HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



TRẦN THANH QUANG

NGHIÊN CỨU, XÂY DỰNG HỆ THỐNG ĐÁNH GIÁ ĐỀ THI HỖ TRỢ VIỆC CÁ NHÂN HÓA CHO NGƯỜI HỌC TẠI CÁC CƠ SỞ GIÁO DỤC

Chuyên ngành: Hệ thống thông tin
Mã số: 8.48.01.04 (Hệ thống thông tin)

TÓM TẮT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

HÀ NỘI - NĂM 2025

Đề án tốt nghiệp được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: TS. ĐÀO NGỌC PHONG

(Ghi rõ học hàm, học vị)

Phản biện 1:

…….

Phản biện 2:

……

Đề án tốt nghiệp sẽ được bảo vệ trước Hội đồng chấm đề án tốt nghiệp thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ ..... ngày ..... tháng .... năm .....

Có thể tìm hiểu đề án tốt nghiệp tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông.

MỤC LỤC

MỤC LỤC ..... **i**
MỞ ĐẦU **1**
1. Lý do chọn đề tài 1
2. Tổng quan về vấn đề nghiên cứu 1
3. Mục đích nghiên cứu 2
4. Đối tượng và phạm vi đề án 2
5. Phương pháp nghiên cứu 3
CHƯƠNG 1: THỰC TRẠNG VÀ PHÂN TÍCH **4**
1.1. Thực trạng hệ thống đánh giá đề thi 4
1.1.1. Phương pháp truyền thống 4
1.1.2. Các ứng dụng công nghệ hiện có 4
1.1.3. Hiệu quả và hạn chế 4
1.2. Phân tích dữ liệu phản hồi từ đề thi 5
1.2.1. Thu thập và tổng hợp dữ liệu 5
1.2.2. Ứng dụng mô hình Rasch 6
1.2.3. Phân loại câu hỏi 6
1.3. Đánh giá các thách thức hiện tại 7
1.3.1. Thiếu dữ liệu phản hồi đa dạng 7
1.3.2. Hạn chế về hạ tầng công nghệ 7
1.3.3. Đặc thù ngôn ngữ và văn hóa 7
1.4. Kết chương 7
CHƯƠNG 2: CƠ SỞ LÝ LUẬN **8**

2.1. Tổng quan về Lý thuyết Đáp ứng câu hỏi (IRT).... 8
2.1.1. Khái niệm và lịch sử phát triển..... 8
2.1.2. Các mô hình IRT 9
2.2. Mô hình Rasch và ứng dụng trong giáo dục 9
2.2.1. Nguyên tắc hoạt động 9
2.2.2. Ưu điểm và hạn chế 9
2.2.3. Ứng dụng trong giáo dục 10
2.3. Các yếu tố ảnh hưởng đến độ khó câu hỏi 11
2.3.1. Khả năng nhận thức và kinh nghiệm học tập 11
2.3.2. Cấu trúc và nội dung câu hỏi 11
2.3.3. Ngữ cảnh văn hóa và ngôn ngữ 11
2.4. Cơ sở lý luận cho hệ thống học tập cá nhân hoá.. 11
2.4.1. Nghiên cứu quốc tế về học tập cá nhân hóa . 11
2.4.2. Vai trò của công nghệ AI..... 11
2.5. Kết chương..... 11
CHƯƠNG 3: THIẾT KẾ VÀ PHÁT TRIỂN HỆ THỐNG 12
3.1. Thiết kế hệ thống đánh giá đề thi..... 12
3.1.1. Cấu trúc hệ thống..... 12
3.1.2. Xây dựng kho câu hỏi ôn tập..... 13
3.2. Phát triển thuật toán rule-based..... 14
3.3. Thử nghiệm và đánh giá hệ thống..... 15
3.3.1. Phương pháp thử nghiệm..... 15

3.3.2. Chỉ số đánh giá 15
3.3.3. Kịch bản kiểm thử..... 16
3.4. Kết chương..... 19
KẾT LUẬN VÀ KIẾN NGHỊ..... 20
1. Kết luận..... 20
2. Kiến nghị..... 21
DANH MỤC CÁC TÀI LIỆU THAM KHẢO..... 24

MỞ ĐẦU

1. Lý do chọn đề tài

Sự phát triển của AI và phân tích dữ liệu tạo cơ hội đổi mới đánh giá giáo dục, trong khi nhiều cơ sở vẫn dùng đề thi cố định chưa phản ánh đúng năng lực học sinh. IRT, đặc biệt mô hình Rasch, cung cấp nền tảng mô hình hóa năng lực và độ khó câu hỏi cho đánh giá thích ứng nhưng chưa được khai thác hiệu quả. Khoảng cách giữa nhu cầu cá nhân hóa và cách đánh giá truyền thống dẫn tới yêu cầu xây dựng hệ thống dựa trên dữ liệu phản hồi và mô hình đo lường hiện đại. Vì vậy, đề tài lựa chọn ứng dụng Rasch nhằm góp phần đổi mới và thúc đẩy chuyển đổi số trong đánh giá giáo dục.

2. Tổng quan về vấn đề nghiên cứu

IRT, đặc biệt mô hình Rasch 1 tham số, cung cấp nền tảng đo lường giúp mô tả mối quan hệ giữa năng lực người học và độ khó câu hỏi, hỗ trợ hiệu chuẩn câu hỏi và xây dựng thang đo năng lực thống nhất. Dù được ứng dụng rộng rãi trong các hệ thống đánh giá lớn nhờ khả năng chuẩn hóa và phân tích chất lượng câu hỏi, phần mềm học tập hiện nay vẫn ít khai thác dữ liệu phản hồi để ước lượng năng lực theo IRT và cá nhân hóa nội dung.

3. Mục đích nghiên cứu

Đề án xây dựng và kiểm thử hệ thống đánh giá dựa trên mô hình Rasch nhằm chuẩn hóa độ khó câu hỏi và ước lượng năng lực học sinh trên một thang đo thống nhất. Hệ thống kết hợp thuật toán chọn câu theo năng lực cập nhật để tạo lộ trình ôn tập cá nhân hóa và đánh giá sự thay đổi năng lực sau một vòng luyện tập thích ứng. Thử nghiệm được dùng để so sánh với cách ôn tập truyền thống, qua đó xác định tính khả thi và hiệu quả của giải pháp. Đồng thời, đề án cung cấp cơ sở tham khảo cho nghiên cứu và triển khai các hệ thống học tập cá nhân hóa trong bối cảnh chuyển đổi số giáo dục.

4. Đối tượng và phạm vi đề án

Đề án nghiên cứu hệ thống đánh giá và thuật toán phân phối câu hỏi dựa trên mô hình Rasch kết hợp rule-based, tập trung vào học sinh THPT, trong đó nhóm học sinh lớp 10 được chọn làm mẫu thử nghiệm. Phạm vi triển khai giới hạn trong một trường THPT để đảm bảo điều kiện thu thập dữ liệu đồng nhất. Môn Toán được chọn làm lĩnh vực thử nghiệm với ngân hàng 800 câu hỏi đã phân loại chủ đề và mức độ nhận thức, dùng cho cả hiệu chuẩn và chạy thuật toán chọn câu. Nghiên cứu được thực hiện qua các

giai đoạn: xây dựng mô hình Rasch, phát triển thuật toán và thử nghiệm trên người học, tạo nền tảng cho việc mở rộng trong tương lai.

5. Phương pháp nghiên cứu

Đề án sử dụng kết hợp phân tích lý thuyết, mô phỏng và thực nghiệm. Trên cơ sở lý thuyết, nghiên cứu tổng hợp tài liệu về IRT và mô hình Rasch để xây dựng quy trình hiệu chuẩn độ khó và thang đo năng lực, đồng thời tham khảo các tiếp cận về học tập cá nhân hóa. Ở phần thực nghiệm, dữ liệu bài làm của học sinh được dùng để ước lượng tham số Rasch bằng R và triển khai thuật toán rule-based để chọn câu hỏi trong vòng ôn tập thích ứng. Hệ thống được kiểm thử qua hai giai đoạn: ước lượng năng lực ban đầu và thực hiện vòng ôn tập dựa trên năng lực cập nhật. Hiệu quả giải pháp được đánh giá thông qua các chỉ số định lượng ($\Delta\theta$, tỷ lệ cải thiện, độ phủ nội dung…), kiểm định thống kê và phản hồi từ học sinh – giáo viên nhằm xác định mức độ phù hợp và định hướng tối ưu hóa.

CHƯƠNG 1: THỰC TRẠNG VÀ PHÂN TÍCH

1.1. Thực trạng hệ thống đánh giá đề thi

1.1.1. Phương pháp truyền thống

- Đề thi được biên soạn trước, áp dụng đồng thời cho toàn bộ học sinh theo ma trận nội dung thống nhất.

- Ưu điểm: dễ tổ chức, đảm bảo đồng bộ giữa các lớp.

- Hạn chế: độ khó phụ thuộc kinh nghiệm giáo viên; khó kiểm soát tính phân hóa khi số lượng lớp và câu hỏi lớn.

- Kết quả thi chủ yếu dưới dạng điểm tổng, chưa hỗ trợ nhận diện rõ chủ đề học sinh yếu.

- Dữ liệu phản hồi ít được khai thác, điều chỉnh câu hỏi dựa chủ yếu vào kinh nghiệm hơn là bằng chứng thống kê.

1.1.2. Các ứng dụng công nghệ hiện có

1.1.3. Hiệu quả và hạn chế

Hệ thống đánh giá hiện tại vận hành ổn định, hỗ trợ tốt các khâu tổ chức thi, chấm điểm và quản lý kết quả, đặc biệt ở các trường đã triển khai thi trực tuyến. Tuy nhiên, hiệu quả chủ yếu tập trung vào thao tác vận hành, trong khi

các yêu cầu về phân tích dữ liệu sâu và cá nhân hóa vẫn chưa được đáp ứng. Dữ liệu phản hồi chưa được khai thác để nhận diện độ khó, xu hướng sai hay mức độ phân hóa của câu hỏi, khiến việc điều chỉnh đề thi giữa các kỳ vẫn dựa nhiều vào kinh nghiệm. Đồng thời, công nghệ hiện có chưa hỗ trợ việc xây dựng đề hoặc nội dung ôn tập dựa trên năng lực từng học sinh, trong bối cảnh chương trình mới đòi hỏi học tập phân hóa. Những hạn chế này cho thấy dư địa lớn cho các hệ thống đánh giá dựa trên mô hình đo lường hiện đại và phân tích dữ liệu câu hỏi.

1.2. Phân tích dữ liệu phản hồi từ đề thi

1.2.1. Thu thập và tổng hợp dữ liệu

- Dữ liệu thu được qua phiếu trả lời hoặc thi trực tuyến chủ yếu gồm điểm tổng, số câu đúng, tỷ lệ đúng-sai theo câu.

- Dữ liệu phản hồi chi tiết theo từng học sinh thường không được lưu đủ, gây khó khăn cho việc phân tích đặc tính câu hỏi.

- Khi phải tổng hợp thủ công, khối lượng công việc lớn khiến các phân tích nâng cao khó thực hiện thường xuyên.

- Dữ liệu thường lưu theo từng kỳ kiểm tra riêng lẻ, thiếu chuỗi thời gian để theo dõi độ khó câu hỏi và tiến bộ của học sinh.

1.2.2. Ứng dụng mô hình Rasch

Mô hình Rasch có khả năng chuẩn hóa độ khó câu hỏi và ước lượng năng lực học sinh, nhưng việc áp dụng trong trường phổ thông còn hạn chế do yêu cầu dữ liệu chi tiết và công cụ phân tích chuyên dụng. Các hệ thống thi hiện nay thường chỉ thống kê mô tả, không cung cấp tham số độ khó (b) hay chỉ số phù hợp mô hình, khiến giáo viên thiếu cơ sở định lượng để đánh giá chất lượng câu hỏi. Việc triển khai Rasch cũng đòi hỏi sử dụng các phần mềm như R, vốn chưa phổ biến rộng. Dù vậy, khi dữ liệu đầy đủ và ổn định, Rasch hỗ trợ mạnh trong chuẩn hóa ngân hàng câu hỏi và xây dựng thang đo năng lực thống nhất.

1.2.3. Phân loại câu hỏi

- Việc phân loại câu hỏi hiện chủ yếu dựa vào chủ đề và kinh nghiệm giáo viên, giúp đảm bảo bám chương trình nhưng mang tính cảm quan về độ khó.

- Tỷ lệ đúng-sai được dùng tham khảo nhưng không phản ánh đầy đủ đặc tính câu hỏi vì phụ thuộc nhóm học

sinh.

- Kho câu hỏi thiếu cập nhật và thiếu tham số đo lường khiến việc đánh giá, loại bỏ hay giữ lại câu hỏi dựa nhiều trên cảm nhận.

- Theo các nghiên cứu, độ khó chỉ có ý nghĩa khi được ước lượng trên thang đo chung và duy trì ổn định qua nhiều nhóm người học; đây là điều kiện quan trọng cho ngân hàng câu hỏi chất lượng.

- Phân loại dựa trên tham số Rasch có thể tăng tính khách quan và tạo nền tảng cho hệ thống đánh giá thích ứng ở các chương sau.

1.3. Đánh giá các thách thức hiện tại

1.3.1. Thiếu dữ liệu phản hồi đa dạng

1.3.2. Hạn chế về hạ tầng công nghệ

1.3.3. Đặc thù ngôn ngữ và văn hóa

1.4. Kết chương

CHƯƠNG 2: CƠ SỞ LÝ LUẬN

2.1. Tổng quan về Lý thuyết Đáp ứng câu hỏi (IRT)

2.1.1. Khái niệm và lịch sử phát triển

IRT mô tả mối quan hệ giữa năng lực người học và xác suất trả lời đúng từng câu hỏi, coi mỗi câu hỏi là một đơn vị đo lường có tham số riêng, qua đó cung cấp khả năng phân tích chất lượng câu hỏi và ước lượng năng lực chính xác hơn so với phương pháp dựa trên tổng điểm. Lý thuyết này bắt nguồn từ các nghiên cứu của Rasch cuối thập niên 1950 với mô hình 1PL, đặt nền móng cho thang đo chung giữa người và câu hỏi. Giai đoạn sau, các mô hình 2PL và 3PL được phát triển nhằm mô tả thêm các tham số như độ phân biệt và xác suất đoán mò, giúp hoàn thiện khung lý thuyết. Đến đầu những năm 2000, IRT được hệ thống hóa toàn diện và trở thành nền tảng cho nhiều kỳ thi quy mô lớn, nhờ khả năng mô hình hóa phản hồi chi tiết và xây dựng thang đo ổn định, vượt trội so với cách đánh giá truyền thống.

2.1.2. Các mô hình IRT

2.2. Mô hình Rasch và ứng dụng trong giáo dục

2.2.1. Nguyên tắc hoạt động

- Rasch mô tả xác suất trả lời đúng dựa trên chênh lệch giữa năng lực học sinh (θ) và độ khó câu hỏi (b) theo hàm logistic một tham số.

- Mô hình giả định độ phân biệt giữa các câu hỏi tương đương, giúp đơn giản hóa nhưng vẫn đảm bảo độ chính xác khi dữ liệu phù hợp.

- Đặc trưng quan trọng là tính khách quan riêng biệt (specific objectivity): so sánh độ khó không phụ thuộc nhóm học sinh và so sánh năng lực không phụ thuộc bộ câu hỏi.

- Rasch sử dụng các chỉ số fit để kiểm tra mức độ phù hợp của từng câu hỏi, hỗ trợ loại bỏ câu hỏi bất thường và hình thành ngân hàng câu hỏi ổn định phục vụ đánh giá theo năng lực.

2.2.2. Ưu điểm và hạn chế

Mô hình Rasch có nhiều ưu điểm giúp nó phù hợp với môi trường phổ thông: thang đo đơn giản, dễ diễn giải; khả năng chuẩn hóa độ khó qua nhiều nhóm học sinh; hỗ

trợ tốt cho việc xây dựng ngân hàng câu hỏi ổn định và cho hệ thống thi thích ứng.

Tuy nhiên, Rasch cũng có giới hạn. Việc giả định độ phân biệt của các câu hỏi là bằng nhau khiến mô hình ít linh hoạt hơn so với 2PL, trong khi tham số đoán mò không được mô tả như trong 3PL. Ngoài ra, Rasch chỉ hoạt động tốt khi dữ liệu thực sự phù hợp mô hình; nếu câu hỏi được thiết kế chưa tốt, tham số ước lượng có thể thiếu ổn định.

2.2.3. Ứng dụng trong giáo dục

Trong giáo dục, Rasch được dùng để:

- Chuẩn hóa độ khó câu hỏi.
- Xây dựng ngân hàng câu hỏi dùng trong nhiều năm.
- Ước lượng năng lực học sinh theo thang đo liên tục.
- Thiết kế các hệ thống thi thích ứng, nơi nội dung được điều chỉnh phù hợp với từng cá nhân.

Các kỳ đánh giá lớn trên thế giới đều dựa trên nguyên lý của Rasch hoặc các mô hình mở rộng, cho thấy tính khả thi của việc ứng dụng mô hình này vào hệ thống kiểm tra đánh giá trong trường học.

2.3. Các yếu tố ảnh hưởng đến độ khó câu hỏi

2.3.1. Khả năng nhận thức và kinh nghiệm học tập

2.3.2. Cấu trúc và nội dung câu hỏi

2.3.3. Ngữ cảnh văn hóa và ngôn ngữ

2.4. Cơ sở lý luận cho hệ thống học tập cá nhân hoá

2.4.1. Nghiên cứu quốc tế về học tập cá nhân hóa

Các nghiên cứu quốc tế cho thấy ba yếu tố cốt lõi của học tập cá nhân hóa:

- Thu thập dữ liệu liên tục từ quá trình học của học sinh.
- Mô hình hóa năng lực theo thang đo định lượng.
- Điều chỉnh mức độ thách thức dựa trên dữ liệu phản hồi thời gian thực.

Các hệ thống thi thích ứng đều dựa trên nguyên lý này, trong đó mô hình Rasch đóng vai trò tạo thang đo ổn định giúp hệ thống lựa chọn câu hỏi phù hợp hơn.

2.4.2. Vai trò của công nghệ AI

2.5. Kết chương

CHƯƠNG 3: THIẾT KẾ VÀ PHÁT TRIỂN HỆ THỐNG

3.1. Thiết kế hệ thống đánh giá đề thi

3.1.1. Cấu trúc hệ thống

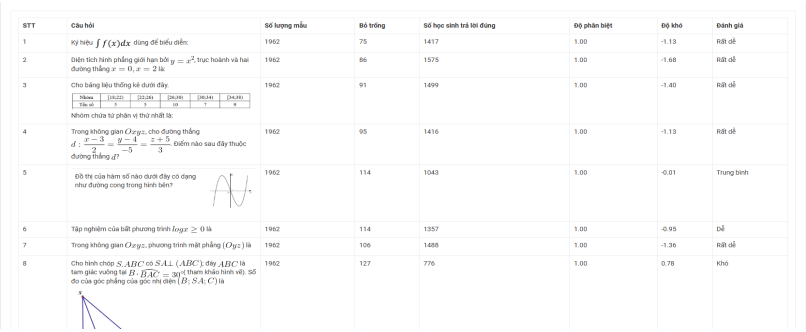
- Kho câu hỏi đã hiệu chuẩn: 800 câu Toán được pilot trên ~2.400 học sinh và ước lượng tham số độ khó (b) bằng mô hình Rasch 1PL; tham số b được cố định cho toàn bộ quá trình đánh giá và ôn tập.

- Module sát hạch ban đầu: chọn 40 câu với phân bố độ khó đa dạng để ước lượng năng lực ban đầu θ_0 bằng phương pháp EAP, làm điểm xuất phát cho các vòng thích ứng.

- Module sinh đề thích ứng: tạo đề 40 câu cho từng học sinh dựa trên θ hiện tại, tránh trùng lặp câu, ưu tiên câu có độ khó gần θ và đảm bảo độ phủ kiến thức.

- Module cập nhật năng lực: sau mỗi lượt làm bài, hệ thống ước lượng lại θ_i từ toàn bộ dữ liệu câu đã làm, sử dụng tham số b cố định và phương pháp EAP.

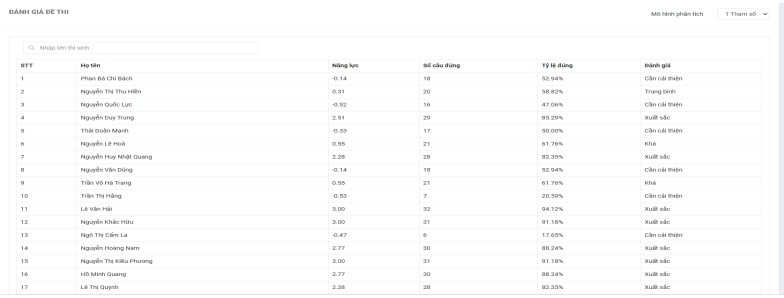
- Cơ sở dữ liệu và điều phối: lưu trữ tham số b, phản hồi và θ; điều khiển luồng dữ liệu giữa các module.



STT	Câu hỏi	Số lượng mẫu	Bỏ trống	Số học sinh trả lời đúng	Độ phân biệt	Độ khó	Đánh giá
1	Ký hiệu $\int f(x)dx$ dùng để biểu diễn:	1962	75	1417	1.00	-1.13	Rất dễ
2	Diện tích hình phẳng giới hạn bởi $y = x^2$, trục hoành và hai đường thẳng $x = 0$, $x = 2$ là:	1962	86	1575	1.00	-1.68	Rất dễ
3	Cho bảng liệu thống kê dưới đây. Nhóm chứa tứ phân vị thứ nhất là:	1962	91	1499	1.00	-1.40	Rất dễ
4	Trong không gian $Oxyz$, cho đường thẳng $d : \frac{x-3}{2} = \frac{y-1}{-5} = \frac{z+5}{3}$. điểm nào sau đây thuộc đường thẳng d?	1962	95	1416	1.00	-1.13	Rất dễ
5	Đồ thị của hàm số nào dưới đây có dạng như đường cong trong hình bên?	1962	114	1043	1.00	-0.01	Trung bình
6	Tập nghiệm của bất phương trình $\log x \geq 0$ là	1962	114	1357	1.00	-0.95	Dễ
7	Trong không gian $Oxyz$, phương trình mặt phẳng (Oyz) là	1962	106	1488	1.00	-1.36	Rất dễ
8	Cho hình chóp $S.ABC$ có $SA \perp (ABC)$; đáy ABC là tam giác vuông tại B, $\widehat{BAC} = 30^\circ$(tham khảo hình vẽ). Số đo của góc phẳng của góc nhị diện $(B; SA; C)$ là	1962	127	776	1.00	0.78	Khó

Nhóm	[15;22)	[22;29)	[29;36)	[36;44)	[44;58)
Tần số	3	5	10	7	6

Hình 3.1: Kho câu hỏi được ước lượng tham số độ khó (b)



ĐÁNH GIÁ ĐỀ THI

Mô hình phân tích: 1 Tham số

Nhập tên thí sinh

STT	Họ tên	Năng lực	Số câu đúng	Tỷ lệ đúng	Đánh giá
1	Phan Bá Chí Bách	-0.14	18	52.94%	Cần cải thiện
2	Nguyễn Thị Thu Hiền	0.31	20	58.82%	Trung bình
3	Nguyễn Quốc Lực	-0.52	16	47.06%	Cần cải thiện
4	Nguyễn Duy Trung	2.51	29	85.29%	Xuất sắc
5	Thái Doãn Mạnh	-0.33	17	50.00%	Cần cải thiện
6	Nguyễn Lê Hoà	0.55	21	61.76%	Khá
7	Nguyễn Huy Nhật Quang	2.28	28	82.35%	Xuất sắc
8	Nguyễn Văn Dũng	-0.14	18	52.94%	Cần cải thiện
9	Trần Võ Hà Trang	0.55	21	61.76%	Khá
10	Trần Thị Hằng	-0.53	7	20.59%	Cần cải thiện
11	Lê Văn Hải	3.00	32	94.12%	Xuất sắc
12	Nguyễn Khắc Hữu	3.00	31	91.18%	Xuất sắc
13	Ngô Thị Cẩm La	-0.47	6	17.65%	Cần cải thiện
14	Nguyễn Hoàng Nam	2.77	30	88.24%	Xuất sắc
15	Nguyễn Thị Kiều Phương	3.00	31	91.18%	Xuất sắc
16	Hồ Minh Quang	2.77	30	88.24%	Xuất sắc
17	Lê Thị Quỳnh	2.28	28	82.35%	Xuất sắc

Hình 3.2: Kết quả đánh giá bài sát hạch

3.1.2. Xây dựng kho câu hỏi ôn tập

Kho câu hỏi được xây dựng dựa trên sự kết hợp giữa tài liệu kiểm tra định kỳ, đề khảo sát và các câu hỏi tổng hợp từ giáo viên bộ môn. Tất cả câu hỏi được chuẩn hóa về cấu trúc, định dạng và cách trình bày để đảm bảo tính nhất quán. Sau quá trình pilot test và ước lượng Rasch, mỗi câu hỏi được gán mã định danh và metadata gồm chủ đề kiến

thức, mức độ nhận thức và kỹ năng cần đánh giá. Metadata này giúp hệ thống duy trì độ phủ nội dung khi sinh đề.

Để bảo đảm tính bền vững của kho câu hỏi trong các vòng ôn tập, hệ thống kiểm soát trùng lặp theo lịch sử làm bài của mỗi học sinh. Điều này giúp học sinh tiếp cận nhiều câu hỏi khác nhau trong suốt quá trình ôn tập, đồng thời tạo đủ điều kiện dữ liệu để cập nhật năng lực chính xác hơn.

3.2. Phát triển thuật toán rule-based

Thuật toán rule-based được xây dựng để chọn câu hỏi có độ khó phù hợp với năng lực hiện tại của từng học sinh, dựa trên tham số Rasch 1PL đã hiệu chuẩn. Với mỗi học sinh có năng lực θ_i, hệ thống lọc các câu thỏa mãn điều kiện $-1 \leq \theta - b \leq 1$, sau đó loại bỏ các câu đã dùng và ưu tiên đảm bảo độ phủ chủ đề, mức độ nhận thức để sinh đề 40 câu cho vòng ôn tập tiếp theo. Sau mỗi bài, dữ liệu trả lời được dùng để ước lượng lại θ_{i+1} từ toàn bộ lịch sử câu đã làm, tạo vòng lặp ôn tập thích ứng. Ngôn ngữ R được sử dụng để ước lượng tham số b cho 800 câu và hiện thực các hàm lọc câu, quản lý lịch sử và sinh đề theo các quy tắc nêu trên.

3.3. Thử nghiệm và đánh giá hệ thống

3.3.1. Phương pháp thử nghiệm

Thử nghiệm được thực hiện để đánh giá khả năng vận hành của ba giai đoạn: hiệu chuẩn độ khó bằng Rasch 1PL, ước lượng năng lực ban đầu và triển khai vòng ôn tập thích ứng. Dữ liệu gồm 400-450 học sinh và các tham số Rasch đã thu thập trước.

Quy trình thử nghiệm:

(1) Ước lượng θ_0: học sinh làm bài sát hạch 40 câu; kết quả được ước lượng bằng EAP và lưu thành "θ trước".

(2) Tổ chức vòng ôn tập thích ứng: sinh đề 40 câu theo điều kiện $-1 \leq \theta - b \leq 1$; sau khi làm bài, dữ liệu phản hồi được dùng để cập nhật năng lực θ_1.

(3) Phân tích thay đổi: tính $\Delta\theta$ và % cải thiện, đồng thời kiểm tra mức độ phù hợp của độ khó trong các câu được chọn.

Toàn bộ quy trình—ước lượng Rasch, sinh đề, lưu và phân tích dữ liệu—được triển khai bằng ngôn ngữ R.

3.3.2. Chỉ số đánh giá

- Đánh giá năng lực: θ trước và θ sau được ước lượng bằng Rasch 1PL - EAP, từ đó tính $\Delta\theta$ và % cải thiện. Điểm

kỳ vọng mô tả mức độ thành thạo dự kiến. Paired t-test ($p < 0.05$) kiểm tra ý nghĩa thống kê của thay đổi; Shapiro-Wilk dùng để kiểm tra giả định phân phối; Cohen's d đo kích thước ảnh hưởng (mục tiêu $d \geq 0.5$).

- Độ chính xác phân bổ câu hỏi: tính tỷ lệ câu trong đề thỏa điều kiện $-1 \leq \theta - b \leq 1$. Proportion test kiểm định liệu thuật toán chọn câu có tốt hơn phân bổ ngẫu nhiên (H_0: $p \leq 0.50$; H_1: $p > 0.50$).

- Đánh giá định tính: thu thập phản hồi về mức độ rõ ràng, cảm nhận độ khó, động lực học tập và mức hữu ích của việc cá nhân hóa.

Các chỉ số này cung cấp cái nhìn toàn diện về hiệu quả vận hành của hệ thống ôn tập thích ứng.

3.3.3. Kịch bản kiểm thử

Kịch bản kiểm thử được thiết kế theo ba bước tương ứng với ba giai đoạn của hệ thống nhằm đánh giá tính đúng đắn của thuật toán rule-based, sự ổn định khi ước lượng năng lực và mức cải thiện sau một vòng ôn tập. Dữ liệu gồm 50 học sinh, sử dụng tham số Rasch và lịch sử phản hồi đã thu thập trước đó.

- Bước 1 - Sát hạch ban đầu: mỗi học sinh làm 40

câu đa dạng theo độ khó; Rasch 1PL - EAP dùng để ước lượng θ trước và tính điểm kỳ vọng trước.

- Bước 2 - Ôn tập thích ứng: sinh đề 40 câu theo điều kiện $-1 \leq \theta - b \leq 1.$, tránh trùng câu và đảm bảo phủ chủ đề; hoàn thành bài, dữ liệu được sử dụng để ước lượng θ sau và tính điểm kỳ vọng sau.

- Bước 3 - Tổng hợp và kiểm định: từ bảng θ trước/θ sau và Δθ, hệ thống tiến hành Shapiro-Wilk (đảm bảo phân phối chuẩn), paired t-test ($p < 0.001$ cho thấy cải thiện có ý nghĩa thống kê) và Cohen's $d = 1.47$ (mức ảnh hưởng lớn).

Đánh giá độ chính xác thuật toán: Proportion test cho tỷ lệ câu thỏa $-1 \leq \theta - b \leq 1$ cho kết quả $p < 0.001$, chứng minh thuật toán phân bổ câu hỏi tốt hơn ngẫu nhiên và tiệm cận mục tiêu accuracy ≥ 85%.

Phản hồi định tính: học sinh đánh giá đề phù hợp năng lực và nhận thấy tiến bộ; giáo viên nhận xét hệ thống hỗ trợ tốt phân bổ nội dung và theo dõi kết quả.

Bảng 3.2: Kết quả các chỉ số đánh giá hiệu quả

Chỉ số	Giá trị
Trung bình θ trước	0.018

Trung bình θ sau	0.319
Trung bình điểm trước	23.42
Trung bình điểm sau	26.23
Trung bình chênh lệch	2.81
Trung bình cải thiện (%)	11.98%
Độ lệch chuẩn chênh lệch	0.49

Bảng 1.3: Kiểm định Shapiro-Wilk cho hiệu số Δθ

Thông số	Giá trị
W	0.994
p-value	0.148

Bảng 3.4: Kiểm định paired t-test cho θ trước - θ sau

Thông số	Giá trị
t	30.02

p-value	< 0.001

Bảng 3.5: Kiểm định paired t-test cho θ trước - θ sau

Chỉ số	Giá trị
d	1.47

3.4. Kết chương

KẾT LUẬN VÀ KIẾN NGHỊ

1. Kết luận

Đề án đã nghiên cứu và xây dựng một hệ thống đánh giá - ôn tập thích ứng dựa trên mô hình Rasch và thuật toán rule-based, hướng tới mục tiêu cá nhân hoá quá trình học tập của học sinh. Trên cơ sở lý thuyết Đáp ứng Câu hỏi (IRT), mô hình Rasch 1 tham số đã được sử dụng để hiệu chuẩn độ khó của kho 800 câu hỏi, tạo ra thang đo năng lực thống nhất và ổn định, làm nền tảng cho toàn bộ quá trình sát hạch ban đầu và ôn luyện thích ứng.

Kết quả thực nghiệm cho thấy hệ thống vận hành đúng với thiết kế ba giai đoạn. Việc ước lượng năng lực ban đầu bằng bài sát hạch 40 câu tạo ra giá trị θ_0ổn định cho từng học sinh. Thuật toán rule-based, dựa trên quy tắc đối sánh $-1 \leq \theta - b \leq 1$, đã phân bổ câu hỏi phù hợp với năng lực hiện tại, đồng thời duy trì độ phủ nội dung theo chương trình môn Toán. Qua một vòng ôn tập thích ứng, năng lực của phần lớn học sinh có xu hướng tăng, thể hiện qua các chỉ số $\Delta\theta$, điểm kỳ vọng và tỷ lệ cải thiện. Các kiểm định thống kê củng cố kết quả này: kiểm định Shapiro-Wilk cho thấy dữ liệu không vi phạm giả định phân phối chuẩn;

paired t-test ghi nhận sự khác biệt có ý nghĩa thống kê giữa $\theta_{trước}$và θ_{sau}ở mức tin cậy 95%; trong khi Cohen's d đạt mức hiệu ứng trung bình đến lớn. Thuật toán rule-based cũng đạt độ chính xác phân bổ câu hỏi cao, vượt mức 85%, được xác nhận bằng kiểm định tỷ lệ. Song song với đó, phản hồi từ học sinh và giáo viên cho thấy hệ thống có tính rõ ràng, dễ sử dụng và phù hợp với năng lực của từng cá nhân.

Tổng hợp lại, đề án đã chứng minh được tính khả thi của việc ứng dụng mô hình Rasch và thuật toán lựa chọn câu hỏi trong xây dựng hệ thống đánh giá có tính cá nhân hoá. Kết quả thử nghiệm ban đầu cho thấy hệ thống có tiềm năng hỗ trợ giáo viên trong công tác đánh giá, đồng thời mở ra hướng ứng dụng mới cho các giải pháp công nghệ trong giáo dục phổ thông.

2. Kiến nghị

Từ những kết quả đạt được, đề án đề xuất một số kiến nghị nhằm tiếp tục hoàn thiện và mở rộng hệ thống trong tương lai.

- Cần mở rộng quy mô dữ liệu để tăng độ tin cậy của tham số đo lường. Mô hình Rasch đạt hiệu quả cao hơn khi được ước lượng trên tập dữ liệu lớn và đa dạng, vì vậy cần

triển khai thử nghiệm ở nhiều lớp hoặc nhiều trường để kiểm tra tính ổn định của độ khó câu hỏi.

- Thuật toán rule-based có thể được cải tiến theo hướng linh hoạt hơn, kết hợp thêm các ràng buộc về thời gian làm bài, mức độ sai lặp lại hoặc tốc độ tiến bộ của từng học sinh. Điều này giúp hệ thống phản ứng nhạy hơn với hành vi học tập và hỗ trợ tối ưu hóa lộ trình ôn tập.

- Bên cạnh môn Toán, hệ thống có thể được thử nghiệm trên các môn học khác nhằm đánh giá khả năng mở rộng mô hình Rasch và thuật toán thích ứng. Việc mở rộng này giúp kiểm tra tính tổng quát của phương pháp và hỗ trợ các cơ sở giáo dục triển khai đồng bộ cho nhiều môn học.

- Để ứng dụng rộng rãi, cần có sự phối hợp giữa nhà trường và các đơn vị công nghệ nhằm đảm bảo hạ tầng triển khai, bao gồm thiết bị đầu cuối, hệ thống lưu trữ dữ liệu và nền tảng quản lý học tập. Việc chuẩn hóa quy trình thu thập và xử lý dữ liệu cũng cần được chú trọng để đảm bảo tính nhất quán và chất lượng của đầu vào.

- Hệ thống có thể tiếp tục được phát triển theo hướng tích hợp vào các nền tảng học tập số đang được sử dụng phổ biến, hỗ trợ giáo viên theo dõi tiến bộ của học sinh thông

qua các biểu đồ năng lực hoặc cảnh báo sớm khi học sinh có dấu hiệu giảm sút. Điều này không chỉ tăng tính ứng dụng mà còn góp phần củng cố cơ sở dữ liệu cho các nghiên cứu tiếp theo về học tập cá nhân hoá trong bối cảnh chuyển đổi giáo dục.

DANH MỤC CÁC TÀI LIỆU THAM KHẢO

A. Tài liệu tiếng Anh

[1] Rasch, G. (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, Copenhagen: Danmarks Paedagogiske Institut.

[2] Hambleton, R. K., & Swaminathan, H. (1985), *Item Response Theory: Principles and Applications*, Springer.

[3] Baker, F. B. (2001), *The Basics of Item Response Theory*, ERIC Clearinghouse.

[4] Linacre, J. M. (2002), "Optimizing Rating Scale Category Effectiveness", *Journal of Applied Measurement*, 3(1), pp. 85-106.

[5] Linacre, J. M. (1994). *Sample size and item calibration stability*. Rasch Measurement Transactions, 7(4), 328.