ISSN: 1859-1779

Nghiên cứu Y học
Tạp chí Y học Thành phố Hồ Chí Minh; 27(5):90-99
https://doi.org/10.32895/hcjm.m.2024.05.11

Ứng dụng mô hình đo lường Rasch trong xây dựng ngân hàng câu hỏi đánh giá mức kiến thức người học bằng trắc nghiệm thích nghi trên máy tính

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

[1]Đại học Y Dược Thành phố Hồ Chí Minh, Thành phố Hồ Chí Minh, Việt Nam

Tóm tắt

Đặt vấn đề: Ngân hàng câu hỏi và thuật toán trắc nghiệm thích nghi là hai thành phần quan trọng nhất của hệ thống trắc nghiệm thích nghi trên máy tính (CAT). Chất lượng đo lường của trắc nghiệm thích nghi phụ thuộc vào chất lượng của tập hợp các câu hỏi trắc nghiệm đã được hiệu chuẩn được đưa vào thực hiện lượng giá. Nghiên cứu này tập trung vào đánh giá chức năng, hiệu quả của ngân hàng câu hỏi UMPItemBank nhằm cải tiến hệ thống trắc nghiệm thích nghi UMPCAT.

Mục tiêu: Đánh giá chức năng và hiệu quả vận hành của phần mềm UMPItemBank tích hợp mô hình đo lường Rasch trong phát triển ngân hàng câu hỏi trắc nghiệm thích nghi trên máy tính đánh giá mức kiến thức đầu vào của người học.

Đối tượng và phương pháp nghiên cứu: Phần mềm UMPItemBank được thiết kế chức năng chính quản lý câu hỏi và tạo đề thi, sử dụng ngôn ngữ lập trình PHP và cơ sở dữ liệu MySQL, hệ điều hành Windows 11, lưu trữ đám mây kết hợp, bảo mật xác thực và ủy quyền OAuth 2.0 và WAF. Nội dung ngân hàng câu hỏi được phát triển dựa trên lý thuyết học tập kiến tạo, câu hỏi sau biên soạn được rà soát và thực nghiệm, tích hợp mô hình Rasch, đưa vào bài kiểm tra thích nghi trên máy tính. Phương pháp đánh giá quá trình và nghiên cứu trường hợp được sử dụng nhằm cung cấp thông tin phản hồi liên tục trong bối cảnh thực tế cho nghiên cứu và phát triển phần mềm. Số liệu được quản lý và phân tích bằng Excel MS 365, JASP 0.19.1, Heuristic Lab Optimizer 3.3.16.1786.

Kết quả: Nghiên cứu chỉ ra ngân hàng câu hỏi hoạt động đúng thiết kế, có hiệu quả và tương thích với thuật toán CAT, cung cấp bằng chứng về tính giá trị nội dung, cỡ và dạng ngân hàng, tính công bằng, hiệu quả chi phí, tính khả thi và tiềm năng ứng dụng trong quản lý khảo thí và bài kiểm tra trắc nghiệm thích ứng. Tuy nhiên có một số vấn đề cần khắc phục như kiểm soát ngưỡng sai số chuẩn và mức độ phơi lộ, phân tích sai biệt vận hành câu hỏi, giao diện sử dụng.

Kết luận: Kết quả nghiên cứu ghi nhận tính hiệu quả và hiệu suất của thuật toán UMPItemBank cũng như một số đề xuất cải tiến chất lượng ngân hàng câu hỏi.

Từ khóa: ngân hàng câu hỏi; đánh giá quá trình; trắc nghiệm thích nghi năng lực (CAT); mô hình Rasch; lý thuyết đáp ứng câu hỏi (IRT); giải thuật di truyền; hồi quy hàm số

Ngày nhận bài: 30-09-2024 / Ngày chấp nhận đăng bài: 11-11-2024 / Ngày đăng bài: 13-11-2024
*Tác giả liên hệ: Nguyễn Anh Vũ. Đại học Y Dược Thành phố Hồ Chí Minh, Thành phố Hồ Chí Minh, Việt Nam. E-mail: nguyenanhvu@ump.edu.vn

Abstract

APPLICATION OF RASCH MODEL IN DEVELOPING COMPUTERIZED ADAPTIVE TESTING ITEM BANK

Vinh Son, Tran Thi Dieu, Hoang Dao Bao Tram, Pham Duong Uyen Binh, Nguyen Khanh Chi, Pham Le An, Nguyen Anh Vu

Background: The item bank and adaptive testing algorithm are the two most important components of a Computerized Adaptive Testing (CAT) system. The measurement quality of adaptive testing depends on the quality of the calibrated set of test items included in the assessment. This study focuses on evaluating the functionality and effectiveness of the UMPItemBank to improve the UMPCAT adaptive testing system.

Objective: To evaluate the functionality and operational efficiency of the UMPItemBank software, which utilizes the Rasch model to develop a CAT item bank for assessing initial knowledge levels of learners.

Methods: UMPItemBank software is primarily designed for question management and test generation. It is developed using PHP programming language and MySQL database, operating on Windows 11 with hybrid cloud storage. Security is ensured through OAuth 2.0 authentication and authorization, and WAF firewall. All item content were developed based on constructivist learning theory, undergo a rigorous review and pilot testing process, integrated with the Rasch model for CAT. Processing formative evaluation and case study methods are employed to provide continuous feedback in real-world contexts for software research and development. The data were managed and analyzed using Excel MS 365, JASP 0.19.1, and Heuristic Lab Optimizer 3.3.16.1786.

Results: The study indicated that the item bank had operated as designed, effective and compatible with the CAT algorithm, providing evidence of content validity, bank size and format variety, fairness, cost-effectiveness, feasibility, and potential applications in testing management and computer adaptive testing. However, there were some issues to be addressed, such as threshold control of standard error of measurement, item exposure rate, differential item functioning analysis, and user interface.

Conclusion: The research results indicate the effectiveness and efficiency of the UMPItemBank algorithm as well as some suggestions for improving the quality of the item bank.

Keywords: item bank; formative evaluation; computer adaptive testing; Rasch model; item response theory; genetic algorithm; symbolic regression

1. ĐẶT VẤN ĐỀ

Kỹ thuật đo lường giáo dục học kết hợp với công nghệ thông tin đang góp phần quan trọng đảm bảo chất lượng đào tạo nói chung và chất lượng khảo thí nói riêng, đặc biệt là trong việc nghiên cứu, phát triển và triển khai vận hành các hệ thống kiểm tra trắc nghiệm [1]. Tại Đại học Y Dược Thành phố Hồ Chí Minh, hình thức trắc nghiệm truyền thống đã được chuyển đổi sang trắc nghiệm trên máy tính (CBT), trong đó trắc nghiệm thích nghi (CAT) đã được triển khai thử nghiệm nhiều năm trên đối tượng sinh viên đại học với hệ thống phần mềm mạng UMPCAT.

Trắc nghiệm thích nghi (CAT) là một dạng CBT được xây dựng dựa trên nguyên tắc điều chỉnh linh hoạt độ khó của các câu hỏi theo thời gian thực dựa trên kết quả trả lời của từng thí sinh để cá nhân hóa việc tạo lập, cấp phát và quản lý các bài kiểm tra đánh giá kết quả học tập. CAT phân phát cho mỗi thí sinh một bộ câu hỏi riêng phù hợp với mức trình độ kiến thức. Khi hai thí sinh có cùng số câu trả lời đúng thì thí sinh trả lời đúng nhiều câu khó hơn có đánh giá cao hơn [2,3]. Trắc nghiệm thích nghi bắt đầu từ một câu hỏi có độ khó trung bình sau đó dựa trên kết quả trả lời câu hỏi này để ước lượng của thí sinh, điều chỉnh độ khó của câu hỏi tiếp và lựa chọn từ ngân hàng câu hỏi. Quá trình được tiếp tục lặp lại, mức năng lực người học được ước lượng cập nhật sau mỗi bước, quá

trình dừng lại khi sai số ước lượng đủ mức chính xác cần thiết hoặc người học trả lời hết số câu hỏi giới hạn [4].

CAT được nghiên cứu từ 1970 và được sử dụng lần đầu năm 1985, trở nên phổ biến từ những năm 1990 trong nhiều lĩnh vực như giáo dục, tuyển dụng, đánh giá tâm lý và sức khỏe [5-7], có đủ tính năng giúp giải quyết các khuyết điểm của trắc nghiệm cố định [8,9]. Tại Việt Nam, CAT được triển khai ứng dụng tại một số cơ sở từ những năm 2000 và được ghi nhận là một phương pháp đánh giá hữu hiệu [10,11]. Tuy nhiên, việc tổ chức thực hiện trên thực tế đòi hỏi sự đầu tư nguồn lực và công nghệ lâu dài để có thể phát triển và khai thác hiệu quả [12].

Ngân hàng câu hỏi và thuật toán trắc nghiệm thích nghi là hai thành phần quan trọng nhất, cần được xây dựng và phát triển đồng bộ để hoạt động tương hỗ chặt chẽ để hệ thống CAT vận hành lượng giá hiệu quả, cá nhân hóa và công bằng. Chất lượng đo lường của CAT phụ thuộc vào chất lượng của tập hợp các câu hỏi trắc nghiệm đã được hiệu chuẩn được đưa vào thực hiện lượng giá. Mô hình Rasch giúp việc đo lường thích nghi vững bền và linh hoạt hơn [13]. Mức độ đạt kết quả học của người học có thể được so sánh với nhau ngay cả khi mỗi người học nhận được một bộ câu hỏi duy nhất, thông qua phương pháp cân bằng đề thi để liên kết các bộ câu hỏi và hiệu chuẩn cân bằng số đo logit của các câu hỏi trên một thang đo duy nhất [14].

UMPIteamBank là một thành phần của hệ thống UMPCAT, được thiết kế như một ngân hàng câu hỏi để hỗ trợ lượng giá trắc nghiệm thích nghi. Các chức năng chính của UMPItemBank gồm có: lưu trữ và quản lý câu hỏi trắc nghiệm theo các tiêu chí tổ chức và đo lường khác nhau; hỗ trợ người dùng xác định nội dung, chuẩn đầu ra học phần và tạo ma trận liên kết giữa nội dung và chuẩn đầu ra; hỗ trợ tạo đề thi trắc nghiệm tự động hoặc thủ công từ ngân hàng câu hỏi phù hợp với mục tiêu đánh giá và tùy chọn của người dùng; phân tích thống kê đánh giá chất lượng câu hỏi và hiệu quả của đề thi.

Nghiên cứu này tập trung đánh giá hiệu quả ngân hàng câu hỏi UMPItemBank, từ quy trình kiến tạo nội dung, kiến trúc hệ thống, nguyên tắc hoạt động đến hiệu quả vận hành thực tế thông qua trường hợp ứng dụng UMPCAT đánh giá năng lực đầu vào học phần Tin Học Ứng Dụng, nhằm cải tiến hệ thống trắc nghiệm thích nghi UMPCAT.

2. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Đối tượng nghiên cứu

Phần mềm ngân hàng câu hỏi UMPItemBank, cấu trúc và khả năng đáp ứng vận hành hệ thống phần mềm trắc nghiệm thích nghi UMPCAT. Bao gồm các đơn nguyên: trình quản lý ngân hàng câu hỏi, thư viện thông tin câu hỏi, điều hướng và liên kết nội tuyến. Cấu hình máy tính chạy UMPCAT gồm CPU Intel Core i5/i7, AMD Ryzen 5/7; RAM 16GB; hệ điều hành Window 11, dung lượng bộ nhớ SSD trống 250 GB, trình duyệt Chrome.

2.1.1. Kiến tạo nội dung ngân hàng câu hỏi

Lý thuyết học tập được sử dụng là thuyết kiến tạo, kiến thức mục tiêu là kiến thức cấu trúc về mạng lưới quan hệ giữa các khái niệm lý thuyết và thao tác trên máy tính. Quy trình biên soạn câu hỏi gồm 05 giai đoạn: định nghĩa mục tiêu học tập, mô tả đặc trưng nội dung học tập, biên soạn câu hỏi, rà soát câu hỏi, tổ chức và vận hành đề thi, đánh giá lại độ phù hợp của câu hỏi và mục tiêu học tập. Bộ câu hỏi sau biên soạn được tổ chức phân loại và lưu trữ theo thư viện, chỉ được đưa vào ngân hàng câu hỏi sau khi được chuẩn hóa tham số theo mô hình Rasch và IRT bằng mô phỏng kết hợp kiểm thử thực địa. Sau khi hình thành nội dung, ngân hàng câu hỏi được kết nối thuật toán thích nghi để thử nghiệm chức năng. Kết quả thử nghiệm được phân tích lại để xác định đặc trưng câu hỏi phù hợp với mô hình đo lường và những điểm cần cải tiến. Quy trình biên soạn câu hỏi được thực hiện định kỳ sau mỗi học kỳ hoặc sau mỗi kỳ thi tùy mục đích sử dụng.

Kiến trúc hệ thống

Các thành phần cốt lõi của UMPItemBank bao gồm ngôn ngữ lập trình PHP, cơ sở dữ liệu MySQL, lưu trữ đám mây kết hợp, bảo mật xác thực và ủy quyền OAuth 2.0 và WAF, mô hình đo lường Rasch nhị phân với thang đo logit sử dụng tham số độ khó câu hỏi (Bảng 1).

Bảng 1. Đặc điểm của phần mềm UMPItemBank

Thành phần phần mềm	Thông tin
Ngôn ngữ lập trình	PHP
Cơ sở dữ liệu	MySQL
Mô hình đo lường	Rasch nhị phân, sử dụng một tham số độ khó logit.

https://doi.org/10.32895/hcjm.m.2024.05.11

Thành phần phần mềm	Thông tin
Ngân hàng câu hỏi	Tin Học Ứng Dụng, 654 câu, chuẩn độ logit theo Rasch.
Bảo mật	OAuth 2.0 và WAF
Thư viện câu hỏi	Nội dung câu hỏi: trắc nghiệm 4 lựa chọn A-B-C-D; Dữ liệu câu hỏi: mã số, phân loại, tham số đo lường, mức độ đo theo thang Bloom, từ khóa, chuẩn đầu ra có liên quan; Lịch sử câu hỏi: biên soạn, cập nhật, sử dụng.
Thư viện liên kết	Quan hệ logic giữa các câu hỏi Phân loại các câu hỏi theo CLO, phân vùng kiến thức, thang Bloom, phiên bản câu hỏi đôi
Quản lý ngân hàng	Quyền truy cập có phân cấp; Rà soát và phê duyệt sử dụng; Phân tích câu hỏi và báo cáo
Tích hợp hệ thống CAT	Kết nối dữ liệu động metadata; Kiểm soát mức độ phơi lộ câu hỏi; Cập nhật động thư viện câu hỏi

2.2. Phương pháp nghiên cứu

2.2.1. Thiết kế nghiên cứu

Nghiên cứu trường hợp và đánh giá quá trình. Trong đó có mô tả cắt ngang tại thời điểm đánh giá, đánh giá tính nhất quán và tính đáp ứng mục tiêu trên từng câu hỏi cũng như trên đề thi, đo lường hiệu chuẩn câu hỏi thông qua kiểm nghiệm thực địa.

2.2.2. Biến số nghiên cứu

Độ khó câu hỏi theo Rasch (logit), độ khó câu hỏi theo trắc nghiệm cổ điển (CTT).

Độ dài bài kiểm tra (câu).

Kích cỡ ngân hàng câu hỏi (câu).

Thời gian làm bài của thí sinh (phút).

Thời gian phân tích đề (giây).

Thời gian truy xuất câu từ cơ sở dữ liệu (giây).

2.2.3. *Địa điểm nghiên cứu*

Nghiên cứu được thực hiện tại Đại học Y Dược Thành phố Hồ Chí Minh từ tháng 12 năm 2023 đến tháng 9 năm 2024.

2.2.4. Nhân lực tham gia

Giảng viên Bộ môn Tin học và Bộ môn Toán, gồm 01 giảng viên điều phối, 01 giảng viên quản trị mạng, 05 giảng viên hỗ trợ.

2.2.5. *Triển khai vận hành hệ thống*

Cấu trúc hệ thống đã được triển khai trên nền tảng web, được phát triển bằng ngôn ngữ lập trình PHP với nhiều ưu điểm như chạy nhẹ, linh hoạt và đa nền tảng, có thể được sử dụng cho nhiều hệ điều hành khác nhau. Trường hợp nghiên cứu là ngân hàng câu hỏi học phần Tin Học Ứng Dụng. Sau khi được đánh giá cấu trúc và nội dung, thử nghiệm chức năng được thực hiện trên máy tính, thuật toán chọn lựa câu hỏi được sử dụng với các tham số đề thi số câu hỏi là 45, sai số chuẩn SE của ước lượng năng lực thí sinh là 0,3 logit, thời gian làm bài 60 phút. Các tham số được xác định dựa vào kết quả mô phỏng năng lực thí sinh với phân phối chuẩn trong khoảng -3,0 – 3,0 logit, thời gian làm bài được ước tính dựa vào kinh nghiệm của giảng viên phụ trách học phần.

2.2.6. Quy trình nghiên cứu

Đánh giá quá trình theo Tessmer M (1993) gồm 4 giai đoạn [15]: đánh giá chuyên gia, đánh giá cá nhân, đánh giá nhóm nhỏ, đánh giá thực nghiệm thực địa. Ngân hàng câu hỏi được đánh giá cấu trúc và đánh giá chức năng thông qua vận hành trong hệ thống UMPCAT. Trường hợp nghiên cứu được thực hiện với ngân hàng câu hỏi của học phần Tin Học Ứng Dụng.

- Đánh giá chuyên gia được thực hiện chuyên gia đo lường giáo dục học và giáo dục học, chuyên gia nội dung lĩnh vực, chuyên gia đảm bảo chất lượng phần mềm. Chuyên gia hiểu rõ về bản chất và mức độ tác động của nghiên cứu, tự nguyện tham gia và nhận được đầy đủ thông tin về nghiên cứu, quyền của người tham gia và bảo mật thông tin cá nhân. Nhóm chuyên gia thực hiện thiết kế mô phỏng kiểm thử, xác định tải trọng người dùng và nút nghẽn luồng, theo dõi quá trình và kết quả kiểm tra thực địa, đánh giá quy trình phát triển, cấu hình và phiên bản, giao diện người dùng, kiểm tra thâm nhập và lỗi xác thực danh tính. Chuyên gia đề ra khuyến nghị cải thiện toàn diện hệ thống UMPCAT trong đó có quản lý và truy xuất cơ sở dữ liệu UMPItemBank. Phương pháp phỏng vấn bán cấu trúc được sử dụng khi tiếp xúc với chuyên gia.

- Đánh giá cá nhân kết hợp nhóm nhỏ được thực hiện với nhóm sinh viên tự nguyện. Sinh viên tham gia đánh giá hiểu rõ về bản chất và mức độ tác động của nghiên cứu, tự nguyện tham gia và nhận được đầy đủ thông tin về nghiên cứu, quyền của người tham gia và bảo mật thông tin cá nhân. Ngân hàng câu hỏi Tin Học Ứng Dụng được sử dụng. Đối tượng được

giới thiệu về mục đích và cách hoạt động của phần mềm UMPCAT, sau đó được trải nghiệm cá nhân với một số câu hỏi. Sau khi được trải nghiệm cá nhân, sinh viên được làm thử một bài kiểm tra thích nghi. Mỗi sinh viên mô tả trải nghiệm của mình, chia sẻ quan điểm cá nhân về ưu điểm khuyết điểm của các câu hỏi được sử dụng. Sau khi có kết quả đánh giá, sinh viên được phỏng vấn bán cấu trúc về trải nghiệm sử dụng, bao gồm cảm nhận chung về đề kiểm tra, đặc trưng các câu hỏi được sử dụng gồm cảm nhận chủ quan về độ phức tạp và tính liên quan với nội dung học tập, tính hợp lý của chuỗi câu hỏi, tính rõ ràng mạch lạc, tính thực tiễn, tính sáng tạo mức độ thách thức, mức tư duy cần thiết, mức độ căng thẳng. Sinh viên cũng được khuyến khích đưa ra ý kiến nên phát triển bộ câu hỏi theo hướng nào.

- Đánh giá thực nghiệm thực địa được thực hiện trong môi trường thực tế giảng dạy, dựa trên đánh giá mức kiến thức đầu vào của sinh viên đăng ký học phần Tin Học Ứng Dụng. Quy trình gồm 6 bước: (1) nhập câu hỏi trắc nghiệm vào phần mềm; (2) xác định tham số thích nghi của đề kiểm tra bao gồm số câu hỏi giới hạn, thời gian làm bài giới hạn, độ khó câu hỏi khởi đầu; (3) khởi động thuật toán lựa chọn câu hỏi thích nghi của phần mềm UMPCAT; (4) quản lý vận hành phần mềm trong thời gian trắc nghiệm; (5) rà soát các trường hợp quá hạn tham số đề; (6) thu thập và phân tích số liệu.

2.2.7. Quy mô mẫu và lấy mẫu

Mẫu toàn thể: 1308 câu hỏi trắc nghiệm tham chiếu chuẩn mực (norm-referenced) với 4 lựa chọn trong ngân hàng câu hỏi Tin Học Ứng Dụng.

Mẫu ngẫu nghiên phân tầng tỷ lệ xác suất (Proportionate Stratified Random Sampling): 654 câu trích xuất từ ngân hàng câu hỏi Tin Học Ứng Dụng.

Dành cho đánh giá chuyên gia: lấy mẫu có mục đích 05 chuyên gia.

Dành cho đánh giá cá nhân và nhóm nhỏ: mẫu tiện lợi có mục đích 05 sinh viên.

Dành cho đánh giá thực nghiệm thực địa: mẫu tiện lợi 99 sinh viên.

Tiêu chuẩn chọn vào: sinh viên đăng ký học phần Tin Học Ứng Dụng.

Tiêu chuẩn loại ra: sinh viên học lại; hoặc đã có chứng chỉ tin học ứng dụng trình độ B trở lên, chứng chỉ quốc tế ICDL, Cisco, IC3, MOS.

2.2.8. Phương pháp thống kê

Số liệu mô phỏng và số liệu trắc nghiệm được quản lý bằng Excel MS 365, mô tả và kiểm định thống kê bằng R 4.3.1 có phối kiểm lại bằng JASP 0.19.1.0, hồi quy hàm số bằng Heuristic Lab Optimizer 3.3.16.1786.

Giải thuật di truyền Offspring Selection được dùng với tập huấn luyện và tập kiểm tra chứa 64% và 34% dữ liệu.

Biến định tính được mô tả bằng tần số và tỷ lệ phần trăm, so sánh bằng kiểm định $\chi 2$ hoặc Fisher nếu tần số kỳ vọng dưới 5. Biến định lượng được mô tả bằng giá trị trung bình và độ lệch chuẩn, trung vị và khoảng phân vị, kiểm định Shapiro-Wilk kiểm tra phân phối chuẩn, kiểm định Student so sánh trung bình khi có phân phối chuẩn, kiểm định Mann-Whitney so sánh trung vị khi phân phối không chuẩn.

Khác biệt có ý nghĩa thống kê khi $p < 0,05$.

3. KẾT QUẢ

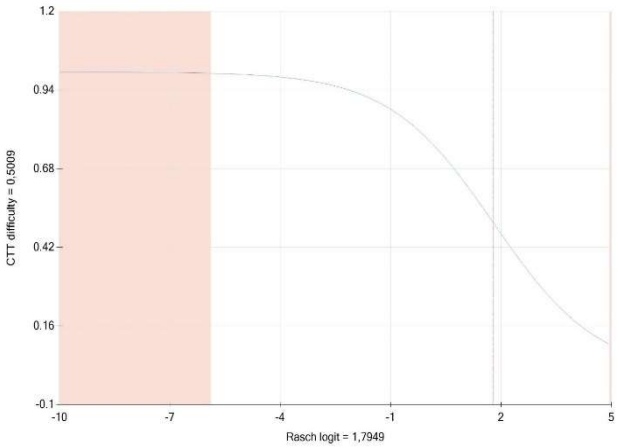
3.1. Trường hợp nghiên cứu

Ngân hàng câu hỏi Tin Học Ứng Dụng gồm 1308 câu hỏi trắc nghiệm tham chiếu tiêu chuẩn với 4 lựa chọn, được chọn lọc từ thư viện câu hỏi 3538 câu, sau quá trình biên tập chỉnh đốn liên tục từ năm 2015 đến thời điểm thực hiện nghiên cứu. Các câu hỏi này được tổ chức theo chủ đề kiến thức và chuẩn đầu ra học phần, dùng để đánh giá mức kiến thức đầu vào của người học (Bảng 2).

Bảng 2. Độ khó của câu hỏi ngân hàng câu hỏi Tin Học Ứng Dụng 1308 câu

Tham số	Trung bình	Độ lệch chuẩn	Mode	Min	Max	Trung vị	Q25	Q75
Độ khó theo Rasch (logit)	-0,0909	1,4535	-2,8163	-5,8815	4,950	-0,1604	-1,0409	0,8025
Độ khó cổ điển	0,7518	0,1700	0,962	0,0993	0,9954	0,7977	0,6678	0,8796

https://doi.org/10.32895/hcjm.m.2024.05.11



Hình 1. Mối quan hệ giữa độ khó cổ điển và độ khó logit theo Rasch trong ngân hàng câu hỏi

Mô hình hồi quy hàm số quan hệ giữa độ khó logit và độ khó cổ điển cho thấy kết quả có dạng xấp xỉ hàm logistic khá rõ rệt, độ khó CTT đạt 0,5 tại giá trị logit 1,79 (ảnh hưởng biến số 0,995, training và test R-squared 0,999, training RMSE 0,0000139 và test RMSE 0,0000126), độ chính xác cao hơn so với mô hình tuyến tính (training R-squared 0,902, test R-squared 0,993, training RMSE 0,00261 và test RMSE 0,00234) (Hình 1).

Mối quan hệ giữa độ khó cổ điển và logit cho thấy tính đơn chiều của bộ câu hỏi. Những câu dễ và rất dễ có vị trí rất sát nhau trên thang đo độ khó cổ điển, nhưng có độ khó logit phân phối đều từ – 6,6 logit đến -10 logit cho thấy đo lường Rasch phân biệt tốt hơn ở những câu hỏi này. Quan hệ hàm số giữa hai loại độ khó trên cũng cho thấy sự tương đồng giữa đo lường trắc nghiệm cổ điển và trắc nghiệm Rasch. Việc thực hiện chỉnh đốn chọn lọc bộ câu hỏi theo trắc nghiệm cổ điển trong một thời gian đủ lâu dài sẽ đạt kết quả giống như sử dụng mô hình Rasch.

3.2. Đánh giá chuyên gia

Về vận hành, phần mềm chạy trên nền tảng mạng, gọn nhẹ, dễ sử dụng và quản lý. Về giao diện người dùng, bố cục trực quan đơn giản, các yếu tố quan trọng được sắp xếp dễ sử dụng và dễ theo dõi, giao diện có phản hồi nhanh và dễ nhận biết. Thiết kế giao diện phù hợp với mục đích sử dụng phần mềm và nhu cầu khảo thí. Tuy nhiên cần chú ý yếu tố thẩm mỹ và sử dụng màu sắc thu hút thị giác của người dùng. Cấu trúc giao diện cố định, chưa có chức năng tùy chỉnh, do đó tính tùy biến đáp ứng người dùng chưa cao.

Về hiệu suất, phần mềm có khả năng chịu tải lượng người dùng ổn định, tuy nhiên có khoảng cách truy xuất không đều, một số trường hợp có mức kiến thức xấp xỉ điểm cực đại thông tin nhưng thời gian ước lượng gần mức cao nhất. Điều cần thiết là cải thiện cả thuật toán chọn câu hỏi, thuật toán ước lượng và truy vấn cơ sở dữ liệu. Về tính năng, phần mềm đáp ứng được các kịch bản sử dụng khác nhau, có tương thích với thuật toán lựa chọn câu hỏi trong hệ thống UMPCAT, kết quả ước lượng mức kiến thức của sinh viên đủ chính xác. Độ chính xác tính toán 15 chữ số thập phân, báo cáo kết quả với 5 chữ số thập phân. Thời gian truy xuất cơ sở dữ liệu xấp xỉ 0,02 giây mỗi câu. Kịch bản tối ưu là 120 câu hỏi và 300 thí sinh, thời gian phân tích kết quả khoảng 19-28 giây theo CTT và khoảng 95-120 giây theo Rasch.

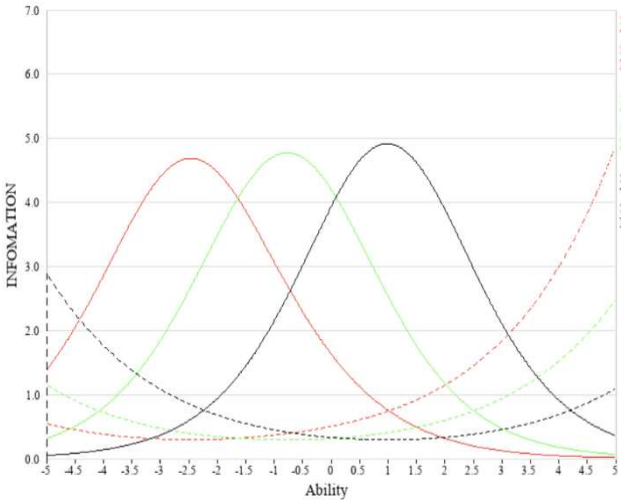
Tuy nhiên, với ngưỡng sai số chuẩn SE 0,3 hàm thông tin câu hỏi không đạt đỉnh tại mức trình độ thí sinh, cho thấy ước lượng trình độ thí sinh chưa được đảm bảo mức sai số tối ưu. Với ngưỡng SE 0,2 có thể cải thiện vấn đề, tuy nhiên số câu hỏi kiểm tra sẽ tăng lên, đặt ra vấn đề về kích cỡ ngân hàng câu hỏi khi có số lượng thí sinh trên 1000. Sau khi phát hiện câu hỏi có tần suất chọn lặp cao, cần có biện pháp nhận diện và ngăn chặn lựa chọn lặp lại, giúp đảm bảo tính bảo mật của ngân hàng câu hỏi. Đồng thời, cần chú ý phân tích độ sai biệt vận hành của câu hỏi hoặc nhóm câu hỏi đối với các nhóm thí sinh khác nhau để đảm bảo tính công bằng khảo thí. Trong quá trình kiến tạo nội dung ngân hàng câu hỏi, cần có hệ số định lượng độ phù hợp của câu hỏi với mục tiêu để hỗ trợ người biên soạn. Chức năng phân tích và báo cáo kết quả chi tiết về mức năng lực ước tính, độ khó câu hỏi, độ phân cách, hàm thông tin câu hỏi, hàm sai số ước lượng, trực quan hóa dữ liệu. Tuy nhiên hỗ trợ xuất dữ liệu ở các định dạng khác nhau chưa được thiết kế.

Về tổng thể, UMPItemBank vận hành đúng chức năng được thiết kế. Dạng câu hỏi trắc nghiệm đồng nhất không yêu cầu thao tác phản hồi phức tạp. Phần mềm có tiềm năng ứng dụng rất cao trong cả quản lý khảo thí và đánh giá thành quả học tập theo trắc nghiệm thích nghi. Tuy nhiên để bảo đảm UMPItemBank vận hành tốt trong môi trường đánh giá thích nghi của UMPCAT cần chú ý kiểm soát khả năng phơi lộ, ngưỡng sai số chuẩn và kích cỡ ngân hàng. Để nâng cao tính cá nhân hóa, cải thiện hiệu suất với số liệu lớn, và đặc biệt là hỗ trợ biên soạn câu hỏi mới cũng như chỉnh đốn các câu hỏi chưa đạt tính giá trị nội dung hoặc đo lường, nhóm phát triển phần mềm nên xem xét việc tích hợp các tính năng học máy

vào UMPItemBank.

3.3. Đánh giá kết hợp cá nhân và nhóm nhỏ

Trong nghiên cứu này, đánh giá cá nhân được kết hợp với đánh giá nhóm nhỏ. Về ưu điểm, sinh viên tham gia đánh giá hài lòng với mức độ thực tế, rõ ràng, chính xác và cụ thể của câu hỏi, mức độ phức tạp và tính thách thức không quá cao nhưng cũng không hoàn toàn đơn giản, thời gian hiển thị câu hỏi khá nhanh và không gián đoạn, độ dài bài kiểm tra và thời gian làm bài đều giảm đi rõ rệt. Về những hạn chế, sinh viên cho biết giao diện người dùng thiếu yếu tố thẩm mỹ đồ họa. Tính tương tác với thí sinh chưa cao, đánh giá kết quả trả lời mỗi câu hỏi cũng như nội dung câu hỏi tiếp theo không hiển thị ngay sau khi thí sinh đã hoàn thành trả lời mỗi câu hỏi. Các câu trong chuỗi câu hỏi có mối liên quan khá chặt chẽ, bao phủ kiến thức khá đầy đủ nhưng chủ đề chưa được sắp xếp thống nhất.



Hình 2. Thông tin ước lượng thích nghi mức kiến thức thí sinh với CATItemBank

Hình 2 là một biểu đồ minh họa điển hình thông tin ước lượng mức kiến thức thí sinh dựa trên ngân hàng câu hỏi Tin Học Ứng Dụng, quản lý nội dung và truy xuất câu hỏi bằng phần mềm CATItemBank, trích xuất dữ liệu của ba sinh viên có các mức kiến thức thấp (θ = - 2,56), trung bình (θ = -0,33) và trên trung bình (θ = 0,68).

Cả ba bài kiểm tra đánh giá đều có hàm thông tin phân biệt nhau, cho thấy kết quả đo lường của các thí sinh độc lập nhau, mỗi thí sinh có bài kiểm tra riêng biệt phù hợp mức năng lực của mình. Hàm thông tin đạt đỉnh tại mức giá trị xấp xỉ với mức kiến thức của sinh viên, cho thấy ước lượng mức kiến thức của mỗi thí sinh đều đạt độ chính xác cao, sai số ước lượng đủ nhỏ tuy chưa đạt được mức nhỏ nhất.

3.4. Đánh giá thực địa

Đánh giá thực địa sử dụng 654 câu hỏi trong ngân hàng để nạp vào hệ thống UMPCAT, số lượng câu hỏi như vậy cao hơn so với số câu hỏi đã thu được bằng mô phỏng. Điều này nhằm thu được tham số độ khó câu hỏi của mẫu tương đồng với toàn bộ ngân hàng câu hỏi, đồng thời tránh hiện tượng chọn lặp câu hỏi có thể xảy ra, cũng như đảm bảo tính liên tục của dải độ khó câu hỏi (Bảng 3).

Độ khó cổ điển có phân phối lệch trái rõ rệt, cho thấy tỷ lệ trả lời đúng tính trên mẫu rất lớn các thí sinh là khá cao. Độ khó logit có phân phối đối xứng khá rõ và nhọn so với phân phối chuẩn, cho thấy so với phân phối chuẩn thì tỷ lệ thấp hơn ở các câu hỏi đánh giá mức trình độ kiến thức trung bình trong khi tỷ lệ câu hỏi nhằm đến mức trình độ kiến thức cao và thấp lại cao hơn ở hai biên. Đặc điểm này phù hợp với đánh giá mức trình độ kiến thức đầu vào của sinh viên mới đăng ký học phần lần đầu.

Bảng 3. Độ khó của câu hỏi trong mẫu 654 câu hỏi Tin Học Ứng Dụng

Tham số	Trung bình ± ĐLC	Min	Max	Trung vị (Q25 - Q75)	Shapiro Wilk	P
Độ khó theo Rasch (logit)	-0,091 ± 1,454	-5,881	4,950	-0,16 (-1,04; 0,80)	0,993	< 0,001
Độ khó cổ điển	0,75 ± 0,17	0,1	0,99	0,79 (0,67; 0,88)	0,905	< 0,001

Kiểm định Shapiro-Wilk cho phân phối chuẩn

Bảng 4. Mức trình độ kiến thức thí sinh, độ dài bài thi và thời gian làm bài

Biến số	Chung (n = 86)	Nam (n = 54)	Nữ (n = 31)	P
Năng lực thí sinh (logit) Trung bình ± ĐLC Min – Max	 -0,743 ± 1,002 -3,534 – 1,411	 -0,766 ± 1,081 -3,534; 1,411	 -0,704 ± 0,865 -2,358; 1,038	 0,769[a] -

https://doi.org/10.32895/hcjm.m.2024.05.11

Biến số	Chung (n = 86)	Nam (n = 54)	Nữ (n = 31)	P
Trung vị	-0,758	-0,698	-0,797	-
Q25 – Q75	-1,273; -0,039	-1,27; -0,039	-1,204; -0,133	-
Độ dài bài kiểm tra				
Trung bình ± ĐLC	44,68 ± 4,96	44,80 ± 5,40	44,60 ± 4,3	0,606[b]
Min – Max	20; 58	20; 58	20; 58	-
Trung vị	45	44	45	-
Q25 – Q75	43 – 47	43 – 47	43 – 47	-
Thời gian làm bài (phút)				
Trung bình ± ĐLC	28,8 ± 8,8	28,00 ± 8,8	30,1 ± 8,6	0,236[a]
Min – Max	9,7 - 50	9,7 – 50	14,0 – 45,6	-
Trung vị	28,4	28,1	28,7	-
Q25 – Q75	23,0; 34,7	22,3; 33,6	24,2; 37,9	-

So sánh trung bình bằng: a Kiểm định Student; b Kiểm định Mann-Whitney

Mức trình độ kiến thức thí sinh có phân phối chuẩn (p=0,592 đối với nữ, p=0,284 đối với nam), số đo năng lực biến động trong phạm vi từ -3,5 đến 1,4 logits. Độ dài bài kiểm tra trung bình 45 câu, biến động trong phạm vi từ 20 câu đến 58 câu. Thời gian làm bài có phân phối chuẩn, với trung bình 28,7 phút, thời gian ngắn nhất là 9,7 phút và dài nhất là 50 phút. Mức trình độ kiến thức đầu vào, số câu hỏi được kiểm tra và thời gian làm bài không khác biệt có ý nghĩa thống kê giữa hai nhóm giới tính (Bảng 4).

4. BÀN LUẬN

Kết quả nghiên cứu cho thấy bằng chứng quan trọng đảm bảo chất lượng và độ tin cậy vận hành phần mềm UMPItemBank. Các bài kiểm tra thể hiện đặc điểm cá nhân hóa rõ rệt, kết quả ước lượng mức kiến thức thí sinh có tính độc lập không có sự ảnh hưởng lẫn nhau. Mô hình đo lường Rasch vừa có tính bền vững vừa tương đối đơn giản để triển khai trong phần mềm máy tính. 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. Những đặc điểm này phù hợp với việc thiết kế và phát triển đo lường khách quan và ứng dụng thuật toán trắc nghiệm thích nghi UMPCAT.

Tính khả dụng và tương thích hệ thống cao cho phép phần mềm dễ vận hành, không cạnh tranh với các phần mềm khác. Phần mềm có các chức năng theo dõi đánh giá độ phù hợp của câu hỏi với mô hình đo lường, phạm vi và phân phối độ khó câu hỏi, khả năng phân cách của câu hỏi, tuy tính công bằng của câu mới chỉ phân tích sai biệt vận hành câu hỏi đối với nhóm giới tính, chưa được phân tích trên các yếu tố khác. Chức năng lưu trữ và truy xuất câu hỏi hoạt động hiệu quả và tương thích với thuật toán lựa chọn câu hỏi. Chức năng chỉnh sửa và cập nhật câu hỏi mới dễ sử dụng đề câu, tuy nhiên hỗ trợ nhập và xuất dữ liệu câu hỏi từ các định dạng file khác nhau chưa được quan tâm.

Trên nền tảng tương tác với ngân hàng câu hỏi, thuật toán lựa chọn câu hỏi hoạt động hiệu quả, cho phép ước tính mức kiến thức thí sinh với độ chính xác cao tuy chưa đạt mức tối ưu, số lượng câu hỏi cần thiết để đạt độ chính xác cần thiết giảm đi rõ rệt. So với một bài kiểm tra CBT thông thường gồm từ 100 đến 120 câu hỏi, và thường có SE khoảng từ 0,5 đến 0,7, bài kiểm tra thích nghi CAT giảm thời gian làm bài trung bình 20 phút, chỉ còn 40 đến 60 câu đủ đạt kết độ chính xác cao hơn với SE 0,3 logit. Kết quả này phù hợp các nghiên cứu của Wainer H (1999), Callear D và King T (1997), Linacre JM (2000) [3,16,17].

Bài kiểm tra CAT thành công trong việc hạn chế đưa ra câu hỏi quá khó hoặc quá dễ đối với từng cá nhân, đặc biệt giảm mức độ căng thẳng tâm lý trong thi cử là một yếu tố gây nhiễu khi đo lường năng lực thực hiện công việc đã học. Tuy nhiên bài kiểm tra thích nghi có thể gây ra những hiệu ứng tâm lý khác liên quan đến trình tự của dãy câu hỏi và không thể quay lại các câu hỏi đã trả lời trước đó, phù hợp với nghiên cứu của Linacre JM (2000) và Colwell NM (2013) [17,18]. Điều này gợi ra một số cải tiến cần thiết về tổ chức ngân hàng câu hỏi phù hợp cho việc mở rộng kiểm tra thích ứng đa giai đoạn, trong đó thí sinh có thể bỏ qua một số câu hỏi cũng như sửa câu trả lời trong phạm vi nhất định.

Bên cạnh đó, nghiên cứu cũng chỉ ra một số hạn chế của ngân hàng câu hỏi, thuật toán lựa chọn câu hỏi, và thuật toán ước lượng. Tính thẩm mỹ, tùy biến của giao diện người dùng

và tính tương tác với người dùng còn thấp. Đây cũng là những hạn chế đã được khảo sát trong nghiên cứu của Economides AA và Roupas C (2007), Bridegeman B (2017) [19,20]. Một số hướng tiếp cận mới có thể giúp cải tiến phương pháp thiết kế thuật toán và giải quyết vấn đề chất lượng ứng dụng CAT [13,21-23].

5. KẾT LUẬN

Nghiên cứu đã ghi nhận bằng chứng về tính giá trị, độ tin cậy và tính tương thích của phần mềm ngân hàng câu hỏi UMPItemBank với thuật toán lựa chọn câu hỏi trong hệ thống trắc nghiệm thích nghi UMPCAT. Tuy nhiên phần mềm UMPItemBank có một số hạn chế hiệu suất hoạt động và tính tương tác và phản hồi cần khắc phục, cụ thể gồm có tính tùy biến, thẩm mỹ và nhất quán trong giao diện người dùng để tạo trải nghiệm liền mạch, kiểm soát và hạn chế tỷ lệ phơi lộ câu hỏi, chức năng phân tích độ phù hợp của câu hỏi với mục tiêu, phân tích sai biệt chức năng câu hỏi và tối ưu hóa sai số ước lượng. Việc tích hợp chức năng trí tuệ nhân tạo cũng được đề xuất để phát triển phần mềm.

Lời cảm ơn

Xin chân thành cảm ơn Đại học Y Dược Thành phố Hồ Chí Minh đã tài trợ cho nghiên cứu này. Tập thể tác giả gửi lời cảm ơn sâu sắc đến Phòng Đảm bảo Chất lượng Giáo dục và Khảo thí, Phòng Khoa học Công nghệ, Khoa Khoa học Cơ bản đã tạo điều kiện thuận lợi, hỗ trợ cho nghiên cứu.

Nguồn tài trợ

Nghiên cứu nhận được kinh phí tài trợ từ Đại học Y Dược Thành phố Hồ Chí Minh theo hợp đồng số 73/2022/HĐ-ĐHYD.

Xung đột lợi ích

Không có xung đột lợi ích tiềm ẩn nào liên quan đến bài viết này được báo cáo.

ORCID

Vĩnh Sơn

https://orcid.org/0009-0005-4864-8601

Trần Thị Diệu

https://orcid.org/0009-0002-2605-9003

Phạm Dương Uyển Bình

https://orcid.org/0000-0003-3398-3210

Nguyễn Khánh Chi

https://orcid.org/0009-0001-2656-8409

Phạm Lê An

https://orcid.org/0000-0003-1186-0543

Nguyễn Anh Vũ

https://orcid.org/0009-0003-7148-8840

Đóng góp của các tác giả

Ý tưởng nghiên cứu: Hoàng Đạo Bảo Trâm, Phạm Lê An

Đề cương và phương pháp nghiên cứu: Phạm Dương Uyển Bình, Nguyễn Khánh Chi, Nguyễn Anh Vũ

Thu thập dữ liệu: Vĩnh Sơn, Trần Thị Diệu, Nguyễn Khánh Chi

Giám sát nghiên cứu: Hoàng Đạo Bảo Trâm, Phạm Lê An

Nhập dữ liệu: Trần Thị Diệu, Nguyễn Khánh Chi

Quản lý dữ liệu: Vĩnh Sơn

Phân tích dữ liệu: Nguyễn Anh Vũ

Viết bản thảo đầu tiên: Nguyễn Anh Vũ

Góp ý bản thảo và đồng ý cho đăng bài: Hoàng Đạo Bảo Trâm, Vĩnh Sơn, Phạm Lê An, Nguyễn Anh Vũ, Phạm Dương Uyển Bình, Trần Thị Diệu, Nguyễn Khánh Chi

Cung cấp dữ liệu và thông tin nghiên cứu

Tác giả liên hệ sẽ cung cấp dữ liệu nếu có yêu cầu từ Ban biên tập.

TÀI LIỆU THAM KHẢO

1. Oakleaf M. Dangers and Opportunities: A Conceptual Map of Information Literacy Assessment Approaches. Libraries and the Academy. 2008;8(3):233–253.
2. Weiss DJ. Adaptive testing by computer. Journal of Consulting and Clinical Psychology, 1985;53(6):774–789.
3. Wainer H, et al. Computerized Adaptive Testing: A Primer. Lawrence Erlbaum Associates.1990.
4. Meijer RR, Nering ML. Computerized Adaptive Testing: Overview and Introduction. Applied Psychological Measurement. 1999;23(3):187–194.

https://doi.org/10.32895/hcjm.m.2024.05.11

5. Drasgow F. The work ahead: A psychometric infrastructure for computerized adaptive tests. In CN Mills, MT Potenza, JJ Fremer & WC Ward (Eds.), Computer-based testing: Building the foundation for future assessments, Hillsdale, NJ: Lawrence Erlbaum, pp.67–88. 2002.
6. Larson JW, Madsen HS. Computer-adaptive language testing: Moving beyond computer-assisted testing. CALICO Journal. 1985;2(3):32-6.
7. Grigoriadou M, Papanikolaou K, Kornilakis H, Magoulas G. INSPIRE: An intelligent system for personalized instruction in a remote environment. Proceedings of 3rd Workshop on Adaptive Hypertext and Hypermedia, Sonthoven, Germany. 2001.
8. Ling G, Attali Y, Finn B, Stone EA. Is a Computerized Adaptive Test More Motivating Than a Fixed-Item Test? Applied Psychological Measurement. 2017;41(7):495–511.
9. Thomson N (2007). A Practitioner's Guide for Variable-length Computerized Classification Testing. Practical Assessment, Research & Evaluation. 2007;12 (1):1-13.
10. Lê Thái Hưng, Trần Thị Hoa, Đặng Thị Mây, Hoàng Lan Hương. 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. 2019;24(12):54-59.
11. Lê Xuân Tài, Đặng Hoài Phương. Xây dựng mô hình trắc nghiệm thích nghi trên cơ sở lý thuyết đáp ứng câu hỏi. Tạp chí Khoa học Đại học Huế, 2015;97(9):1-17.
12. Travitzky R, Meneghetti DDR, Alavarse OM, Catalani EMT (2018). How to build a Computerized Adaptive Test with free software and pedagogical relevance? Proceedings of IAC 2018 in Vienna, Vienna, Austria. 2018.
13. Eggen T. Computerized classification testing with the Rasch model. Educational Research and Evaluation. 2011;5(17):361-371.
14. Boone WJ, Staver RJ. Advances in Rasch analyses in Human Sciences, 306 p. Springer. 2020.
15. Tessmer M. Planning and conducting formative evaluations, p 25-45. Kogan Page. 1993.
16. Callear D, King T. Using computer based tests for information science. ALT-J. 1997;5(1): 27-32.
17. Linacre JM. Computer-adaptive testing: A methodology whose time has come. MESA Memoramdum No.69. In S Chae, U Kang, E Jeon & JM Linacre (Eds.), Development of computerized middle school achievement test, Komesa Press, Seoul, South Korea. 2000.
18. Colwell NM. Test anxiety, computer-adaptive testing and the common core. Journal of Education and Training Studies. 2013;1(2):51-60.
19. Economides AA, Roupas C. Evaluation of computer adaptive testing systems. International Journal of Web Web-Based Learning and Teaching Technologies. 2007;2(1):70-87.
20. Bridgeman B, Lennon ML, Jackenthal A. Effects of Screen Size, Screen Resolution, and Display Rate on Computer-Based Test Performance. Applied Measurement in Education. 2003;16(3):191-205.
21. Delgado-Gomez D, Laria JC, Ruiz-Hernandez D. Computerized adaptive test and decision trees: a unifying approach. Expert Systems with Applications. 2018;117:358-266.
22. Chen SY. Controlling Item Exposure and Test Overlap in Computerized Adaptive Testing. Applied Psychological Measurement. 2005;3(29):204-217.
23. Dang Hoai Phuong, Shabalina OA, Kamaev VA. Adaptive testing algorithm design methods. In Proceedings VSTU, Series "Actual problems of management, computer science and informatics in technical systems", pp.107-113. 2012.