the item bank. Additionally, some results of CAT implementing for first-year medical students were provided.

Conclusion: The findings show that the CAT web app performs effectively as intended. However, there are certain limitations in the algorithm such as the estimation method, a limited item bank, and a high exposure rate for certain items. So it is necessary to enhance the algorithm in order to improve the item bank security.

Keywords: computerized adaptive tests, Rasch measurement model, simulation, web application

I. ĐẶT VẤN ĐỀ

Bài kiểm tra trắc nghiệm đánh giá năng lực với độ dài cố định tuy là hình thức trắc nghiệm rất phổ biến nhưng có một số khuyết điểm. Chẳng hạn như, số câu hỏi trắc nghiệm trong bài kiểm tra phải lớn và thời gian làm bài thường dài. Đề kiểm tra thường chứa những câu hỏi quá khó hoặc quá dễ. Điều này không phù hợp với một số yêu cầu bài kiểm tra, như kiểm tra đánh giá năng lực của học viên trước khóa học. Đồng thời sai số điểm số đánh giá cũng không đồng đều giữa các thí sinh[3,9]. Hình thức thi trắc nghiệm hiện nay cũng thường áp dụng những bộ câu hỏi tương tự cấu trúc và nội dung cho toàn bộ các thí sinh dự thi cùng kỳ thi. Điều này có thể tạo điều kiện cho thí sinh trao đổi đáp án với nhau, đặc biệt khi thi trực tuyến không tập trung thí sinh tại phòng thi. Phương pháp trắc nghiệm thích nghi trên máy tính có thể giải quyết các khuyết điểm của trắc nghiệm cố định.[4,6]. Khác với trắc nghiệm cố định, trắc nghiệm thích nghi bắt đầu từ một câu hỏi có độ khó trung bình. Nếu trả lời đúng, thí sinh sẽ nhận câu hỏi có độ khó cao hơn. Nếu trả lời sai, thí sinh sẽ nhận câu hỏi có độ khó thấp hơn. Sau mỗi lần trả lời, năng lực thí sinh được ước lượng cùng với sai số tiêu chuẩn, quá trình lặp tiếp tục cho đến khi năng lực thí sinh được ước tính đủ chính xác.[1,2,8] Nghiên cứu này nhằm thiết kế và phát triển phát triển thuật toán hệ thống UMPCAT, một phần mềm web trắc nghiệm thích nghi năng lực thí sinh, phục vụ cho sinh viên năm thứ nhất tại Đại học Y Dược Thành phố Hồ Chí Minh.

II. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU

1. Mô tả cắt ngang: Thực nghiệm thăm dò tính giá trị của UMPCAT, dựa trên một bài trắc nghiệm thích nghi trên tập hợp 654 câu hỏi đã chuẩn độ theo đo lường Rasch, với mẫu tiện lợi gồm 99 sinh viên khoa Y khóa 2021, đánh giá năng lực đầu vào môn Tin học.

2. Các thành phần của UMPCAT:

Thành phần phần mềm	Thông tin
Ngôn ngữ lập trình	PHP
Cơ sở dữ liệu	MySQL
Mô hình đo lường	Lưỡng phân Rasch, sử dụng một tham số độ khó logit.
Ngân hàng câu hỏi	Môn Tin học, 654 câu, đã chuẩn độ logit theo Rasch.
Câu hỏi khởi đầu	Tập con của tập các câu hỏi có độ khó gần mức θ = 0 logit

Thành phần phần mềm	Thông tin
Quy tắc chọn câu hỏi	Fisher maximum information.
Phương pháp lượng giá	Maximum likelihood
Quy tắc dừng thuật toán	4 quy tắc: độ chính xác; số câu hỏi được dùng; thời gian làm bài; buộc dừng.

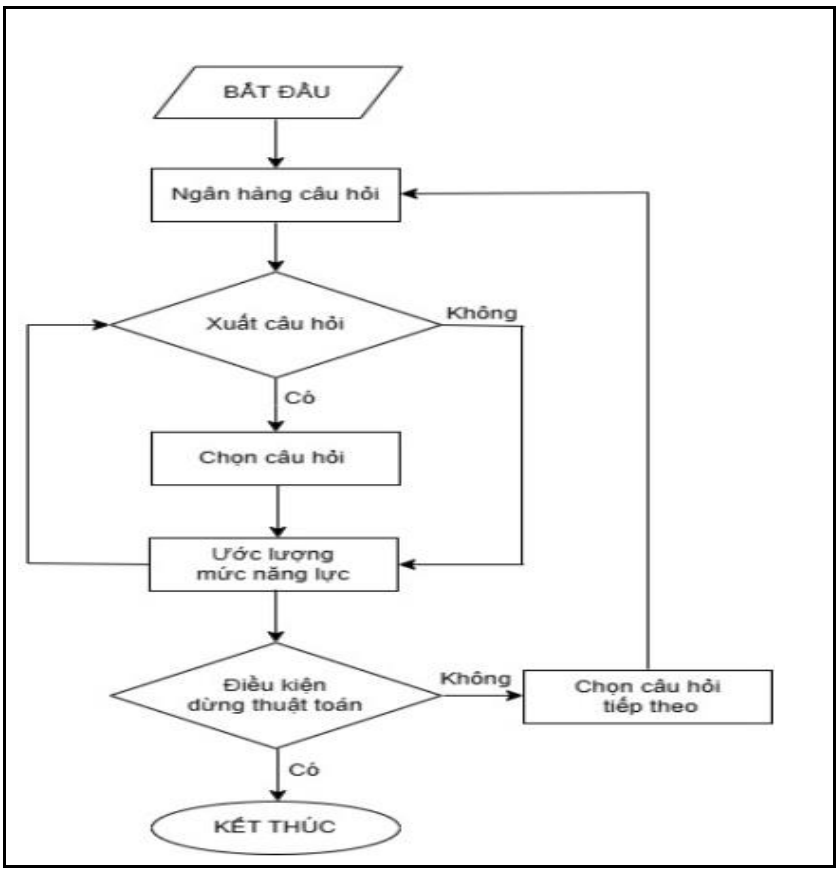
Quy tắc dừng thuật toán	Nội dung
Quy tắc 1	Độ chính xác ước lượng đạt yêu cầu
Quy tắc 2	Số câu hỏi vượt quá số câu thiết lập
Quy tắc 3	Thời gian làm bài vượt quá thời gian thiết lập
Quy tắc 4	Bắt buộc dừng thuật toán: (1) Thí sinh trả lời đúng hoặc sai tất cả câu hỏi Hoặc (2) Toàn bộ câu hỏi trong ngân hàng đã được dùng

3. Cấu hình máy tính chạy UMPCAT: CPU Intel Core i5/i7, AMD Ryzen 5/7; RAM 16GB; hệ điều hành Window 10, dung lượng bộ nhớ SSD trống 250 GB, trình duyệt Chrome.

4. Quy trình thực nghiệm:

Bước	Hoạt động	Điều kiện
1	Phần cứng, phần mềm mạng. Địa điểm, thời gian, người tham gia.	Phòng máy 5, Nhà E,Đại học Y Dược TPHCM Bắt đầu khóa học môn Tin học, tháng 11 năm 2021. Sinh viên Y 2021, tổ 25 – 48.
2	Nạp câu hỏi kiểm tra vào UMPCAT.	Tập hợp 654 câu hỏi trắc nghiệm. Dạng câu hỏi lưỡng phân đúng – sai với 4 lựa chọn.
3	Định tham số thích nghi vào UMPCAT: - Tham số câu hỏi - Tham số đề kiểm tra - Độ khó câu hỏi đầu	 - Mã câu hỏi, độ khó logit, đáp án. - Số câu hỏi giới hạn, độ chính xác ước lượng. - Ngẫu nhiên trong khoảng -0,8 – 0,8 logit.
4	Chạy phần mềm UMPCAT.	Kết nối phần mạng ổn định.
5	Rà soát quá hạn tham số đề.	Phát hiện trường hợp cần kiểm tra lại.
5	Thu thập dữ liệu làm bài.	Trích xuất bộ nhớ với quyền hệ thống cấp 2.
6	Phân tích dữ liệu và báo cáo.	Nhóm nghiên cứu Đo lường Giáo dục học.

Để xác định tham số đề kiểm tra, bài kiểm tra giả lập được mô phỏng trên phần mềm UMPCAT và tính toán dựa vào mô hình Rasch. Số thí sinh giả lập là 500, mức năng lực phân bố ngẫu nhiên từ -1,5 đến 2,5 dựa vào kinh nghiệm khảo thí, câu hỏi khởi đầu có độ khó 0,5 logit. UMPCAT xuất câu hỏi kế tiếp dựa trên kết quả trả lời trước đó của thí sinh. Khả năng trả lời đúng mỗi câu hỏi của mỗi thí sinh giả lập được tính toán theo mô hình Rasch, được quyết định đúng sai dựa vào ngưỡng phân biệt 0,5. Thử nghiệm được lặp 1000 lần. Kết quả mô phỏng cho thấy câu hỏi trung bình là 45 và SE = 0,3.



Hình 1. *Quy trình thực hiện trích xuất ngân hàng đề và chỉ định câu hỏi kiểm tra.*

III. KẾT QUẢ NGHIÊN CỨU

Bảng 1. *Độ khó logit câu hỏi trong ngân hàng câu hỏi thử nghiệm môn Tin học 654 câu*

Tham số	Trung bình (ĐLC)	Min	Max	Trung vị (Q25-Q75)	Shapiro Wilk	P
Độ khó logit của câu hỏi	-0,0909 (1,4541)	-5,881	4,950	-0,1591 (-1,041 ; 0,803)	0,993	0,006

Nhận xét: Độ khó logit câu hỏi có phân phối đối xứng khá rõ và nhọn so với phân phối chuẩn.

Bảng 2. Mức năng lực thí sinh, độ dài bài thi và thời gian làm bài

Biến số	Chung (n = 99)	Nam (n = 62)	Nữ (n = 37)	p
Năng lực thí sinh, (logit) Trung bình ± ĐLC	-0,743 ± 1,002	-0,766 ± 1,081	-0,704 ± 0,865	0,769a
Số câu hỏi được dùng Trung bình ± ĐLC	44,68 ± 4,96	44,80 ± 5,40	44,60 ± 4,3	0,606b
Thời gian làm bài(phút) Trung bình ± ĐLC	28,8 ± 8,7	28,00 ± 8,8	30,1 ± 8,6	0,236a

[a] *Kiểm định Student, biểu diễn số liệu với trung bình ± độ lệch chuẩn*
[b] *Kiểm định Mann-Whitney, biểu diễn số liệu với trung bình ± độ lệch chuẩn*

Nhận xét: Năng lực thí sinh có phân phối chuẩn, phạm vi giá trị từ -3,5 đến 1,4 logit. Số câu hỏi được sử dụng từ 20 đến 58 câu, trung bình là 45 câu. Thời gian làm bài có phân phối chuẩn, trung bình là 28,7 phút, thời gian ngắn nhất là 9,7 phút và dài nhất là 50 phút. Giữa hai nhóm giới tính không có khác biệt có ý nghĩa thống kê.

Bảng 3. Lý do dừng thuật toán theo nhóm giới tính

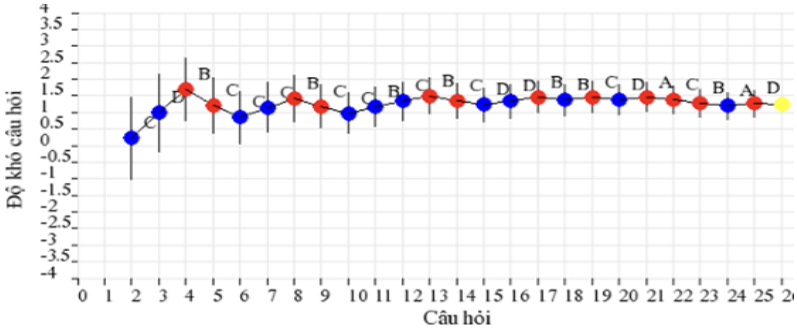
Thí sinh	Lý do dừng thuật toán				p
	Chung	Đủ chính xác	Quá số câu	Quá thời gian	
Nữ, n(%)	37 (37,4)	31 (83,8)	0 (0)	6 (16,2)	0,438
Nam, n(%)	62 (62,6)	54 (87,1)	2 (3,2)	6 (9,7)	
Chung, n(%) KTC95%	99 (100)	85 (85,9) 77,7 – 91,4	2 (2,0) 0,6 – 7,1	12 (12,1) 7,1 – 20,0	

Kiểm định Fisher Exact, biểu diễn số liệu bằng tần số (tỷ lệ)

Nhận xét: Tỷ lệ dừng thuật toán khác biệt không có ý nghĩa thống kê giữa hai nhóm giới tính. Tỷ lệ dừng thuật toán do đạt độ chính xác cần thiết đạt gần 86%. Khác biệt giữa nam và nữ do quá thời gian không có ý nghĩa thống kê (p = 0,340). Hai thí sinh quá số câu do nguyên nhân tâm lý, làm sai toàn bộ hoặc phần lớn các câu có độ khó thấp hơn mức năng lực. Kết quả làm bài lần sau θ = -3,314, SE = 0,305 và θ = 1,695, SE = 0,305.

Biểu đồ 1. Quá trình trắc nghiệm thích nghi năng lực thí sinh với UMPCAT

Bước	Năng lực	SE	idCauHoi	Độ khó (b)	Độ p.cách (a)	Đoán mò (c)	Xác suất (P)	Trả lời	Đúng
0	0.50000	3.00000	-1	0.50000	1	0	0.50000	@	0
1	0.50000	3.00000	-1	0.50000	1	0	0.50000	@	1
2	1.09991	1.22719	18645	0.21068	1.00000	0.00000	0.57183	C	1
3	1.66760	1.16368	19074	0.98858	1.00000	0.00000	0.52780	D	1
4	1.20001	0.94310	19563	1.68107	1.00000	0.00000	0.49663	B	0
5	0.84226	0.84097	19505	1.19632	1.00000	0.00000	0.50092	C	0
6	1.14616	0.78338	18832	0.85053	1.00000	0.00000	0.49793	C	1



Biểu đồ minh họa một điển hình của diễn trình ước lượng mức năng lực thí sinh của phần mềm UMPCAT, kết quả trích xuất từ dữ liệu của một sinh viên có mức năng lực trên trung bình. Mức năng lực thí sinh được ước lượng bằng độ khó logit của câu hỏi cuối cùng trong dãy câu hỏi. Hai câu hỏi khởi đầu có độ khó logit trùng lặp và sai số ước lượng năng lực thí sinh khá lớn. Trong 5 câu hỏi đầu, sai số ước lượng năng lực khá lớn và mức độ giảm sai số tương đối nhỏ.

IV. BÀN LUẬN

Mô hình Rasch có những đặc điểm phù hợp với việc thiết kế và phát triển ứng dụng trắc nghiệm thích nghi CAT vì một số lý do sau. Đây là mô hình đo lường mạnh đồng thời tương đối đơn giản để triển khai trong phần mềm máy tính.[7] Mô hình Rasch cũng giúp hiệu chuẩn các câu hỏi kiểm tra theo một thang đo logit chung, kể cả khi ngân hàng câu hỏi được cập nhật liên tục. Điều này giúp dễ dàng chọn các câu hỏi trắc nghiệm phù hợp với mức năng lực thí sinh, dễ hiệu chỉnh và tối ưu hóa hiệu suất thuật toán.[5,8]

Kết quả nghiên cứu cho thấy so với bài thi trắc nghiệm không thích nghi, thời gian thi giảm đi trung bình 20 phút và kết quả lượng giá có độ chính xác cao hơn với SE là 0,3 và 40 – 60 câu hỏi. Một bài thi không thích nghi thông thường gồm 120 câu hỏi thường có SE khoảng từ 0,5 đến 0,7. Bài kiểm tra thích nghi không có hoặc rất ít những câu hỏi quá khó hoặc quá dễ đối với từng cá nhân, vì vậy rất ít trường hợp căng thẳng tâm lý, một yếu tố gây nhiễu khi đo lường năng lực thực hiện công việc đã được dạy.

Nghiên cứu chỉ ra một số hạn chế của thuật toán và dữ liệu câu hỏi thi ảnh hưởng đến chất lượng đánh giá năng lực thí sinh. Thứ nhất, sai số ước lượng những câu đầu khá lớn, điều này có thể dễ dàng làm cho chọn lựa câu hỏi kế tiếp bị lệch hoặc lặp lại. Sai số lớn dẫn đến khả năng cao là thuật toán chọn lựa câu hỏi kế tiếp trong khoảng logit lân cận, mặt khác thuật toán chỉ lựa chọn trong đó câu hỏi có mức thông tin cao nhất. Đây là nguyên nhân tỷ lệ bộc lộ cao ở một số câu đã được ghi nhận. Thứ hai, ngân hàng câu hỏi còn hạn chế và phổ phân bố độ khó logit còn chỗ hổng dẫn đến khả năng thí sinh có thể nhận được nhiều lần cùng một câu hỏi. Thứ ba, có một số câu hỏi có tần suất được chọn cao hơn những câu hỏi khác. Những điểm yếu này có thể bị dễ dàng lợi dụng làm giảm tính bảo mật của ngân hàng câu hỏi, hoặc làm đề thi vô hiệu[4].

V. KẾT LUẬN

Phần mềm UMPCAT góp phần đáng kể nâng cao chất lượng khảo thí. Tuy nhiên cần cải thiện thuật toán để nâng cao tính bảo mật của ngân hàng câu hỏi, giảm tỷ lệ hiển thị một số câu hỏi nhất định, nâng cao độ chính xác ước lượng, tăng số lượng trong ngân hàng câu hỏi.

TÀI LIỆU THAM KHẢO

1. Lê Thái Hưng, Trần Thị Hoa, Đặng Thị Mây, Hoàng Lan Hương. (2019). Phát triển ngân hàng trắc nghiệm thích ứng để đánh giá năng lực đọc hiểu môn Ngữ văn của học sinh lớp 10 trung học phổ thông. *Tạp chí Khoa học Giáo dục Việt Nam*, Số 24, Tháng 12, 54-59.
2. Lê Xuân Tài, Đặng Hoài Phương (2015). Xây dựng mô hình trắc nghiệm thích nghi trên cơ sở lý thuyết đáp ứng câu hỏi. *Tạp chí Khoa học Đại học Huế*, Tập 97, Số 9, 1-17
3. Ling G., Attali Y., Finn B., Stone E.A. (2017). Is a Computerized Adaptive Test More Motivating Than a Fixed-Item Test? *Applied Psychological Measurement*, 41(7): 495–511
4. Delgado-Gomez D., Laria J.C., Ruiz-Hernandez D. (2018). Computerized adaptive test and decision trees: a unifying approach. *Expert systems with applications*, 117, 358-266
5. Chen, S.Y. (2005). Controlling Item Exposure and Test Overlap in Computerized Adaptive Testing. *Applied Psychological Measurement*, 3(29), 204-217.
6. Eggen, T. J.H.M. (2011). Computerized classification testing with the Rasch model. *Educational Research and Evaluation*, 5(17), 361-371.
7. Magis, D., Yan, D., von Davier, A. (2017). *Computerized Adaptive and Multistage Testing with R: Using Packages catR and mstR*. Springer International Publishing.
8. Travitzky, R., Meneghetti, D.D.R, Alavarse, O.M., Catalani E.M.T. (2018). How to build a Computerized Adaptive Test with free software and pedagogical relevance?. *Proceedings of IAC 2018 in Vienna*, Czech Technical University in Prague.
9. Seo, D.G. (2017). Overview and current management of computerized adaptive testing in licensing/certification examinations. *Journal of Educational Evaluation of Health Professions*, 14-17.