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

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(***Theo định hướng ứng dụng***)

HÀ NỘI - NĂM 2025

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

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT
(*Theo định hướng ứng dụng*)

NGƯỜI HƯỚNG DẪN KHOA HỌC: TS. ĐÀO NGỌC PHONG

HÀ NỘI - NĂM 2025

LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi.

Các số liệu, kết quả nêu trong đề án tốt nghiệp là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tác giả đề án tốt nghiệp

Trần Thanh Quang

LỜI CẢM ƠN

Trong quá trình học tập và thực hiện đề án Thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông, em đã nhận được sự quan tâm, hỗ trợ và động viên từ nhiều thầy cô, cơ quan, bạn bè và gia đình. Đây là những nguồn động lực quý giá giúp em hoàn thành đề tài nghiên cứu.

Trước hết, em xin bày tỏ lòng biết ơn sâu sắc tới **TS. Đào Ngọc Phong**, người thầy đã tận tình hướng dẫn, định hướng khoa học và đồng hành cùng em trong suốt quá trình thực hiện đề tài. Những góp ý, chỉ dẫn và sự hỗ trợ kịp thời của thầy đã giúp em vượt qua nhiều khó khăn, hoàn thiện đề án theo đúng mục tiêu đặt ra.

Em cũng xin trân trọng cảm ơn các thầy cô của Học viện Công nghệ Bưu chính Viễn thông, những người đã truyền đạt cho em hệ thống kiến thức nền tảng và chuyên sâu trong suốt thời gian học tập, tạo điều kiện để em có đủ năng lực tiếp cận và triển khai đề tài.

Cuối cùng, em xin gửi lời cảm ơn chân thành tới gia đình, bạn bè và đồng nghiệp đã luôn ở bên, động viên và hỗ trợ em trong quá trình học tập và thực hiện đề án.

Em xin trân trọng cảm ơn!

MỤC LỤC

LỜI CAM ĐOAN ..... **i**
LỜI CẢM ƠN ..... **ii**
DANH SÁCH CÁC THUẬT NGỮ, CHỮ VIẾT TẮT ..... **v**
DANH SÁCH BẢNG ..... **v**
DANH SÁCH HÌNH VẼ ..... **vii**
MỞ ĐẦU ..... **vii**
1. Lý do chọn đề tài 1
2. Tổng quan về vấn đề nghiên cứu 2
3. Mục đích nghiên cứu 3
4. Đối tượng và phạm vi nghiên cứu 4
5. Phương pháp nghiên cứu 4
CHƯƠNG 1: THỰC TRẠNG VÀ PHÂN TÍCH **6**
1.1. Thực trạng hệ thống đánh giá đề thi 6
1.1.1. Phương pháp truyền thống 6
1.1.2. Các ứng dụng công nghệ hiện có 7
1.1.3. Hiệu quả và hạn chế 7
1.2. Phân tích dữ liệu phản hồi từ đề thi 8
1.2.1. Thu thập và tổng hợp dữ liệu 9
1.2.2. Ứng dụng mô hình Rasch 9
1.2.3. Phân loại câu hỏi 10
1.3. Đánh giá các thách thức hiện tại 11
1.3.1. Thiếu dữ liệu phản hồi đa dạng 11
1.3.2. Hạn chế về hạ tầng công nghệ 12
1.3.3. Đặc thù ngôn ngữ và văn hóa 12
1.4. Kết chương 13
CHƯƠNG 2: CƠ SỞ LÝ LUẬN **15**
2.1. Tổng quan về Lý thuyết Đáp ứng câu hỏi (IRT) 15
2.1.1. Khái niệm và lịch sử phát triển 15

2.1.2. Các mô hình IRT 16
2.2. Mô hình Rasch và ứng dụng trong giáo dục 18
2.2.1. Nguyên tắc hoạt động 18
2.2.2. Ưu điểm và hạn chế 20
2.2.3. Ứng dụng trong giáo dục 21
2.3. Các yếu tố ảnh hưởng đến độ khó câu hỏi 22
2.3.1. Khả năng nhận thức và kinh nghiệm học tập 23
2.3.2. Cấu trúc và nội dung câu hỏi 23
2.3.3. Ngữ cảnh văn hóa và ngôn ngữ 24
2.4. Cơ sở lý luận cho hệ thống học tập cá nhân hoá 25
2.4.1. Nghiên cứu quốc tế về học tập cá nhân hóa 25
2.4.2. Vai trò của công nghệ AI 26
2.5. Kết chương 27
CHƯƠNG 3: THIẾT KẾ VÀ PHÁT TRIỂN HỆ THỐNG 29
3.1. Thiết kế hệ thống đánh giá đề thi 29
3.1.1. Cấu trúc hệ thống 29
3.1.2. Xây dựng kho câu hỏi ôn tập 34
3.2. Phát triển thuật toán rule-based 35
3.3. Thử nghiệm và đánh giá hệ thống 37
3.3.1. Phương pháp thử nghiệm 37
3.3.2. Chỉ số đánh giá 38
3.3.3. Kịch bản kiểm thử 40
3.4. Kết chương 45
KẾT LUẬN VÀ KIẾN NGHỊ 47
1. Kết luận 47
2. Kiến nghị 47
DANH MỤC CÁC TÀI LIỆU THAM KHẢO 49

DANH SÁCH CÁC THUẬT NGỮ, CHỮ VIẾT TẮT

Ký hiệu / Chữ viết tắt	Tiếng Anh	Tiếng việt
1PL	One-Parameter Logistic Model	Mô hình IRT 1 tham số (mô hình Rasch), dùng để hiệu chuẩn độ khó câu hỏi.
2PL	Two-Parameter Logistic Model	Mô hình IRT 2 tham số (độ khó và độ phân biệt câu hỏi).
3PL	Three-Parameter Logistic Model	Mô hình IRT 3 tham số (độ khó, độ phân biệt và đoán mò).
AI	Artificial Intelligence	Trí tuệ nhân tạo, công nghệ được nhắc đến trong bối cảnh chuyển đổi số giáo dục.
EAP	Expected A Posteriori	Phương pháp ước lượng tham số năng lực θ trong mô hình Rasch.
IRT	Item Response Theory	Lý thuyết Đáp ứng Câu hỏi - cơ sở lý thuyết cho mô hình Rasch và các mô hình đo lường hiện đại.
R	R Programming Language	Ngôn ngữ lập trình dùng để ước lượng mô hình Rasch, xử lý và phân tích dữ liệu.
THPT	Trung học phổ thông	Cấp học là bối cảnh triển khai thử nghiệm hệ thống đánh giá - ôn tập thích ứng.
b	Item Difficulty Parameter	Tham số độ khó của câu hỏi trong mô hình Rasch 1 tham số.

θ (theta)	Ability Parameter	Tham số năng lực của học sinh trên thang đo Rasch.
Δθ	Change in Ability	Độ chênh lệch năng lực trước và sau khi tham gia một vòng ôn tập thích ứng.
1PL	One-Parameter Logistic Model	Mô hình IRT 1 tham số (mô hình Rasch), dùng để hiệu chuẩn độ khó câu hỏi.
2PL	Two-Parameter Logistic Model	Mô hình IRT 2 tham số (độ khó và độ phân biệt câu hỏi).
3PL	Three-Parameter Logistic Model	Mô hình IRT 3 tham số (độ khó, độ phân biệt và đoán mò).

DANH SÁCH BẢNG

Bảng 3.1. Tổng hợp dữ liệu đánh giá năng lực học sinh 41
Bảng 3.2. Kết quả các chỉ số đánh giá hiệu quả 42
Bảng 3.3. Kiểm định Shapiro–Wilk cho hiệu số Δθ 43
Bảng 3.4. Kiểm định paired t-test cho θ trước – θ sau 44
Bảng 3.5. Kích thước hiệu ứng Cohen's d 44

DANH SÁCH HÌNH VẼ

Hình 2.1. Các đường cong ICC của các câu hỏi 19
Hình 2.2. Đồ thị hàm thông tin bài thi 20
Hình 2.3. Biểu đồ so sánh năng lực thí sinh và độ khó câu hỏi 22
Hình 3.1. Kho câu hỏi được ước lượng tham số độ khó (b) 30
Hình 3.2. Kết quả đánh giá bài sát hạch 31
Hình 3.3. Hệ thống cho phép học sinh tham gia bài thi ôn tập 32
Hình 3.4. Học sinh tham gia làm bài ôn tập 33

MỞ ĐẦU

1. Lý do chọn đề tài

Cùng với quá trình chuyển đổi số đang diễn ra trên toàn thế giới, các công nghệ như trí tuệ nhân tạo (AI), học máy và phân tích dữ liệu đã tác động ngày càng rộng rãi đến nhiều lĩnh vực, bao gồm cả giáo dục. Những công nghệ này tạo ra các hướng tiếp cận mới cho hoạt động dạy và học, đặc biệt hữu ích trong các nhiệm vụ cần xử lý lượng dữ liệu lớn hoặc đưa ra các quyết định mang tính cá nhân hóa. Tuy vậy, tại nhiều cơ sở giáo dục, việc đánh giá vẫn dựa trên các phương pháp truyền thống, sử dụng đề thi cố định và hầu như không được điều chỉnh theo năng lực của từng học sinh. Cách tiếp cận đó khiến việc đánh giá chưa thể hiện đầy đủ năng lực thực tế của người học và chưa tận dụng hiệu quả dữ liệu thu được trong quá trình làm bài.

Trong nghiên cứu đo lường giáo dục, Lý thuyết Đáp ứng Câu hỏi (Item Response Theory - IRT) và đặc biệt là mô hình Rasch được xem là những tiếp cận quan trọng cho phép mô hình hóa mối quan hệ giữa năng lực người học và đặc tính của câu hỏi [1], [2], [4]. Các mô hình này giúp xây dựng thang đo năng lực liên tục, ổn định và hỗ trợ phân tích độ khó dựa trên dữ liệu phản hồi. Khi được kết hợp với công nghệ số, IRT mở ra khả năng xây dựng các hệ thống đánh giá có tính thích ứng, trong đó nội dung được lựa chọn dựa trên năng lực hiện tại của từng người học thay vì áp dụng chung một đề thi cho tất cả.

Dù nhiều nền tảng học trực tuyến đã bổ sung các chức năng gợi ý hoặc phân tầng nội dung, phần lớn hệ thống vẫn vận hành dựa trên dạng bài tập cố định và chưa khai thác hiệu quả dữ liệu phản hồi để suy luận năng lực người học theo mô hình IRT. Điều này tạo ra một khoảng thiếu hụt quan trọng khi phát triển các hệ thống đánh giá và hỗ trợ học tập hướng tới mức độ cá nhân hóa sâu hơn.

Từ những lý do trên, đề tài "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" được lựa chọn với kỳ vọng góp phần cung cấp một tiếp cận hệ thống dựa trên mô hình Rasch và dữ liệu thực nghiệm. Đề tài mang giá trị khoa học khi vận dụng các phương pháp đo lường hiện

đại vào bối cảnh giáo dục tại Việt Nam, đồng thời có ý nghĩa thực tiễn khi hỗ trợ giáo viên trong hoạt động đánh giá và phù hợp với định hướng chuyển đổi số giáo dục đến năm 2030 của Chính phủ.

2. Tổng quan về vấn đề nghiên cứu

Lý thuyết Đáp ứng Câu hỏi (Item Response Theory - IRT) là một khung tiếp cận quan trọng trong lĩnh vực đo lường, cho phép mô tả xác suất trả lời đúng thông qua mối liên hệ giữa mức độ thành thạo của người học và đặc trưng của câu hỏi. Trong số các mô hình thuộc IRT, mô hình Rasch một tham số được xem là nền tảng nhờ cấu trúc đơn giản và khả năng xây dựng thang đo năng lực nhất quán [1], [4]. Mô hình này giả định rằng xác suất trả lời đúng phụ thuộc chủ yếu vào chênh lệch giữa năng lực của người làm bài và độ khó của câu hỏi, qua đó hỗ trợ việc hiệu chuẩn câu hỏi và đánh giá năng lực một cách khách quan.

Nhiều nghiên cứu quốc tế cho thấy các mô hình IRT nói chung và mô hình Rasch nói riêng được áp dụng rộng rãi trong các hệ thống đánh giá quy mô lớn nhờ khả năng chuẩn hóa độ khó và phân tích chi tiết chất lượng từng câu [2], [3]. Các tài liệu kinh điển trong lĩnh vực này, như Hambleton & Swaminathan (1985) hay Baker (2001), nhấn mạnh rằng IRT tạo điều kiện để xây dựng các bài kiểm tra có khả năng phân hóa tốt và phản ánh chính xác hơn mức độ thành thạo của người học [2], [3]. Đây cũng là nền tảng lý thuyết để hình thành các hệ thống học tập có khả năng điều chỉnh nội dung theo năng lực, một hướng tiếp cận ngày càng được quan tâm trong giáo dục hiện đại.

Trong thực tế triển khai tại nhiều cơ sở giáo dục, các phần mềm hỗ trợ học tập hiện nay chủ yếu tập trung vào quản lý điểm số hoặc cung cấp ngân hàng câu hỏi theo cấu trúc tĩnh, chưa tích hợp cơ chế phân tích dữ liệu phản hồi để ước lượng năng lực theo mô hình IRT. Điều này khiến hệ thống chưa phản ánh đầy đủ mức độ tiến bộ của người học và còn hạn chế trong việc cá nhân hóa nội dung ôn tập. Mặt khác, việc phát triển các công cụ đánh giá theo IRT đòi hỏi dữ liệu phản hồi đủ lớn, quá trình kiểm định chất lượng câu hỏi chặt chẽ và thuật toán lựa chọn câu hỏi phù hợp với trình độ của từng cá nhân. Bên cạnh đó, khi áp dụng các mô hình đo lường vào bối

cảnh Việt Nam, một số yêu cầu đặc thù cần được xem xét, như đảm bảo chất lượng dữ liệu đầu vào, tiêu chuẩn hóa kho câu hỏi và sự phù hợp của thuật toán lựa chọn câu hỏi với cách trình bày nội dung trong chương trình phổ thông. Ngoài ra, hạ tầng công nghệ giữa các cơ sở giáo dục còn có sự khác biệt nhất định, dẫn đến yêu cầu về giải pháp triển khai phải có tính linh hoạt và chi phí hợp lý. Việc tận dụng các công cụ mã nguồn mở như R để ước lượng mô hình Rasch, xây dựng bộ quy tắc lựa chọn câu hỏi và xử lý dữ liệu phản hồi là một hướng đi phù hợp, góp phần giảm chi phí và tăng khả năng ứng dụng.

Có thể thấy việc nghiên cứu và phát triển hệ thống đánh giá dựa trên IRT kết hợp thuật toán lựa chọn câu hỏi phù hợp là cần thiết nhằm nâng cao tính cá nhân hóa trong học tập. Đây cũng là khoảng trống mà đề tài hướng tới, với mục tiêu điều chỉnh các tiếp cận lý thuyết quốc tế để bám sát với bối cảnh thực tế và nhu cầu triển khai trong giáo dục phổ thông Việt Nam.

3. Mục đích nghiên cứu

Đề án tập trung phát triển và kiểm thử một hệ thống đánh giá đề thi vận hành theo mô hình Rasch, nhằm hỗ trợ xây dựng lộ trình ôn tập tương ứng với mức năng lực của mỗi học sinh. Việc áp dụng mô hình Rasch cho phép chuẩn hóa độ khó câu hỏi và ước lượng năng lực trên một thang đo thống nhất, từ đó tạo cơ sở khoa học cho quá trình cá nhân hoá trong học tập. Bên cạnh việc phát triển mô hình đánh giá, đề án tập trung thiết kế và thử nghiệm một hệ thống ứng dụng thực tế, trong đó thuật toán lựa chọn câu hỏi được điều chỉnh theo năng lực cập nhật của từng học sinh. Thử nghiệm hệ thống nhằm đánh giá mức độ khả thi, tính ổn định và sự thay đổi năng lực người học sau một vòng ôn luyện thích ứng. Kết quả thu được sẽ giúp so sánh với phương pháp ôn tập truyền thống, từ đó xác định mức độ phù hợp của giải pháp trong bối cảnh giáo dục phổ thông.

Bên cạnh kết quả thực nghiệm, đề án còn nhằm đưa ra một tài liệu tham khảo cho các nghiên cứu tiếp theo về hệ thống học tập cá nhân hoá, nhất là trong bối cảnh chuyển đổi số trong giáo dục. Các kết quả đạt được kỳ vọng góp phần hỗ trợ các cơ

sở giáo dục triển khai những công cụ đánh giá có tính linh hoạt hơn, đồng thời gợi mở hướng ứng dụng các mô hình đo lường hiện đại trong thực tiễn dạy và học.

4. Đối tượng và phạm vi nghiên cứu

Đề án tập trung phân tích hệ thống đánh giá đề thi cùng thuật toán phân phối câu hỏi ôn tập dựa trên mô hình Rasch kết hợp phương pháp rule-based. Đối tượng nghiên cứu trọng tâm là học sinh trung học phổ thông, trong đó nhóm học sinh lớp 10 được chọn làm mẫu thử nghiệm vì phù hợp với mục tiêu thiết lập lộ trình học tập cơ bản và dễ tiếp cận các hình thức ôn tập dựa trên dữ liệu phản hồi.

Phạm vi nghiên cứu tập trung vào một nhóm học sinh thuộc một trường trung học phổ thông để bảo đảm sự thống nhất trong điều kiện triển khai và thuận tiện cho việc kiểm soát các yếu tố tác động đến quá trình thu thập dữ liệu. Nghiên cứu tập trung vào môn Toán, với kho dữ liệu gồm 800 câu hỏi trắc nghiệm được phân loại theo chủ đề và mức độ nhận thức theo chương trình Giáo dục phổ thông hiện hành. Các câu hỏi này được sử dụng cho cả giai đoạn hiệu chuẩn độ khó và giai đoạn thử nghiệm thuật toán chọn câu.

Thời gian triển khai nghiên cứu được bố trí theo các giai đoạn: xây dựng mô hình Rasch trên dữ liệu pilot, phát triển thuật toán lựa chọn câu hỏi, và tiến hành thử nghiệm trên người học. Việc giới hạn phạm vi theo nhóm đối tượng và môn học cụ thể giúp đảm bảo độ tin cậy của quy trình thử nghiệm, đồng thời tạo cơ sở cho việc mở rộng phạm vi ứng dụng trong các nghiên cứu tiếp theo

5. Phương pháp nghiên cứu

Đề án kết hợp phương pháp lý thuyết, mô phỏng và thực nghiệm để đảm bảo tính chặt chẽ về mặt khoa học cũng như sự khả thi trong quá trình triển khai. Trước tiên, ở phương diện lý thuyết, nghiên cứu tổng hợp và phân tích các tài liệu về Lý thuyết Đáp ứng Câu hỏi (IRT) và mô hình Rasch một tham số, đặc biệt chú trọng đến đặc tính đo lường, cách kiểm định mức độ phù hợp và các kỹ thuật ước lượng tham số. Các nguồn lý thuyết chủ yếu dựa trên các công trình kinh điển trong lĩnh vực đo lường giáo dục [1], [2], [3], [4], là cơ sở để xây dựng mô hình hiệu chuẩn độ khó câu hỏi và thang đo năng lực người học. Đồng thời, nghiên cứu tham khảo các hướng tiếp

cận về học tập cá nhân hóa và cơ chế điều chỉnh nội dung dựa trên dữ liệu phản hồi, nhằm hình thành nền tảng phương pháp cho phần thử nghiệm.

Đối với phương pháp thực nghiệm, đề án sử dụng dữ liệu từ các bài kiểm tra trắc nghiệm của học sinh để ước lượng tham số Rasch và năng lực từng cá nhân. Ngôn ngữ R được sử dụng để xây dựng quy trình xử lý dữ liệu, bao gồm hiệu chuẩn độ khó câu hỏi, ước lượng năng lực theo từng giai đoạn và triển khai thuật toán rule-based để lựa chọn câu hỏi phù hợp trong vòng ôn tập thích ứng. Thử nghiệm được thực hiện theo hai bước: (i) ước lượng năng lực ban đầu từ bài sát hạch cố định và (ii) tổ chức vòng ôn tập thích ứng dựa trên giá trị năng lực cập nhật, từ đó thu thập dữ liệu phục vụ phân tích.

Cuối cùng, hiệu quả của hệ thống được đánh giá kết hợp giữa chỉ số định lượng và định tính. Các chỉ số định lượng bao gồm sự thay đổi năng lực ($\Delta\theta$), điểm kỳ vọng, tỷ lệ cải thiện, độ chính xác phân bổ câu hỏi và độ phủ nội dung. Các kiểm định thống kê như Shapiro-Wilk, paired t-test và Cohen’s d được sử dụng để xác định mức độ ý nghĩa và quy mô ảnh hưởng của những thay đổi quan sát được. Song song với đó, phản hồi từ học sinh và giáo viên được thu thập nhằm đánh giá mức độ phù hợp của nội dung ôn tập và trải nghiệm sử dụng hệ thống. Kết quả tổng hợp là cơ sở để đề xuất các hướng tối ưu hóa mô hình trong những nghiên cứu tiếp theo.

CHƯƠNG 1: THỰC TRẠNG VÀ PHÂN TÍCH

1.1. Thực trạng hệ thống đánh giá đề thi

Các bài kiểm tra định kỳ với cấu trúc đề thi cố định và phương thức chấm điểm quen thuộc và là hình thức chỉnh để đánh giá kết quả học tập tại các các trường THPT hiện nay. Cách làm này đáp ứng tốt yêu cầu kiểm tra kiến thức cốt lõi, tuy nhiên chưa khai thác sâu dữ liệu phản hồi của học sinh và chưa hỗ trợ nhiều cho đánh giá theo năng lực. Với chương trình giáo dục phổ thông mới trong bối cảnh nhấn mạnh phân hóa và cá nhân hóa học tập, nhu cầu cải tiến hệ thống đánh giá trở nên rõ rệt hơn, đặc biệt ở các môn có lượng câu hỏi lớn như Toán.

Việc phân tích thực trạng hiện nay cho thấy hai thành tố chính ảnh hưởng đến chất lượng đánh giá: (1) Cách tổ chức quy trình thi truyền thống, và (2) Mức độ ứng dụng công nghệ trong xây dựng - quản lý câu hỏi và phân tích dữ liệu.

Hai thành tố này được làm rõ trong các mục dưới đây để cung cấp cái nhìn cụ thể hơn về hệ thống đánh giá đang được áp dụng tại nhiều trường học.

1.1.1. Phương pháp truyền thống

Phần lớn bài kiểm tra ở trường phổ thông được thiết kế theo cách thức quen thuộc: đề thi xây dựng trước, áp dụng đồng thời cho toàn bộ học sinh trong lớp hoặc trong khối. Đề thi thường được biên soạn theo ma trận nội dung do tổ bộ môn thống nhất, dựa trên chương trình đã học và mức độ nhận thức quy định.

Cách thực hiện này thuận lợi về mặt tổ chức và dễ đảm bảo tính đồng bộ giữa các lớp. Tuy vậy, kinh nghiệm của giáo viên ra đề quyết định tính phân hóa của đề thi. Khi số lượng câu hỏi lớn hoặc số lớp nhiều, việc kiểm soát độ khó tương đối giữa các đề có thể gặp khó khăn. Bên cạnh đó, kết quả thu được chủ yếu là điểm tổng hoặc số câu đúng, khiến việc nhận diện những chủ đề học sinh còn gặp khó khăn chưa thật sự thuận lợi.

Một đặc điểm nữa của cách đánh giá truyền thống là dữ liệu thu được sau mỗi kỳ kiểm tra chưa được khai thác một cách hệ thống để cải thiện chất lượng đề thi cho các năm học tiếp theo. Mức độ điều chỉnh câu hỏi thường dựa nhiều vào nhận định

chuyên môn của giáo viên hơn là các phân tích thống kê về hành vi trả lời của học sinh.

1.1.2. Các ứng dụng công nghệ hiện có

Một số trường đã triển khai phần mềm thi trực tuyến hoặc hệ thống quản lý câu hỏi, góp phần giảm tải công việc thủ công trong khâu tạo đề và chấm điểm. Các nền tảng này hỗ trợ tốt thao tác vận hành như trộn mã đề, thu kết quả ngay sau khi làm bài hoặc thống kê cơ bản về điểm số. Tuy nhiên, nhiều hệ thống mới chủ yếu phục vụ mục tiêu tổ chức thi, chưa hỗ trợ phân tích sâu dữ liệu phản hồi theo từng độ khó câu hỏi. Các thông tin như tỷ lệ lựa chọn theo từng phương án, mức độ khó tương đối của câu hỏi, khả năng phân loại hoặc xu hướng sai phổ biến của học sinh chưa được xử lý đầy đủ. Điều này khiến giáo viên khó có cơ sở dữ liệu rõ ràng để đánh giá chất lượng câu hỏi hoặc chuẩn hóa kho đề một cách hệ thống.

Kho câu hỏi trong nhiều nền tảng cũng chưa được gán tham số đo lường cụ thể. Việc phân loại câu hỏi theo mức độ khó thường dựa trên kinh nghiệm chuyên môn, dẫn đến sự khác biệt giữa các đề thi hoặc giữa các năm học. Một số hệ thống cho phép tạo đề ngẫu nhiên, nhưng năng lực của một vài học sinh lại chưa thể đáp ứng được độ khó, đặc biệt trong các giai đoạn ôn tập.

Nhìn chung, ứng dụng công nghệ đã hỗ trợ tốt khâu tổ chức thi, nhưng khả năng phân tích dữ liệu và chuẩn hóa câu hỏi còn hạn chế. Đây là khoảng trống quan trọng trước khi có thể tiến tới các phương pháp đánh giá hiện đại như mô hình Rasch hay các hệ thống học tập thích ứng.

1.1.3. Hiệu quả và hạn chế

Sự kết hợp giữa phương pháp kiểm tra truyền thống và các ứng dụng công nghệ hiện có đã góp phần tạo ra một hệ thống đánh giá tương đối ổn định trong bối cảnh triển khai tại các trường THPT. Việc tổ chức thi, chấm điểm và tổng hợp kết quả nhìn chung thuận lợi, giúp giáo viên và học sinh duy trì nhịp độ dạy - học quen thuộc. Ở những nơi áp dụng phần mềm thi trực tuyến, thời gian xử lý kết quả được rút ngắn và khả năng quản lý điểm số cũng thuận tiện hơn so với trước đây. Tuy nhiên, những hiệu quả này chủ yếu tập trung vào khâu vận hành, còn những yêu cầu

liên quan đến chất lượng phân tích dữ liệu và mức độ cá nhân hóa vẫn chưa được đáp ứng đầy đủ. Hạn chế dễ nhận thấy là dữ liệu phản hồi sau bài kiểm tra chưa được khai thác sâu để nhận diện độ khó tương đối của câu hỏi hoặc đánh giá xu hướng làm bài của học sinh. Khi thông tin chi tiết ở độ khó câu hỏi không được phân tích, việc điều chỉnh đề thi giữa các kỳ kiểm tra thường chỉ dựa vào kinh nghiệm chuyên môn, dẫn đến sự không đồng đều về độ khó hoặc mức độ phân hóa. Bên cạnh đó, công nghệ hiện có chưa hỗ trợ mạnh cho việc tạo đề thi dựa trên mức độ thành thạo của từng học sinh. Nội dung ôn tập vẫn chủ yếu cố định hoặc sinh tự động nhưng không gắn với năng lực, khiến việc hỗ trợ học sinh tiến bộ theo từng chủ đề còn hạn chế. Trong khi đó, nhu cầu điều chỉnh độ khó và phân bổ nội dung phù hợp với trình độ cá nhân ngày càng rõ hơn khi chương trình giáo dục mới nhấn mạnh học tập phân hóa.

Từ những điểm trên, có thể thấy hệ thống đánh giá hiện tại đáp ứng tốt yêu cầu tổ chức thi nhưng còn nhiều dư địa để khai thác dữ liệu phản hồi theo hướng khoa học hơn. Đây chính là khoảng trống cần được phân tích kỹ trong các mục tiếp theo, đặc biệt liên quan đến cách thức thu thập, tổ chức và sử dụng dữ liệu độ khó câu hỏi trong quá trình đánh giá.

1.2. Phân tích dữ liệu phản hồi từ đề thi

Từ dữ liệu học sinh làm các bài kiểm tra trắc nghiệm, nếu được tổng hợp và phân tích ở độ khó câu hỏi, việc đánh giá chất lượng đề thi và nhận diện xu hướng học tập có thể đạt được nhiều thông tin quan trọng. Trong các nghiên cứu về đo lường giáo dục, dữ liệu độ khó câu hỏi được xem là thành phần cốt lõi cho việc xác định độ khó, đánh giá mức độ phù hợp của câu hỏi và đảm bảo tính ổn định của phép đo [1], [2], [5]. Tuy nhiên, tại nhiều trường THPT, dữ liệu thu được sau mỗi bài kiểm tra chủ yếu ở mức tổng hợp, nên các phân tích nâng cao khó được triển khai thường xuyên.

Trong bối cảnh yêu cầu đổi mới đánh giá theo năng lực, việc xem xét cách thức tổng hợp và phân tích dữ liệu phản hồi trở nên quan trọng. Các mục dưới đây trình bày thực trạng thu thập dữ liệu, khả năng ứng dụng các mô hình phân tích hiện đại như Rasch, và mức độ phân loại câu hỏi đang được áp dụng tại một số cơ sở giáo dục.

1.2.1. Thu thập và tổng hợp dữ liệu

Thông qua phiếu trả lời trắc nghiệm hoặc hệ thống thi trực tuyến, dữ liệu phản hồi từ các bài kiểm tra có thể được tổng hợp. Thông tin lưu trữ chủ yếu bao gồm điểm tổng, số câu làm đúng và mã đề. Một số phần mềm có cung cấp thống kê tỷ lệ đúng - sai theo từng câu hỏi, cái nhìn tổng quan về kết quả lớp học sẽ được cung cấp trực quan tới giáo viên. Tuy nhiên, dữ liệu độ khó câu hỏi chi tiết của từng học sinh nhiều khi chưa được lưu trữ đầy đủ, khiến việc phân tích sâu về đặc tính câu hỏi gặp khó khăn. Điều này trái ngược với yêu cầu của các phương pháp đo lường hiện đại, vốn nhấn mạnh tầm quan trọng của dữ liệu phản hồi chi tiết để đảm bảo độ ổn định của tham số câu hỏi theo thời gian [1], [5]. Trong trường hợp dữ liệu được tổng hợp thủ công, khối lượng công việc lớn khiến việc phân tích nâng cao khó triển khai một cách thường xuyên. Ngoài ra, dữ liệu thường được lưu trữ theo từng kỳ kiểm tra riêng lẻ, chưa hình thành chuỗi thời gian dài để theo dõi sự thay đổi của câu hỏi hoặc mức độ tiến bộ của học sinh. Khi thiếu dữ liệu lịch sử, việc so sánh độ khó giữa các năm học hoặc chuẩn hóa kho câu hỏi trở nên thách thức. Khả năng áp dụng bị hạn chế đối với các mô hình phân tích tham số, vốn cần dữ liệu độ khó câu hỏi được thu thập đầy đủ và nhất quán [2].

1.2.2. Ứng dụng mô hình Rasch

Mô hình Rasch được xem là một trong những phương pháp đo lường có khả năng chuẩn hóa độ khó câu hỏi và ước lượng năng lực người học dựa trên dữ liệu độ khó câu hỏi [1], [3]. Mặc dù các nghiên cứu quốc tế đã khẳng định vai trò của mô hình này trong đánh giá giáo dục, thực tế triển khai tại các trường phổ thông còn hạn chế do yêu cầu về dữ liệu, phần mềm và năng lực phân tích.

Nhiều hệ thống thi trực tuyến hiện nay mới chỉ cung cấp thống kê mô tả, chưa hỗ trợ các chỉ số đặc trưng như độ khó (b) hay chỉ số phù hợp mô hình (fit). Khi không có các tham số này, giáo viên khó có cơ sở định lượng để xem xét chất lượng câu hỏi hoặc điều chỉnh đề thi một cách khách quan. Trong khi đó, các tài liệu đo lường giáo dục đều nhấn mạnh rằng mô hình Rasch chỉ hoạt động hiệu quả khi dữ liệu độ khó câu hỏi được thu thập đầy đủ và phản ánh đúng hành vi trả lời của người

học [1], [4]. Ngoài ra, việc áp dụng Rasch đòi hỏi sử dụng các phần mềm hoặc thư viện thống kê chuyên dụng, như R hoặc các công cụ mô hình hóa item-level. Đây là những công cụ chưa phổ biến rộng trong giáo dục phổ thông, dẫn đến việc triển khai còn mang tính thử nghiệm ở quy mô nhỏ. Tuy nhiên, khi dữ liệu được thu thập đầy đủ và ổn định, mô hình Rasch có thể hỗ trợ đáng kể việc chuẩn hóa ngân hàng câu hỏi và ước lượng năng lực học sinh theo một thang đo thống nhất [3].

1.2.3. Phân loại câu hỏi

Phân loại câu hỏi theo chủ đề và mức độ nhận thức là bước quan trọng trong xây dựng ma trận đề thi. Ở nhiều trường, việc phân loại này chủ yếu dựa trên phân phối nội dung chương trình và kinh nghiệm giảng dạy của giáo viên. Cách tiếp cận này giúp đề thi phù hợp với chuẩn kiến thức, nhưng mức độ phân hóa hoặc độ khó của câu hỏi vẫn mang tính cảm quan. Một số cơ sở sử dụng tỷ lệ đúng - sai từ các bài kiểm tra trước để đánh giá độ khó tương đối. Tuy nhiên, tỷ lệ đúng không phản ánh đầy đủ đặc tính của câu hỏi vì nó phụ thuộc vào đối tượng học sinh cụ thể. Trong khi đó, các nghiên cứu về đo lường nhấn mạnh rằng độ khó chỉ thực sự có ý nghĩa khi được ước lượng theo thang đo chung và duy trì tính ổn định qua nhiều nhóm người học [1], [3], [5].

Ngoài ra, sự thiếu nhất quán trong việc cập nhật kho câu hỏi khiến một số câu hỏi không còn phù hợp với chương trình hoặc cấu trúc đề thi hiện tại. Khi chưa có cơ chế đánh giá khách quan như tham số độ khó hoặc mức độ phù hợp theo mô hình, việc loại bỏ hay giữ lại câu hỏi chủ yếu dựa trên kinh nghiệm cá nhân. Chất lượng phân hóa của đề thi có thể bị ảnh hưởng. Việc phân loại câu hỏi dựa trên dữ liệu độ khó câu hỏi và tham số độ khó, nếu được áp dụng một cách hệ thống, có thể giúp ngân hàng câu hỏi trở nên ổn định hơn. Các nghiên cứu về mô hình Rasch cũng cho thấy việc duy trì độ ổn định của tham số câu hỏi là điều kiện quan trọng để đảm bảo tính khách quan của phép đo [4], [5]. Đây là nền tảng quan trọng cho việc xây dựng hệ thống đánh giá thích ứng ở các chương sau.

1.3. Đánh giá các thách thức hiện tại

Từ phân tích thực trạng thu thập dữ liệu và tổ chức đánh giá trong các mục trước, có thể thấy khả năng triển khai các phương pháp đo lường hiện đại cũng như xây dựng hệ thống đánh giá theo năng lực đang bị ảnh hưởng bởi nhiều yếu tố. Những thách thức này không chỉ liên quan đến dữ liệu và hạ tầng kỹ thuật mà còn xuất phát từ đặc thù ngôn ngữ, văn hóa học tập và điều kiện triển khai tại các cơ sở giáo dục. Khi các phương pháp dựa trên dữ liệu độ khó câu hỏi, như mô hình Rasch, yêu cầu thông tin phản hồi chi tiết và ổn định [1], [2], sự thiếu hụt dữ liệu hoặc hạn chế trong môi trường triển khai có thể làm giảm hiệu quả của quá trình phân tích. Những yếu tố ảnh hưởng trực tiếp đến tiềm lực áp dụng các mô hình đánh giá tiên tiến trong bối cảnh phổ thông được thể tóm lực qua ba nhóm thách thức chính dưới đây.

1.3.1. Thiếu dữ liệu phản hồi đa dạng

Dữ liệu phản hồi chưa được thu thập đầy đủ và chưa phản ánh đủ sự đa dạng của đối tượng học sinh là thách thức vô cùng lớn. Trong thực tế, nhiều trường chỉ lưu trữ điểm tổng mà không lưu thông tin chi tiết về từng câu hỏi. Trong khi đó, các nghiên cứu về đo lường giáo dục cho thấy tham số độ khó và mức độ phù hợp của câu hỏi chỉ ổn định khi dữ liệu độ khó câu hỏi được thu thập từ các nhóm học sinh khác nhau và qua nhiều lần kiểm tra [1], [5].

Khi dữ liệu chỉ giới hạn trong một vài bài kiểm tra của một lớp cụ thể, việc phân tích tham số câu hỏi hoặc theo dõi sự ổn định theo thời gian trở nên khó khăn. Mức độ đa dạng về năng lực - một yếu tố quan trọng để ước lượng tham số chính xác - cũng chưa được đảm bảo. Điều này ảnh hưởng trực tiếp đến khả năng phân loại câu hỏi, xây dựng ngân hàng đề chuẩn hóa hoặc triển khai mô hình đánh giá thích ứng.

Bên cạnh đó, dữ liệu của nhiều năm thường không được tổ chức thành một hệ thống liên tục, khiến việc đánh giá lại câu hỏi hoặc so sánh độ khó giữa các năm học chưa thực sự khả thi. Trong bối cảnh yêu cầu phân tích dựa trên dữ liệu độ khó câu hỏi ngày càng được nhấn mạnh [2], điều này trở thành một rào cản đáng kể.

1.3.2. Hạn chế về hạ tầng công nghệ

Cơ sở hạ tầng phục vụ cho công nghệ tại các cơ sở giáo dục phổ thông có sự phân khóa đáng kể giữa các khu vực. Ở một số nơi, hệ thống thi trực tuyến đã được triển khai khá đầy đủ, trong khi nhiều trường vẫn dựa trên quy trình thi truyền thống hoặc sử dụng các phần mềm với chức năng còn hạn chế. Điều này dẫn đến các câu hỏi trong kho thiếu sự không đồng đều trong khả năng thu thập dữ liệu độ khó và quản lý.

Một số nền tảng hiện có chỉ hỗ trợ thống kê mô tả, mức độ phù hợp hoặc ước lượng tham số độ khó của câu hỏi theo mô hình Rasch chưa cung cấp thông công cụ phân tích dữ liệu nâng cao. Trong khi các nghiên cứu cho thấy các tham số này là yếu tố nòng cốt trong việc chuẩn hóa độ khó và đảm bảo tính khách quan của phép đo [1], [3], chúng vẫn chưa được tích hợp rộng rãi trong các phần mềm đang được sử dụng. Ngoài ra, việc triển khai các mô hình phân tích như Rasch thường đòi hỏi sử dụng phần mềm chuyên dụng (ví dụ: R hoặc các ứng dụng chuyên sâu), nằm ngoài mức độ quen thuộc của phần lớn giáo viên phổ thông. Điều này khiến quá trình phân tích dữ liệu sâu chưa thể thực hiện rộng rãi, dù về nguyên tắc, dữ liệu độ khó câu hỏi có thể cung cấp nhiều thông tin giá trị.

1.3.3. Đặc thù ngôn ngữ và văn hóa

Mặc dù ngôn ngữ giảng dạy chính trong nhà trường là tiếng Việt chuẩn theo quy định chung, cách diễn đạt trong câu hỏi trắc nghiệm vẫn có thể ảnh hưởng đến khả năng hiểu của học sinh. Một số câu hỏi có cấu trúc dài, chứa nhiều bước suy luận hoặc bao gồm các biểu thức toán học phức tạp có thể gây khó khăn cho học sinh, không phải vì nội dung kiến thức mà do cách trình bày chưa thật sự rõ ràng. Điều này có thể làm thay đổi tỷ lệ trả lời đúng và ảnh hưởng đến khả năng đánh giá độ khó thực sự của câu hỏi.

Các nghiên cứu về đo lường giáo dục cho thấy rằng chất lượng của câu hỏi không chỉ phụ thuộc vào nội dung mà còn liên quan đến mức độ mạch lạc và tính nhất quán trong cách diễn đạt [1], [4]. Một câu hỏi có thể dễ với học sinh đã quen với dạng biểu đạt đó nhưng lại khó hơn đối với những học sinh ít tiếp xúc với cách trình

bày tương tự. Vì vậy, sự khác biệt không nằm ở vùng miền hay hệ ngôn ngữ, mà xuất phát từ phong cách học tập, mức độ thành thạo trong đọc hiểu toán học và kinh nghiệm làm bài của từng nhóm học sinh.

Ngoài ra, một số học sinh có xu hướng trả lời dựa trên mẹo làm bài hoặc kinh nghiệm suy đoán, thay vì suy luận hoàn toàn theo kiến thức. Điều này có thể làm sai lệch khả năng phân hóa của câu hỏi nếu không có dữ liệu độ khó câu hỏi đầy đủ để phân tích. Trong trường hợp chưa có cơ chế đánh giá khách quan như kiểm định độ phù hợp của câu hỏi, việc điều chỉnh hoặc thay thế các độ khó câu hỏi này phần lớn dựa trên kinh nghiệm của giáo viên. Khía cạnh này cho thấy việc thiết kế câu hỏi cần chú trọng đến sự rõ ràng trong diễn đạt, mức độ cô đọng của dữ kiện và cấu trúc logic của yêu cầu bài toán. Đây là yếu tố quan trọng để đảm bảo câu hỏi phản ánh đúng năng lực cần đánh giá và phù hợp với các khuyến nghị về kiểm định chất lượng độ khó câu hỏi theo hướng tiếp cận Rasch [4], [5].

1.4. Kết chương

Bức tranh tổng quan về hệ thống đánh giá đề thi hiện nay tại các trường THPT đã được khái quát, tập trung vào ba khía cạnh chính: cách thức tổ chức kiểm tra truyền thống, mức độ ứng dụng công nghệ và thực trạng thu thập - phân tích dữ liệu phản hồi. Qua phân tích cho thấy, các quy trình hiện hành nhìn chung đáp ứng tốt yêu cầu vận hành và kiểm tra định kỳ, nhưng còn nhiều hạn chế khi hướng tới mục tiêu đánh giá theo năng lực và cá nhân hóa học tập.

Trong đo lường hiện đại, độ khó câu hỏi nắm vai trò cốt lõi [1], [2], dù vậy, vẫn chưa được thu thập một cách đầy đủ và nhất quán. Bên cạnh đó, hạ tầng công nghệ và công cụ phân tích còn hạn chế, khiến các phương pháp ước lượng tham số câu hỏi hoặc đánh giá mức độ phù hợp theo mô hình Rasch chưa được áp dụng rộng rãi. Việc phân loại câu hỏi và chuẩn hóa độ khó chủ yếu dựa vào kinh nghiệm chuyên môn, chưa dựa trên thang đo thống nhất - trong khi nhiều tài liệu quốc tế nhấn mạnh tính ổn định của tham số là điều kiện thiết yếu cho đánh giá khách quan [4], [5].

Dữ liệu, công nghệ và đặc thù thiết kế câu hỏi cho thấy nhu cầu cấp thiết cần vận dụng các cơ sở lý luận chặt chẽ hơn trong đo lường giáo dục đã được nhận diện

là thách thức. Đây là nền tảng quan trọng để xây dựng hệ thống đánh giá có khả năng phản ánh chính xác mức độ thành thạo của từng học sinh và hỗ trợ quá trình học tập một cách hiệu quả hơn.

Chương tiếp theo sẽ trình bày cơ sở lý luận và các mô hình đo lường hiện đại, trong đó Lý thuyết Đáp ứng Câu hỏi và mô hình Rasch đóng vai trò trung tâm. Những khái niệm này là cơ sở quan trọng để thiết kế hệ thống đánh giá và quy trình phân tích dữ liệu được đề xuất trong chương 3.

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

Lý thuyết Đáp ứng Câu hỏi (Item Response Theory - IRT) là một hướng tiếp cận hiện đại trong đo lường giáo dục, được vận dụng nhằm mô phỏng mối tương quan giữa năng lực của người làm bài và xác suất trả lời đúng từng câu hỏi trong bài kiểm tra. Khác biệt với cách đánh giá truyền thống vốn dựa chủ yếu vào tổng điểm nơi mọi câu hỏi được xem như có mức độ đóng góp bằng nhau trong khi đối với IRT xem mỗi câu hỏi là một "đơn vị đo lường" với các tham số riêng. Nhờ đó, lý thuyết này cho phép phân tích sâu hơn về chất lượng câu hỏi, mức độ phù hợp của đề thi và khả năng phân loại của các độ khó câu hỏi [2], [3]. IRT không chỉ đánh giá học sinh thông qua điểm tổng mà còn mô hình hóa cách các học sinh phản hồi trước từng câu hỏi, qua đó đạt được mức độ chính xác cao hơn trong việc ước lượng năng lực.

Sự phát triển của IRT gắn liền với các công trình nền tảng từ giữa thế kỷ XX. Những nghiên cứu tiên phong của Georg Rasch vào cuối thập niên 1950 và đầu thập niên 1960 đặt nền móng quan trọng cho lĩnh vực này. Trong tác phẩm xuất bản năm 1960, Rasch giới thiệu mô hình một tham số (1PL), trong đó xác suất trả lời đúng được mô tả thông qua tương quan giữa năng lực người học và độ khó của câu hỏi [1]. Điểm nổi bật trong đóng góp của Rasch là đề xuất thang đo chung cho người và câu hỏi, tạo điều kiện cho việc so sánh các bài kiểm tra hoặc nhóm học sinh khác nhau trên cùng một không gian đo lường.

Đến thập niên 1980, Hambleton và Swaminathan tiếp tục mở rộng khung lý thuyết IRT bằng việc phát triển các mô hình nhiều tham số như 2PL và 3PL. Các mô hình này mô tả thêm những khía cạnh khác của câu hỏi, chẳng hạn độ phân biệt (đặc trưng phản ánh khả năng phân loại học sinh có năng lực khác nhau) hoặc xác suất đoán mò [2]. Những nghiên cứu này đã củng cố cơ sở toán học của IRT và mở đường cho việc ứng dụng trong các kỳ thi quy mô lớn cũng như các hệ thống thi dựa trên máy tính.

Bước sang đầu những năm 2000, Baker (2001) đã hệ thống hóa lý thuyết IRT một cách toàn diện, cung cấp nền tảng kiến thức vững chắc cho việc đào tạo và ứng dụng IRT trong nhiều lĩnh vực đo lường giáo dục khác nhau [3]. Các công trình này giúp định hình IRT như một phương pháp có tính khái quát cao, không chỉ áp dụng trong kiểm tra học tập mà còn mở rộng sang phân tích tâm lý, xây dựng thang đo hành vi phản hồi và các nghiên cứu liên quan.

Qua nhiều giai đoạn phát triển, IRT đã chứng minh tầm quan trọng của mình trong lĩnh vực đánh giá giáo dục hiện đại. Nhờ khả năng mô hình hóa hành vi trả lời chi tiết và xây dựng thang đo năng lực ổn định, IRT đã được áp dụng trong nhiều kỳ thi có độ tin cậy cao như các kỳ đánh giá quốc tế, các hệ thống thi dựa trên máy tính hoặc các ngân hàng câu hỏi có tham số rõ ràng. Mặc dù mức độ triển khai còn phụ thuộc điều kiện từng quốc gia và cơ sở giáo dục, so với cách tiếp cận truyền thống, IRT vẫn được xem là một trong những nền tảng quan trọng hỗ trợ phân tích dữ liệu phản hồi và đo lường năng lực người học theo hướng chính xác và linh hoạt hơn.

2.1.2. Các mô hình IRT

Trong Lý thuyết Đáp ứng Câu hỏi (IRT), mỗi câu hỏi được mô hình hóa thông qua các tham số phản ánh đặc tính cơ bản của câu hỏi đối với người làm bài. Dựa trên số lượng và vai trò của các tham số này, ba mô hình được phát triển từ IRT được biết đến rộng rãi là mô hình một tham số (1PL), hai tham số (2PL) và ba tham số (3PL) [2], [3]. Mỗi mô hình có mức độ phức tạp riêng, phù hợp với những mục tiêu phân tích và điều kiện dữ liệu khác nhau.

Mô hình một tham số (1PL)

Mô hình một tham số, thường được biết đến với tên gọi mô hình Rasch, xem độ khó là tham số duy nhất của câu hỏi [1], với giả định rằng tất cả các câu hỏi có khả năng phân biệt tương đương và không xét đến yếu tố đoán mò. Nhờ cấu trúc đơn giản, mô hình Rasch có ưu điểm là dễ diễn giải và yêu cầu dữ liệu đầu vào không quá phức tạp. Hơn nữa, mô hình đặt người học và câu hỏi trên cùng một thang đo, tạo điều kiện cho việc so sánh và sử dụng kết quả trong nhiều bối cảnh đánh giá khác

nhau. Chính vì vậy, dù đơn giản, mô hình 1PL vẫn được sử dụng phổ biến trong nghiên cứu đo lường giáo dục và trong nhiều hệ thống đánh giá có quy mô vừa.

Mô hình hai tham số (2PL)

Mô hình hai tham số mở rộng mô hình Rasch bằng cách bổ sung tham số phân biệt, phản ánh mức độ mà một câu hỏi có thể tách học sinh có năng lực khác nhau thành các nhóm rõ rệt [2]. Nhờ thêm tham số này, mô hình cho phép mô tả hành vi của câu hỏi một cách chi tiết hơn, đặc biệt đối với những câu có mức độ phân biệt cao hoặc thấp so với mặt bằng chung. Tuy nhiên, để kết quả có thể tin cậy, việc ước lượng tham số trong mô hình 2PL thường yêu cầu kích thước mẫu lớn hơn, thuật toán tính toán phức tạp hơn và dữ liệu có chất lượng ổn định. Vì vậy, mô hình này thường được áp dụng trong các nghiên cứu hoặc hệ thống đánh giá có điều kiện dữ liệu đầy đủ.

Mô hình ba tham số (3PL)

Mô hình ba tham số tiếp tục bổ sung tham số mô tả xác suất đoán mò, tức là khả năng người học trả lời đúng câu hỏi ngay cả khi không nắm vững kiến thức. Tham số này đặc biệt quan trọng đối với các dạng câu hỏi trắc nghiệm lựa chọn đáp án, trong đó học sinh có thể chọn đáp án đúng một cách ngẫu nhiên [3]. Nhờ khả năng mô tả hành vi đoán mò, mô hình 3PL mang tính linh hoạt cao và thường được sử dụng ở các bài thi có phạm vi lớn hoặc trong phát triển kho câu hỏi có yêu cầu cao về độ chính xác. Tuy nhiên, việc ước lượng ba tham số đồng thời đòi hỏi dữ liệu phản hồi có kích thước lớn, ổn định và cần đến các thuật toán tính toán chuyên sâu.

Nhìn chung, các mô hình trong IRT cung cấp các mức độ mô tả khác nhau về câu hỏi và năng lực người học. Mô hình càng phức tạp thì khả năng mô phỏng hành vi phản hồi càng cao, nhưng đi kèm với yêu cầu dữ liệu và năng lực xử lý lớn hơn. Tùy thuộc mục tiêu của kỳ đánh giá, nguồn lực công nghệ và chất lượng dữ liệu phản hồi, các cơ sở giáo dục hoặc nhóm nghiên cứu có thể chọn mô hình phù hợp để phân tích chất lượng câu hỏi hoặc phát triển các kho câu hỏi với tham giá trị rõ ràng. IRT vì vậy tạo ra một khung đo lường linh hoạt, cho phép mô tả năng lực người học theo

hướng chính xác và khách quan vượt trội hơn các cách tiếp cận truyền thống vốn chỉ dựa vào tổng điểm.

2.2. Mô hình Rasch và ứng dụng trong giáo dục

Mô hình Rasch là một trong những mô hình cơ bản thuộc khung lý thuyết Đáp ứng Câu hỏi (IRT), được xem như nền tảng quan trọng trong lĩnh vực đo lường giáo dục hiện đại. Xuất phát từ công trình do Georg Rasch phát triển (1960) [1], mô hình này hướng tới việc xây dựng thang đo khách quan, trong đó năng lực của người học cùng với độ khó của câu hỏi được đặt trên cùng một trục đo chung.

Trong bối cảnh giáo dục phổ thông, Rasch đặc biệt hữu ích khi cần chuẩn hóa ngân hàng câu hỏi, ước lượng độ khó ổn định và hỗ trợ các hình thức kiểm tra thích ứng. Các mục sau trình bày nguyên tắc hoạt động của mô hình, ưu điểm - hạn chế và những ứng dụng phổ biến trong thực tiễn giáo dục.

2.2.1. Nguyên tắc hoạt động

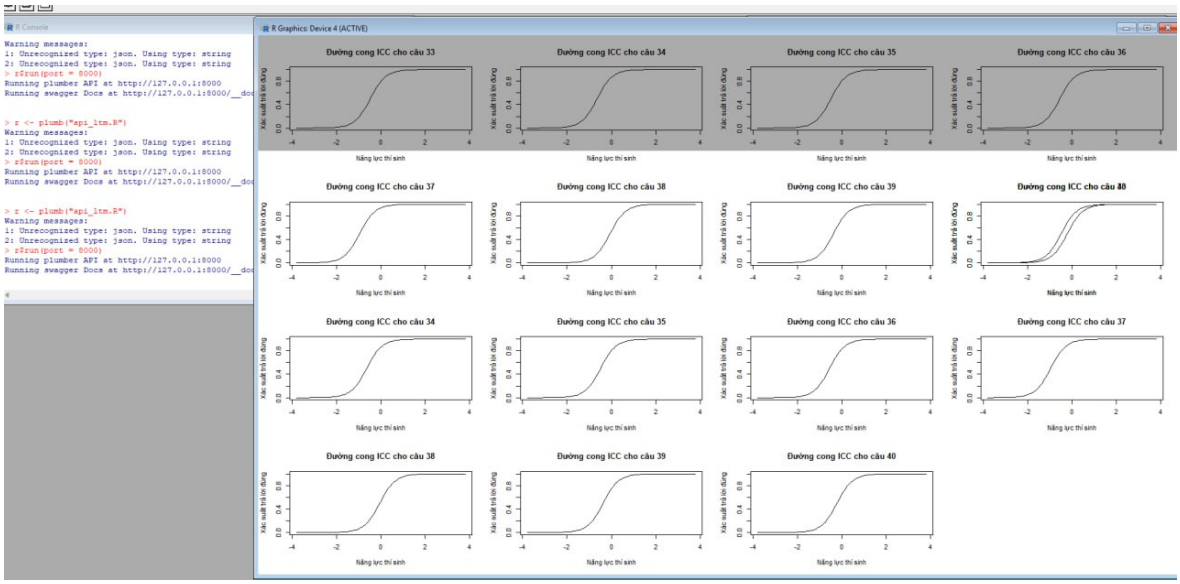
Mô hình Rasch mô tả xác suất trả lời đúng một câu hỏi dựa trên chênh lệch giữa năng lực người học (θ) và độ khó độ khó câu hỏi (b). Công thức của mô hình logistic một tham số được biểu diễn như sau:

$$P(\theta, b) = \frac{e^{(\theta - b)}}{1 + e^{(\theta - b)}} \quad (2.1)$$

Trong đó:

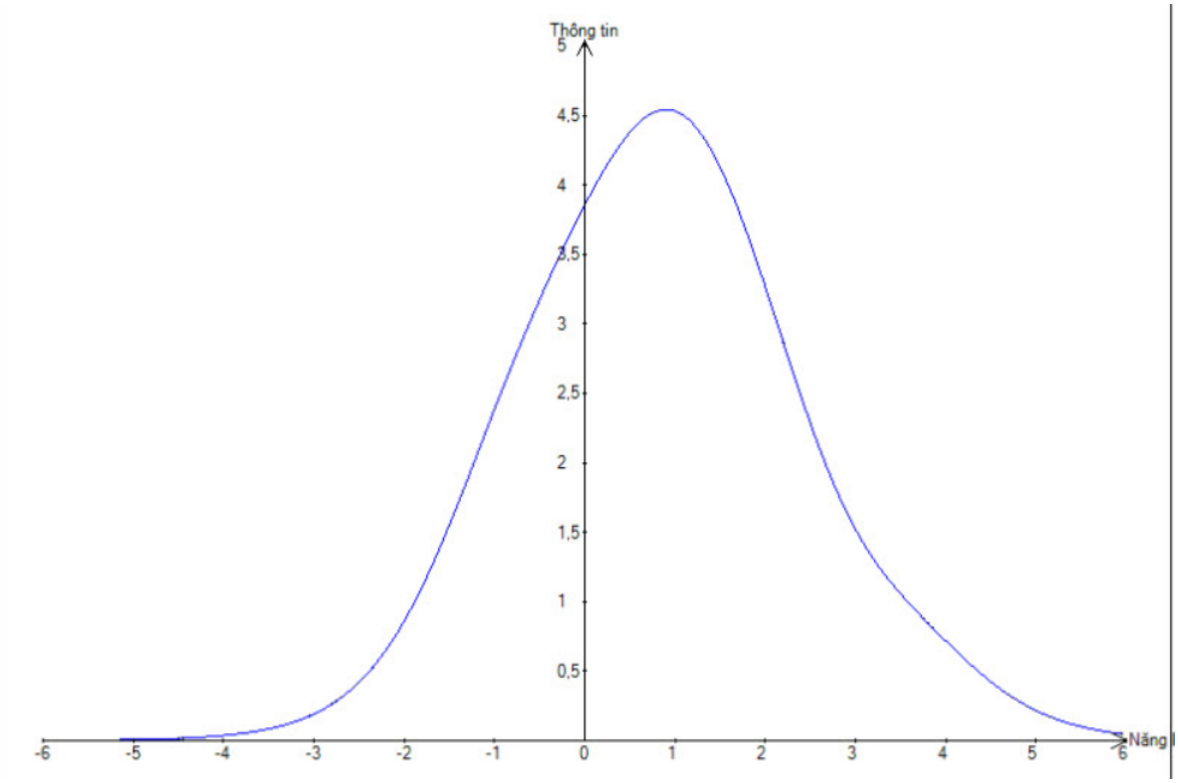
- θ biểu thị năng lực của học sinh trên thang đo chung,
- b là độ khó của câu hỏi.

Công thức cho thấy xác suất trả lời đúng tăng khi năng lực người học lớn hơn độ khó câu hỏi và ngược lại. Điểm đặc biệt của mô hình Rasch là giả định độ phân biệt của các câu hỏi là như nhau, giúp mô hình giữ được sự giản lược cần thiết mà không làm suy giảm độ chính xác của phép đo khi dữ liệu phù hợp [1].



Hình 2.1. Các đường cong ICC của các câu hỏi

Một trong những đặc trưng được nhấn mạnh trong các nghiên cứu của Hambleton & Swaminathan (1985) và Baker (2001) là tính khách quan riêng biệt (specific objectivity), nghĩa là việc so sánh độ khó giữa các câu hỏi không phụ thuộc vào nhóm học sinh cụ thể, và việc so sánh năng lực của học sinh không phụ thuộc vào bộ câu hỏi cụ thể [2], [3]. Điều này chỉ đạt được khi phản hồi dữ liệu tương thích với mô hình. Ngoài ra, Rasch sử dụng các chỉ số kiểm định mức độ phù hợp (fit indices) để xác định xem một câu hỏi có phản ánh đúng hành vi dự kiến hay không. Đây là cơ sở giúp lọc bỏ các câu hỏi bất thường và hình thành kho câu hỏiổn định, thỏa mãn yêu cầu đánh giá theo năng lực [4].



Hình 2.2. Đồ thị hàm thông tin bài thi

2.2.2. Ưu điểm và hạn chế

Ưu điểm:

- Đơn giản và dễ chuẩn hóa: Với chỉ một tham số độ khó, mô hình Rasch dễ triển khai hơn so với các mô hình IRT nhiều tham số. Điều này phù hợp với bối cảnh trường phổ thông, nơi dữ liệu chưa thật sự phong phú [2].

- Xây dựng thang đo thống nhất: Rasch đặt người học và câu hỏi trên cùng một trục đo, giúp việc so sánh năng lực trở nên nhất quán, ngay cả khi các bài kiểm tra khác nhau được sử dụng [1].

- Lọc và kiểm định chất lượng câu hỏi: Các chỉ số phù hợp (fit) giúp nhận diện câu hỏi kém phân hóa hoặc không phản ánh đúng hành vi người học, qua đó cải thiện độ tin cậy của ngân hàng câu hỏi [4].

- Hỗ trợ mạnh cho kiểm tra thích ứng: Khi tham số độ khó được ước lượng ổn định, Rasch giúp lựa chọn câu hỏi phù hợp với năng lực người học, là nền tảng cho các hệ thống thi thích ứng.

Hạn chế

- Yêu cầu dữ liệu phải phù hợp mô hình: Rasch hoạt động hiệu quả khi dữ liệu phản hồi tuân theo cấu trúc logistic một tham số. Các tài liệu đo lường nhấn mạnh rằng khi câu hỏi không phù hợp, tham số ước lượng có thể thiếu ổn định [4], [5].

- Không mô tả được độ phân biệt hoặc đoán mò: Khác với 2PL hoặc 3PL, mô hình Rasch không chứa tham số mô tả độ phân biệt hoặc khả năng đoán mò, nên có thể chưa phản ánh đầy đủ các đặc tính phức tạp của câu hỏi trong một số trường hợp [3].

- Đòi hỏi quy trình kiểm định nghiêm ngặt: Việc xây dựng ngân hàng câu hỏi Rasch yêu cầu kiểm tra mức độ phù hợp liên tục, loại bỏ hoặc điều chỉnh các độ khó câu hỏi chưa đạt chuẩn, khiến quy trình triển khai cần đầu tư thời gian và công sức.

2.2.3. Ứng dụng trong giáo dục

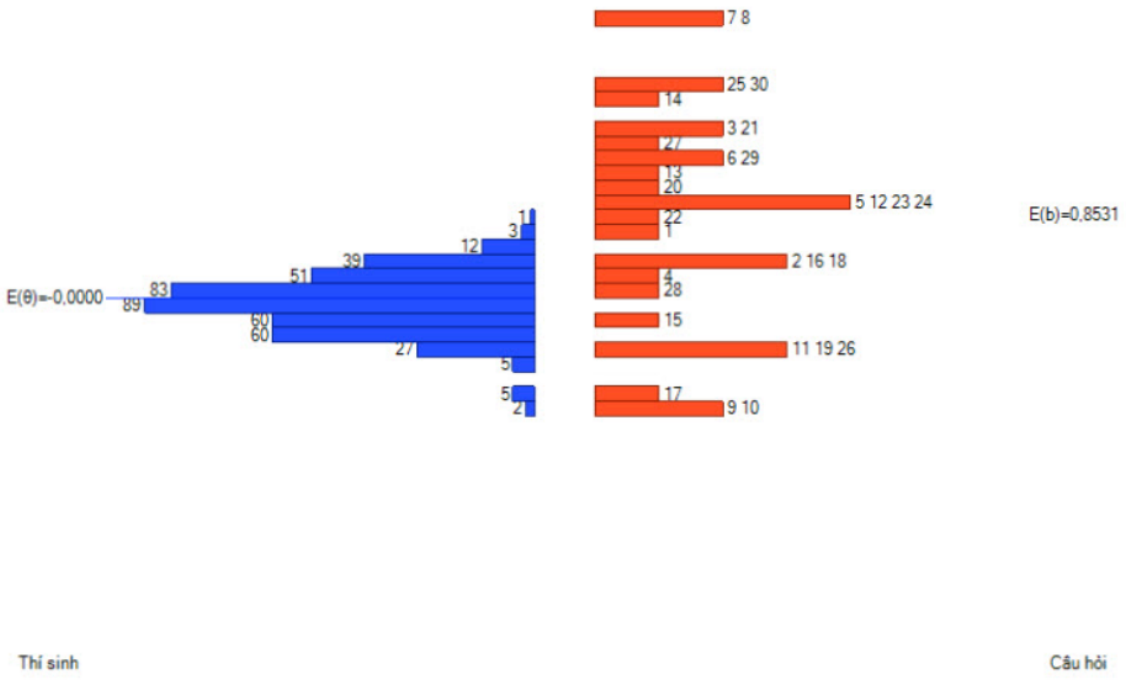
Khung mô hình Rasch được áp dụng phổ biến tại nhiều hệ thống đánh giá quốc tế để xây dựng thang đo năng lực chuẩn hóa. Trong giáo dục phổ thông, mô hình này đặc biệt hữu ích trong các hoạt động sau:

- Chuẩn hóa độ khó và hiệu chuẩn ngân hàng câu hỏi: Rasch giúp xác định độ khó của từng câu một cách ổn định, giảm sự phụ thuộc vào nhóm học sinh cụ thể. Đây là nền tảng quan trọng để phát triển kho câu hỏi lớn, sử dụng qua nhiều năm học [1], [4].

- Đánh giá năng lực trên thang đo chung: Khi câu hỏi được hiệu chuẩn, năng lực học sinh có thể được ước lượng trên trục đo liên tục, tạo điều kiện so sánh năng lực giữa các lớp hoặc giữa các kỳ thi khác nhau.

-Tối ưu hóa đề thi và phân hóa kết quả: Rasch hỗ trợ lựa chọn câu hỏi phù hợp với từng mức năng lực, giúp tăng khả năng phân loại và giảm hiện tượng “đề quá khó/đề quá dễ”.

- Nền tảng của hệ thống thi thích ứng: Các hệ thống thi thích ứng dựa trên việc chọn những câu có mức độ khó gần tương ứng với mức năng lực ước lượng hiện tại của người làm bài. Rasch, nhờ tham số đơn giản và ổn định, phù hợp để triển khai trong môi trường phổ thông [2], [3].



Hình 2.3. Biểu đồ so sánh năng lực thí sinh và độ khó câu hỏi

Trong đề án này, mô hình Rasch đóng vai trò cốt lõi trong Giai đoạn 1 (hiệu chuẩn độ khó câu hỏi) và là cơ sở để triển khai cơ chế chọn câu thích ứng trong Chương 3. Các nguyên lý của Rasch giúp đảm bảo rằng việc phân bổ nội dung ôn tập có thể được cá nhân hóa dựa trên năng lực và tiến bộ của từng học sinh.

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

Mức độ khó của câu trắc nghiệm được xem như tham số quan trọng trong đo lường giáo dục, phản ánh mức độ thách thức mà một câu hỏi đặt ra đối với người học. Trong mô hình Rasch, độ khó được xem là đặc tính nội tại của câu hỏi và được ước lượng dựa trên hành vi trả lời của nhiều nhóm thí sinh [1]. Tuy nhiên, các nghiên cứu về thiết kế câu hỏi cho thấy độ khó không chỉ phụ thuộc vào nội dung kiến thức, mà còn chịu ảnh hưởng bởi nhiều yếu tố như khả năng nhận thức của người học, cấu trúc và hình thức biểu đạt của câu hỏi, cũng như ngữ cảnh ngôn ngữ sử dụng [2], [4]. Những yếu tố này đóng vai trò quan trọng khi phát triển kho câu hỏi và kiểm định tính chất độ khó câu hỏi. Việc hiểu rõ chúng không chỉ hỗ trợ thiết kế câu hỏi phù hợp với mục tiêu đánh giá mà còn góp phần duy trì tính ổn định của tham số đo lường

theo thời gian [1], [5]. Các mục sau đây trình bày ba nhóm yếu tố thường được cân nhắc trong quá trình biên soạn và phân tích mức khó câu hỏi.

2.3.1. Khả năng nhận thức và kinh nghiệm học tập

Khả năng nhận thức của học sinh, bao gồm mức độ hiểu khái niệm, khả năng suy luận và kinh nghiệm tiếp cận dạng bài, có ảnh hưởng trực tiếp đến cách các em xử lý câu hỏi trắc nghiệm. Một câu hỏi dựa trên kiến thức cơ bản có thể được xem là dễ với những học sinh đã thành thạo kỹ năng tương ứng, nhưng lại trở nên khó với học sinh ít kinh nghiệm hoặc chưa quen với dạng yêu cầu đó.

Các nghiên cứu về Rasch chỉ ra rằng độ khó phản ánh mức độ thách thức trung bình của câu hỏi đối với một quần thể người học, nhưng giá trị này vẫn chịu tác động bởi sự phân bố năng lực của nhóm học sinh làm bài [1], [5]. Vì vậy, khi mẫu người làm bài có sự khác biệt lớn về nền tảng hoặc trình độ, tham số độ khó có thể biến động nếu không được kiểm định qua nhiều nhóm khác nhau. Ngoài ra, học sinh có nhiều kinh nghiệm với các dạng câu hỏi quen thuộc thường phát triển chiến lược làm bài nhanh hơn, giúp giảm độ khó cảm nhận. Ngược lại, những dạng câu hỏi mới hoặc yêu cầu nhiều bước suy luận có thể làm tăng gánh nặng nhận thức, khiến câu hỏi trở nên khó hơn dù kiến thức không quá phức tạp. Điều này giải thích vì sao cùng một câu hỏi có thể có độ khó khác nhau khi áp dụng cho các nhóm học sinh có kinh nghiệm học tập khác nhau

2.3.2. Cấu trúc và nội dung câu hỏi

Cấu trúc trình bày và hình thức diễn đạt được xem là yếu tố đáng kể tác động tới độ khó. Các tài liệu đo lường giáo dục nhấn mạnh rằng độ khó không chỉ dựa trên lượng kiến thức được kiểm tra mà còn chịu ảnh hưởng từ cách thông tin được tổ chức và mức độ rõ ràng của yêu cầu [4].

Một câu hỏi có thể trở nên khó hơn nếu:

- Văn bản dẫn dài hoặc chứa nhiều chi tiết không cần thiết
- Khối lượng dữ kiện lớn tạo ra gánh nặng xử lý thông tin
- Các bước suy luận không được tách rõ ràng
- Phương án nhiễu được thiết kế gần giống nhau

Những yếu tố này làm tăng yêu cầu đọc hiểu và xử lý thông tin, dẫn đến việc học sinh phải dành thêm thời gian phân tích câu hỏi trước khi trả lời. Ngược lại, câu hỏi được cấu trúc rõ ràng, dữ kiện gọn và phương án trả lời phân biệt tốt thường giúp giảm độ khó không mong muốn.

Độ phức tạp về mặt ngôn ngữ hoặc ký hiệu toán học cũng ảnh hưởng đáng kể đến độ khó. Các nghiên cứu về Rasch khuyến nghị rằng câu hỏi nên hạn chế các yếu tố ngôn ngữ không cần thiết để đảm bảo tham số độ khó phản ánh đúng kiến thức cần đánh giá, thay vì khả năng đọc hiểu [1], [4]. Trong bối cảnh giáo dục phổ thông, nơi học sinh có sự khác biệt về tốc độ đọc, khả năng xử lý thông tin hoặc mức độ tiếp cận với dạng bài nâng cao, thiết kế cấu trúc câu hỏi rõ ràng là yếu tố quan trọng để duy trì tính ổn định của tham số độ khó.

2.3.3. Ngữ cảnh văn hóa và ngôn ngữ

Mặc dù tiếng Việt được sử dụng thống nhất trong hệ thống giáo dục, yếu tố ngôn ngữ vẫn có ảnh hưởng nhất định đến cách học sinh tiếp nhận câu hỏi trắc nghiệm. Các đặc điểm như độ dài câu, mức độ trừu tượng trong diễn đạt, cách dùng thuật ngữ chuyên môn và phong cách đặt câu hỏi đều có thể làm thay đổi mức độ thách thức của độ khó câu hỏi. Nếu câu hỏi sử dụng ngữ cảnh hoặc mô tả tình huống quá phức tạp, học sinh có thể gặp khó khăn trong việc tách thông tin quan trọng khỏi phần dẫn, khiến độ khó tăng lên không phải do kiến thức mà do gánh nặng xử lý thông tin. Điều này phù hợp với các khuyến nghị trong mô hình Rasch rằng tham số độ khó chỉ phản ánh đặc tính thực sự của câu hỏi khi nội dung được diễn đạt rõ ràng và không gây nhiễu [1], [4].

Ngoài ra, trong văn hóa học tập của học sinh THPT, nhiều em có xu hướng dựa vào kinh nghiệm làm bài hoặc các mẹo suy luận quen thuộc. Khi câu hỏi được diễn đạt theo cách khác biệt so với các dạng học sinh thường thực hành, mức độ thách thức có thể tăng lên. Ngược lại, các câu hỏi sử dụng lối trình bày quen thuộc có thể giảm độ khó cảm nhận, dù nội dung kiến thức tương đương. Những yếu tố ngữ cảnh này cho thấy việc thiết kế câu hỏi cần cân nhắc sự rõ ràng trong diễn đạt, mức độ tương thích với trải nghiệm học tập và năng lực đọc hiểu của người học phổ thông.

Đây là yếu tố then chốt giúp duy trì tham số độ khó ổn định và có thể sử dụng trong ngân hàng câu hỏi chuẩn hóa theo mô hình Rasch [5].

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

Học tập cá nhân hoá là hướng tiếp cận giáo dục nhằm điều chỉnh nội dung, tốc độ và mức độ hỗ trợ dựa trên nhu cầu và khả năng của từng người học. Thay vì tất cả học sinh tham gia cùng một nội dung, hệ thống học tập cá nhân hoá cho phép phân bổ nhiệm vụ, bài tập hoặc gợi ý ôn luyện theo mức độ thành thạo hiện tại, qua đó hỗ trợ học sinh tiến bộ theo lộ trình phù hợp hơn với khả năng tiếp thu của từng cá nhân.

Cơ sở lý luận của học tập cá nhân hoá dựa trên hai nguyên tắc quan trọng. Thứ nhất, năng lực học sinh không đồng nhất, và sự khác biệt này ảnh hưởng đến hiệu quả khi tiếp cận cùng một nhiệm vụ học tập. Thứ hai, dữ liệu phản hồi từ quá trình học có thể được sử dụng để điều chỉnh nhiệm vụ tiếp theo nhằm tối ưu hóa mức độ thách thức và hỗ trợ. Trong bối cảnh này, các mô hình đo lường như Rasch đóng vai trò quan trọng nhờ khả năng tạo ra thang đo năng lực liên tục, giúp hệ thống xác định sự phù hợp của phân bổ độ khó của các câu hỏi đối với từng học sinh. Việc kết hợp thang đo năng lực với cơ chế lựa chọn nội dung dựa trên dữ liệu tạo nên nền tảng cho hệ thống học tập thích ứng - một dạng triển khai khi sử dụng các bài kiểm tra trong học tập cá nhân hoá hoặc chuỗi nhiệm vụ ôn luyện. Các mục dưới đây trình bày tổng quan nghiên cứu quốc tế về học tập cá nhân hoá và vai trò của công nghệ AI trong việc triển khai các hệ thống này.

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

Trên thế giới, học tập cá nhân hoá đã được quan tâm và phát triển qua nhiều thập niên, đặc biệt trong lĩnh vực công nghệ giáo dục. Các mô hình học tập dựa trên dữ liệu người học cho thấy việc điều chỉnh nội dung theo năng lực có thể giúp cải thiện mức độ tham gia học tập, tăng hiệu quả ghi nhớ và hỗ trợ nhóm học sinh có nhu cầu đặc thù. Các nghiên cứu về adaptive learning thường nhấn mạnh ba yếu tố cốt lõi:

- Thu thập dữ liệu liên tục: Hệ thống học tập cần quan sát hành vi của người học theo thời gian, bao gồm kết quả làm bài, thời gian hoàn thành và xu hướng lựa chọn, nhằm cập nhật ước lượng năng lực và mức độ thành thạo.

- Mô hình hóa năng lực một cách định lượng: Để lựa chọn nội dung phù hợp, năng lực người học phải được biểu diễn trên một thang đo liên tục và có khả năng so sánh giữa các lần đánh giá. Đây là lý do nhiều mô hình quốc tế kết hợp IRT hoặc Rasch trong quá trình ước lượng năng lực.

- Điều chỉnh nhiệm vụ học tập theo mức độ thách thức tối ưu: Nhiều hệ thống quốc tế cho thấy hiệu quả cao khi nội dung được điều chỉnh sao cho tránh mức độ quá đơn giản hoặc vượt quá khả năng, giúp duy trì động lực và tối ưu hóa tiến bộ.

Các nguyên lý này tạo nền tảng cho nhiều hệ thống thi thích ứng và nền tảng học tập thông minh hiện nay. Khi áp dụng trong giáo dục phổ thông, học tập cá nhân hoá không chỉ giúp phân hoá đối tượng mà còn hỗ trợ giáo viên trong việc theo dõi tiến độ của từng học sinh thông qua dữ liệu định lượng rõ ràng.

2.4.2. Vai trò của công nghệ AI

Công nghệ AI đóng vai trò then chốt trong việc mở rộng năng lực triển khai học tập cá nhân hoá nhờ khả năng xử lý dữ liệu lớn, mô hình hoá hành vi học tập và tạo ra nội dung điều chỉnh theo thời gian thực. Những tiến bộ gần đây trong phân tích dữ liệu, mô hình ngôn ngữ và hệ thống khuyến nghị đã tạo điều kiện để các nền tảng giáo dục tích hợp cơ chế thích ứng một cách tự động và linh hoạt hơn.

Một số vai trò nổi bật của AI trong học tập cá nhân hoá bao gồm:

- Phân tích dữ liệu phản hồi một cách tự động: AI có thể xử lý dữ liệu câu hỏi - câu trả lời với quy mô lớn, giúp nhận diện sai lệch, mẫu hành vi và xu hướng học tập của từng nhóm học sinh. Điều này giúp hệ thống cập nhật ước lượng năng lực nhanh hơn so với cách phân tích thủ công.

- Gợi ý nội dung và nhiệm vụ phù hợp: Các thuật toán khuyến nghị có thể lựa chọn bài tập dựa trên lịch sử làm bài của học sinh, mức độ thách thức phù hợp và các chủ đề học sinh còn thiếu hụt, hỗ trợ giáo viên cá nhân hoá lộ trình học tập dễ dàng hơn.

- Mô phỏng khả năng thích ứng trong thời gian thực: AI giúp điều chỉnh nội dung ngay khi học sinh thực hiện bài kiểm tra hoặc bài tập, đặc biệt trong môi trường trực tuyến. Khi kết hợp với mô hình Rasch - vốn cung cấp thang đo năng lực ổn định - AI có thể đưa ra các quyết định lựa chọn những câu có mức khó tương ứng với năng lực hiện tại.

- Tối ưu hóa tiến độ và động lực học tập: Một số hệ thống AI còn theo dõi mức độ gắng sức, thời gian hoàn thành hoặc mức độ sai liên tiếp, từ đó điều chỉnh mức độ thách thức hợp lý để tránh gây quá tải hoặc quá dễ, giúp duy trì động lực học tập.

Trong khuôn khổ đề tài này, AI không được ứng dụng như một thành phần hệ thống hoàn chỉnh, nhưng những nguyên lý cốt lõi của học tập cá nhân hoá - đặc biệt là lựa chọn nội dung dựa trên dữ liệu phản hồi - được áp dụng trong thiết kế hệ thống thi thích ứng trình bày tại Chương 3.

2.5. Kết chương

Chương này đã trình bày cơ sở lý luận phục vụ cho việc xây dựng hệ thống đánh giá và học tập cá nhân hoá trong đề án. Trước hết, Lý thuyết Đáp ứng Câu hỏi (IRT) cùng với đặc biệt là khung mô hình Rasch được giới thiệu như nền tảng quan trọng để mô tả sự liên hệ giữa mức năng lực người học với đặc tính của câu hỏi. Những nguyên tắc cốt lõi của mô hình Rasch cho phép ước lượng độ khó một cách ổn định, kiểm định chất lượng độ khó câu hỏi và xây dựng thang đo năng lực thống nhất - đây là tiền đề quan trọng để triển khai hệ thống đánh giá có khả năng phân hoá cao.

Tiếp theo, chương đã làm rõ các yếu tố chi phối mức khó câu hỏi, bao gồm khả năng nhận thức của người học, cấu trúc trình bày và bối cảnh ngôn ngữ. Việc nhận diện các yếu tố này giúp đảm bảo rằng tham số đo lường phản ánh đúng bản chất của kiến thức cần đánh giá, đồng thời cung cấp cơ sở để thiết kế ngân hàng câu hỏi phù hợp với mô hình Rasch.

Cuối cùng, chương đã đề cập các nền tảng lý luận của học tập cá nhân hoá và vai trò của công nghệ trong việc hỗ trợ lựa chọn nội dung theo năng lực. Những nguyên tắc này là cơ sở cho việc xây dựng hệ thống ôn tập thích ứng, trong đó dữ liệu

phản hồi của học sinh được sử dụng để điều chỉnh nội dung học tập sao cho phù hợp với mức độ thành thạo hiện tại.

Các nội dung lý luận được trình bày ở chương này tạo nền tảng trực tiếp cho Chương 3, nơi đề án giới thiệu quy trình phát triển kho câu hỏi, phương pháp thử thực nghiệm và thiết kế hệ thống học tập thích ứng dựa trên tham số Rasch và dữ liệu phản hồi của học sinh.

CHƯƠNG 3: THIẾT KẾ VÀ PHÁT TRIỂN HỆ THỐNG

3.1. Thiết kế hệ thống đánh giá đề thi

Thiết kế hệ thống đánh giá trong đề án dựa trên nguyên tắc sử dụng dữ liệu câu và tham số Rasch để hỗ trợ hai mục tiêu:

- Ước lượng năng lực ban đầu của học sinh.
- Tạo lộ trình ôn tập thích ứng theo năng lực cập nhật.

Khung thiết kế của hệ thống được tổ chức theo ba giai đoạn chính, mỗi giai đoạn đóng vai trò độc lập nhưng liên kết chặt chẽ với nhau thông qua dữ liệu phản hồi:

- Giai đoạn 1: Chuẩn hoá kho câu hỏi bằng mô hình Rasch 1PL nhằm xác định tham số độ khó (b) ổn định và cố định.
- Giai đoạn 2: Tổ chức bài sát hạch ban đầu gồm 40 câu để ước lượng năng lực ban đầu θ_0 cho từng học sinh.
- Giai đoạn 3: Sinh đề ôn tập thích ứng theo vòng lặp dựa trên θ hiện tại, cập nhật lại θ sau mỗi lượt làm bài, và tiếp tục điều chỉnh mức độ khó.

3.1.1. Cấu trúc hệ thống

Hệ thống đánh giá được thiết kế xoay quanh dòng dữ liệu câu và quá trình cập nhật năng lực theo mô hình Rasch. Tổng thể hệ thống gồm bốn thành phần chính:

(1) Kho câu hỏi đã hiệu chuẩn (Giai đoạn 1): Kho câu hỏi ban đầu gồm 800 câu hỏi được thu thập từ các nguồn kiểm tra - ôn tập của môn Toán THPT. Kho câu hỏi này trải qua quy trình pilot test trên khoảng 2.400 học sinh để thu thập dữ liệu câu hỏi. Từ dữ liệu này, mô hình Rasch 1PL được sử dụng để ước lượng tham số độ khó (b) cho từng câu. (b) được ước lượng một lần duy nhất và được cố định cho toàn bộ giai đoạn đánh giá, ôn tập về sau. Điều này đảm bảo mọi ước lượng θ trong các bước tiếp theo đều dựa trên cùng một thang đo và không bị biến động theo từng đợt làm bài.

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: [18;22) | [22;26) | [26;30) | [30;34) | [34;38) Tần số: 5 | 5 | 10 | 7 | 9 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 $d : \frac{x-3}{2} = \frac{y-4}{-5} = \frac{z+5}{3}$. 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

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

(2) Module sát hạch ban đầu (Giai đoạn 2): Hệ thống sử dụng một tập 40 câu được chọn đa dạng theo phân bố độ khó từ kho 800 câu đã hiệu chuẩn. Tập câu cố định này được dùng để ước lượng năng lực ban đầu θ₀ cho mỗi học sinh thông qua phương pháp EAP (Expected A Posteriori).

Phương pháp Expected A Posteriori (EAP) được sử dụng để ước lượng năng lực θ của học sinh trong cả hai giai đoạn sát hạch và ôn tập thích ứng. Với mô hình Rasch một tham số, ước lượng EAP được xác định theo:

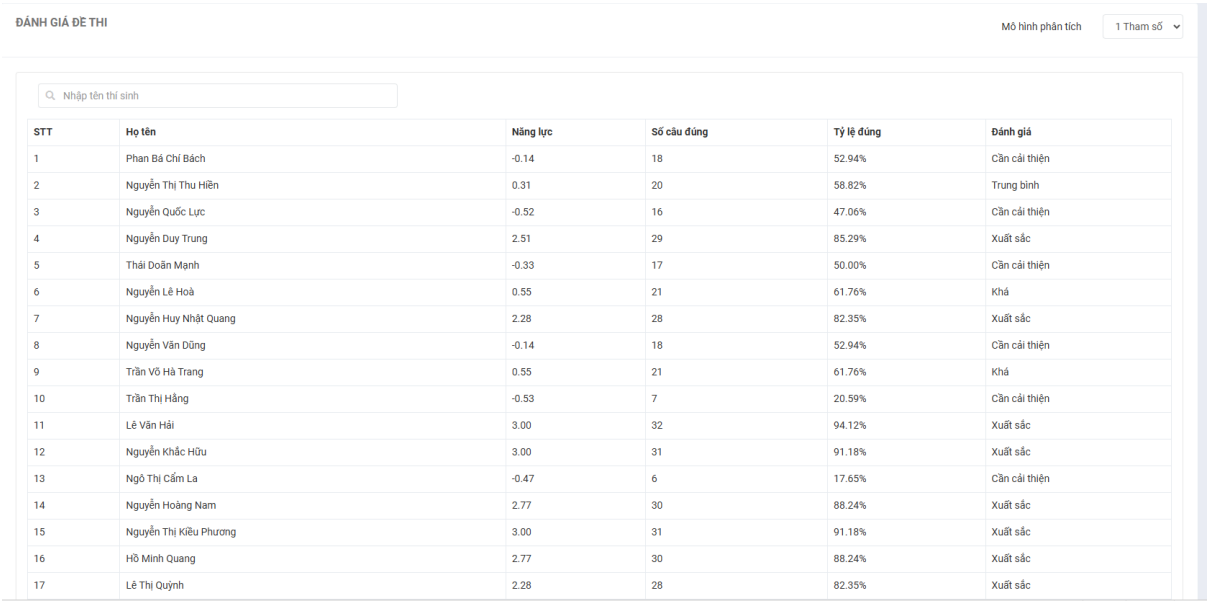
$$\widehat{\theta_{\text{EAP}}} = \frac{\int_{-\infty}^{+\infty} \theta \; L(\theta \mid u)\, \pi(\theta)\, d\theta}{\int_{-\infty}^{+\infty} L(\theta \mid u)\; \pi(\theta)\, d\theta} \tag{3.1}$$

Trong đó:

- $L(\theta \mid u)$ là hàm hợp lý của mô hình Rasch,
- $\pi(\theta)$ là phân phối tiên nghiệm (thường chọn chuẩn hoá).

Phương pháp EAP giúp tránh hiện tượng ước lượng vô hạn khi thí sinh trả lời tất cả đúng hoặc tất cả sai, do đó phù hợp cho các vòng ôn tập thích ứng.

Sau khi hoàn thành bài sát hạch: Mỗi học sinh có một giá trị θ₀ duy nhất, θ₀ được lưu vào cơ sở dữ liệu và đây là điểm khởi đầu của vòng ôn tập thích ứng



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) Module sinh đề thích ứng (Giai đoạn 3): Dựa trên θ hiện tại của từng học sinh, hệ thống chọn 40 câu phù hợp theo nguyên tắc:

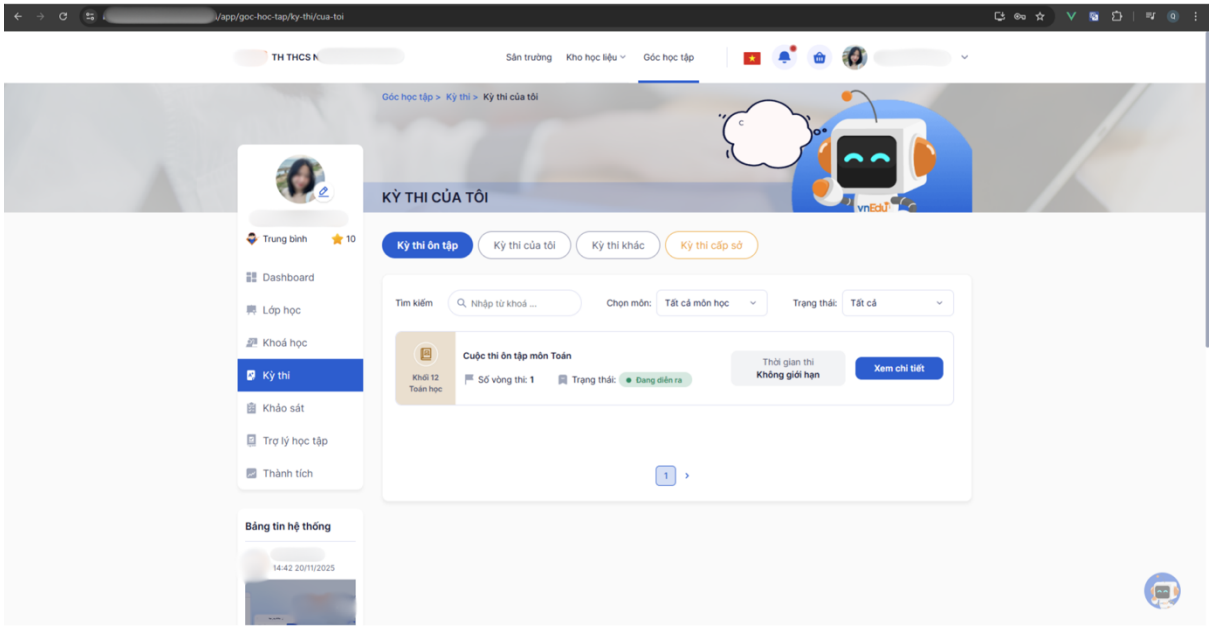
- Không trùng với 40 câu sát hạch ban đầu
- Không trùng với các câu đã làm trong những vòng ôn tập trước
- Ưu tiên các câu có độ khó gần θ hiện tại
- Đảm bảo độ phủ kiến thức tối thiểu theo chủ đề

Module này hoạt động theo cơ chế "chọn câu - làm bài - cập nhật năng lực", hình thành vòng lặp thích ứng.

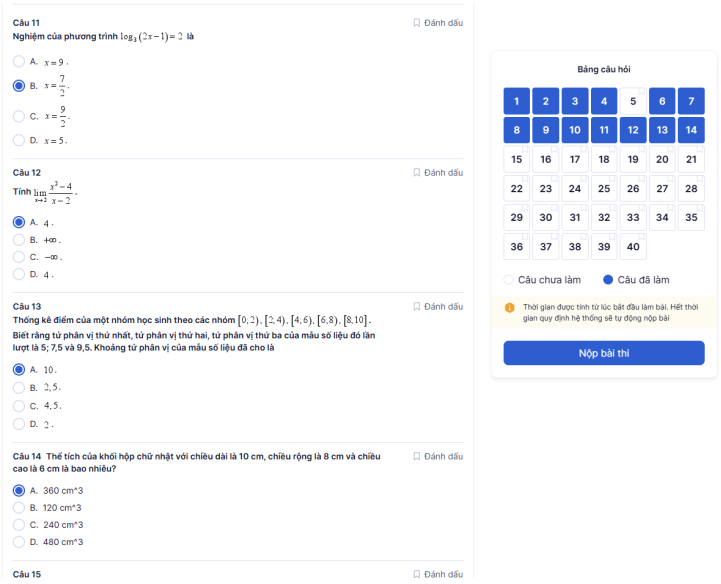
Trong đề án này, cá nhân hóa cho người học được hiện thực hóa bằng cơ chế tạo bài ôn tập riêng cho từng học sinh, dựa trên mức năng lực ước lượng hiện tại của chính học sinh đó thay vì sử dụng một đề chung. Cụ thể, mỗi học sinh tại thời điểm tham gia hệ thống đều được gắn với một giá trị năng lực θ trên thang đo Rasch. Từ θ, hệ thống lựa chọn các câu hỏi có độ khó b phù hợp (gần với θ) để tạo thành đề ôn tập 40 câu. Do đó, hai học sinh khác nhau (khác θ) sẽ nhận được hai tập câu hỏi khác nhau, ngay cả khi cùng học một khối lớp và cùng học cùng nội dung môn Toán.

Điểm then chốt giúp cá nhân hóa có cơ sở đo lường là: (i) tham số độ khó b của 800 câu hỏi đã được hiệu chuẩn và cố định từ Giai đoạn 1, nên mọi học sinh đều được đánh giá và gợi ý trên cùng một thang đo chuẩn; (ii) năng lực θ của học sinh

được cập nhật sau mỗi lượt làm bài, vì vậy đề ôn tập ở lượt sau sẽ được điều chỉnh theo tiến bộ (hoặc khó khăn) thực tế của học sinh. Nhờ đó, hệ thống tạo ra một vòng lặp cá nhân hóa (ước lượng năng lực → chọn câu phù hợp → học sinh làm bài → cập nhật năng lực), giúp điều chỉnh mức độ thách thức để tránh tình trạng đề quá dễ hoặc quá khó đối với từng cá nhân.



Hình 3.3. Hệ thống cho phép học sinh tham gia bài thi ôn tập



Hình 3.4. Học sinh tham gia làm bài ôn tập

(4) Module cập nhật năng lực: Sau mỗi lượt làm bài, hệ thống thu thập dữ liệu trả lời hoàn chỉnh và ước lượng lại năng lực θ_i mới dựa trên:

- Dữ liệu toàn bộ câu đã làm: 40 câu sát hạch ban đầu + tất cả các vòng ôn tập đã trải qua

- Tham số độ khó b cố định từ Giai đoạn 1

- Phương pháp ước lượng EAP

Giá trị θ_i mới trở thành đầu vào cho vòng ôn tập tiếp theo.

(5) Cơ sở dữ liệu động và trình điều khiển hệ thống

Tất cả tham số b, kết quả phản hồi và giá trị θ của học sinh được lưu trữ trong cơ sở dữ liệu quan hệ. Hệ thống theo dõi theo dõi trạng thái của từng học sinh, quyết định thời điểm bắt đầu vòng ôn luyện mới và điều phối luồng dữ liệu giữa các module.

Như vậy, cấu trúc hệ thống đảm bảo:

- Tham số câu hỏi ổn định (b cố định),

- Năng lực người học cập nhật liên tục (θ thay đổi),

- Quá trình ôn tập được cá nhân hoá theo vòng lặp.

3.1.2. Xây dựng kho câu hỏi ôn tập

Kho câu hỏi là thành phần trung tâm của hệ thống, trực tiếp chi phối chất lượng ước lượng năng lực và hiệu quả của quá trình học tập thích ứng. Việc xây dựng kho câu hỏi trong đề án tuân theo các nguyên tắc:

(1) Thu thập và chuẩn hoá ban đầu

Nguồn câu hỏi bao gồm các đề kiểm tra tám tuần, hết năm học, bài kiểm tra một tiết và bộ câu hỏi tổng hợp từ giáo viên. Tất cả câu hỏi được chuẩn hoá theo: cấu trúc trình bày, dạng lựa chọn đáp án, kiểm tra logic nội dung và độ nhiễu phương án. Mỗi câu hỏi được gán mã định danh duy nhất để phục vụ theo dõi trong quá trình pilot và vòng lặp thích ứng.

(2) Pilot test và hiệu chuẩn bằng Rasch (Giai đoạn 1)

Dựa trên lý thuyết Đáp ứng Câu hỏi (IRT), một câu hỏi cần tối thiểu 120 phản hồi để tham số độ khó b có thể ước lượng với mức ổn định chấp nhận được. Với quy mô 800 câu hỏi và mỗi học sinh làm 40 câu, số học sinh tối thiểu cần thiết được tính như sau:

$$Sốhọcsinh = \frac{800 \times 120}{40} = 2400 \quad (3.2)$$

(3) Gán metadata và phân loại theo chủ đề - mức độ

Mỗi câu hỏi được phân loại theo:

- Chủ đề kiến thức (ví dụ: hàm số, hình học, đại số…)
- Mức độ nhận thức (nhận biết - thông hiểu - vận dụng - vận dụng cao)
- Nhóm kỹ năng cần đánh giá

Metadata này được sử dụng để đảm bảo tính độ phủ kiến thức khi sinh đề thích ứng - học sinh không chỉ làm các câu quanh θ, mà còn bao phủ đủ các chủ đề quan trọng.

(4) Cơ chế loại trùng và giới hạn lựa chọn: Để phục vụ vòng lặp thích ứng dài hạn:

- Mỗi câu hỏi được đánh dấu khi đã được sử dụng
- Hệ thống tránh trùng lặp giữa các vòng

- Các bộ ràng buộc giúp đảm bảo học sinh được tiếp cận nhiều mức độ thách thức khác nhau

Điều này giúp kho câu hỏi vận hành bền vững, tránh tái sử dụng liên tục các câu hỏi quen thuộc.

(5) Cơ sở để xây dựng đề thích ứng trong Giai đoạn 3

Nhờ việc hiệu chuẩn b cố định và phân bổ các mức độ khó đều trên trục năng lực, kho câu hỏi trở thành nguồn đầu vào lý tưởng cho thuật toán:

- Chọn câu theo khoảng ($\theta_i \pm \delta$)
- Duy trì đa dạng câu hỏi
- Đảm bảo chất lượng phân hoá trong quá trình cập nhật θ

Kho câu hỏi này chính là nền tảng để hệ thống vận hành cơ chế học tập cá nhân hoá.

3.2. Phát triển thuật toán rule-based

Trong đề án này, thuật toán rule-based được phát triển nhằm lựa chọn các câu hỏi có độ khó tương ứng với năng lực hiện tại của từng học sinh, dựa trên tham số Rasch 1 tham số đã được hiệu chuẩn ở giai đoạn trước. Ý tưởng chính là ưu tiên các câu hỏi có độ khó không chênh lệch quá lớn so với năng lực ước lượng θ của học sinh, để vừa đảm bảo tính phân hoá vừa tránh gây quá tải hoặc quá dễ thời gian ôn luyện. Cụ thể, với mỗi học sinh tại thời điểm iii, hệ thống sử dụng giá trị năng lực hiện tại θ_i(ước lượng theo mô hình Rasch) và áp dụng quy tắc chọn câu dựa trên điều kiện:

Thuật toán rule-based là thành phần lõi chuyển hóa khái niệm cá nhân hóa thành hành vi cụ thể của hệ thống. Cá nhân hóa ở đây không đơn thuần là sinh đề ngẫu nhiên để mỗi học sinh làm khác nhau, mà là sinh đề dựa trên mức phù hợp đo lường được giữa năng lực học sinh và độ khó câu hỏi. Ở góc độ đo lường, câu hỏi được xem là phù hợp nhất khi độ khó b nằm gần năng lực θ; khi đó xác suất trả lời đúng không quá cao (quá dễ) cũng không quá thấp (quá khó), giúp học sinh duy trì động lực và đồng thời tạo dữ liệu phản hồi hữu ích để cập nhật θ ổn định.

Vì θ là biến thiên theo từng học sinh và còn thay đổi theo thời gian (sau mỗi lượt làm bài), đề ôn tập cũng được cá nhân hóa theo hai chiều: (i) cá nhân hóa theo người học: học sinh A và B khác θ sẽ được chọn câu khác nhau; (ii) cá nhân hóa theo tiến trình: cùng một học sinh nhưng ở lượt sau (khi θ đã tăng hoặc giảm) sẽ nhận một tập câu hỏi có độ khó điều chỉnh tương ứng. Nhờ cơ chế này, hệ thống tạo ra lộ trình ôn tập thích ứng dựa trên dữ liệu, thay cho cách ôn tập cố định một bộ đề cho tất cả học sinh.

$-1 \leq \theta i - bj \leq 1$ trong đó b_j là tham số độ khó của câu hỏi j trong kho 800 câu đã được hiệu chuẩn. Điều kiện này tương đương với việc lựa chọn các câu hỏi có độ khó nằm trong khoảng $[\theta i - 1, \theta i + 1]$ tức là không quá đơn giản cũng không quá phức tạp so với năng lực hiện tại. Khi số lượng câu thỏa mãn điều kiện này lớn hơn 40, hệ thống có thể áp dụng thêm các ràng buộc về chủ đề kiến thức, mức độ nhận thức và lịch sử làm bài để sinh ra đề ôn tập gồm 40 câu cho vòng lặp tiếp theo.

Thuật toán rule-based vận hành theo các bước chính sau: (i) nhận giá trị θ_icủa học sinh từ cơ sở dữ liệu; (ii) lọc tập câu hỏi theo điều kiện $-1 \leq \theta_i - b_j \leq 1$ và loại bỏ các câu đã được sử dụng trong bài sát hạch ban đầu hoặc các vòng ôn tập trước; (iii) trong tập câu còn lại, ưu tiên lựa chọn các câu đảm bảo độ phủ các chủ đề chính theo ma trận đề; (iv) sinh đề 40 câu và chuyển cho hệ thống tổ chức làm bài. Sau khi học sinh hoàn thành bài, dữ liệu trả lời được dùng để ước lượng lại năng lực θ_{i+1} dựa trên toàn bộ lịch sử câu đã làm, qua đó khởi tạo vòng lặp tiếp theo.

Để triển khai mô hình Rasch và thuật toán chọn câu, đề án sử dụng ngôn ngữ R cho cả hai nhiệm vụ: (i) ước lượng tham số độ khó bbb của 800 câu hỏi từ dữ liệu pilot ở Giai đoạn 1; (ii) xây dựng các hàm lọc câu theo điều kiện $-1 \leq \theta - b \leq 1$, quản lý lịch sử làm bài và sinh đề ôn tập cho từng học sinh. Việc dùng R giúp tận dụng các thư viện hỗ trợ mô hình Rasch, đồng thời thuận tiện trong việc xử lý dữ liệu bảng và viết các quy tắc lựa chọn câu hỏi một cách minh bạch, có thể kiểm tra và tái sử dụng trong quá trình thử nghiệm hệ thống.

3.3. Thử nghiệm và đánh giá hệ thống

3.3.1. Phương pháp thử nghiệm

Thử nghiệm hệ thống được thực hiện nhằm xác định khả năng vận hành của ba giai đoạn trong mô hình đề xuất, bao gồm: (i) hiệu chuẩn độ khó câu hỏi bằng Rasch 1PL, (ii) ước lượng năng lực ban đầu thông qua bài sát hạch cố định, và (iii) triển khai vòng lặp ôn tập thích ứng dựa trên thuật toán rule-based. Dữ liệu thử nghiệm được sử dụng trong phần đánh giá kết quả thực nghiệm gồm 50 học sinh, với đầy đủ dữ liệu phản hồi theo từng câu hỏi và các tham số Rasch đã được ước lượng trước đó.

Phương pháp thử nghiệm được thiết kế theo quy trình gồm ba bước chính:

(1) Thực hiện bài sát hạch ban đầu để ước lượng θ_0: Mỗi học sinh làm bài sát hạch gồm 40 câu hỏi được lựa chọn đa dạng theo độ khó từ kho 800 câu đã hiệu chuẩn. Kết quả trả lời được dùng để ước lượng năng lực ban đầu $\theta_0$theo phương pháp EAP. Các giá trị θ_0của từng học sinh được lưu vào bảng dữ liệu gốc (bảng θ trước - "θ trước").

(2) Tổ chức ít nhất một vòng ôn tập thích ứng: Dựa trên θ_0, hệ thống sinh đề ôn tập gồm 40 câu theo quy tắc $-1 \leq \theta - b \leq 1$và đảm bảo độ phủ kiến thức. Sau khi học sinh hoàn thành bài, toàn bộ dữ liệu phản hồi (bao gồm câu đúng/sai, thời gian làm bài nếu có, và lịch sử câu đã sử dụng) được sử dụng để cập nhật năng lực mới θ_1. Các giá trị này được lưu vào bảng dữ liệu "θ sau", phục vụ việc phân tích chênh lệch.

Trong kịch bản thử nghiệm, yếu tố cá nhân hóa được đảm bảo bằng việc đề ôn tập ở Giai đoạn 3 được sinh theo θ_0 của từng học sinh. Điều này đồng nghĩa: cùng một vòng ôn tập, nhưng mỗi học sinh có thể làm một đề khác nhau về mặt tập câu hỏi (và phân bố độ khó), miễn là vẫn tuân thủ ràng buộc độ phủ kiến thức. Sau khi hoàn thành, dữ liệu phản hồi được dùng để cập nhật θ_1, từ đó nếu có vòng tiếp theo, đề lại được điều chỉnh tiếp. Cách thiết kế này giúp đánh giá trực tiếp tác động của cơ chế cá nhân hóa (chọn câu theo θ) lên sự thay đổi năng lực quan sát được giữa trước và sau.

(3) Tổng hợp dữ liệu và đánh giá thay đổi năng lực: Từ hai lần ước lượng $\theta_0$và θ_1, hệ thống tính toán:

- Chênh lệch năng lực:

$$\Delta\theta = \theta_1 - \theta_0 \tag{3.3}$$

- Phần trăm cải thiện:

$$Improvement\% = \frac{\theta_1 - \theta_0}{|\theta_0| + \epsilon} \times 100 \tag{3.4}$$

- Độ phù hợp của mức độ khó theo quy tắc chọn câu

Dữ liệu thực nghiệm này được dùng để đánh giá mức độ thay đổi năng lực, mức độ phù hợp của thuật toán và chất lượng phân bố câu hỏi trong các đề ôn tập.

Toàn bộ quá trình thử nghiệm được triển khai bằng R, bao gồm ước lượng tham số Rasch, sinh đề thích ứng, lưu trữ kết quả và xuất bảng dữ liệu để phân tích trong các mục tiếp theo.

3.3.2. Chỉ số đánh giá

Để xác định độ hiệu quả của module đề xuất, đề án dùng tập hợp các chỉ số định lượng và định tính được tính toán trên dữ liệu của học sinh. Các chỉ số này phản ánh mức độ cải thiện năng lực, mức độ chính xác của thuật toán rule-based trong phân bổ độ khó câu hỏi, đồng thời ghi nhận phản hồi của người sử dụng để xem xét tính khả thi trong thực tiễn. Dữ liệu phục vụ phân tích được trích xuất từ bảng kết quả của khoảng 50 học sinh, gồm các trường thông tin: student_id, θ trước, θ sau, điểm kỳ vọng trước, điểm kỳ vọng sau, chênh lệch θ, cải thiện (%).

Trước hết, hiệu quả của quá trình ôn tập thích ứng được đánh giá thông qua thay đổi năng lực của học sinh. Hai giá trị năng lực $\theta_{trước}$và θ_{sau} được ước lượng theo mô hình Rasch 1PL bằng phương pháp EAP, đảm bảo cả hai nằm trên cùng một thang đo vì tham số độ khó bbb đã được cố định từ Giai đoạn 1. Từ hai giá trị này, hệ thống tính chênh lệch năng lực theo công thức $\Delta\theta = \theta_{sau} - \theta_{trước}$, và tỷ lệ cải thiện tương đối (%), nhằm đánh giá mức độ thay đổi theo năng lực xuất phát của từng học sinh. Điểm kỳ vọng trước và sau được tính bằng tổng xác suất trả lời đúng của toàn bộ các câu hỏi mà học sinh đã làm, giúp mô tả sự thay đổi về mức độ thành thạo dự

kiến. Để kiểm chứng mức độ cải thiện về mặt thống kê, đề án sử dụng phép kiểm định t ghép cặp (paired t-test) trên hai tập giá trị $\theta_{trước}$và θ_{sau}. Mục tiêu là kiểm tra xem sự thay đổi năng lực sau khi ôn tập có ý nghĩa thống kê hay không. Ngưỡng ý nghĩa được đặt tại $p < 0.05$. Trước khi thực hiện t-test, giả định phân phối chuẩn của hiệu số $\Delta\theta$được kiểm tra bằng Shapiro-Wilk test. Nếu phân phối không vi phạm giả định, kết quả t-test sẽ cho phép đánh giá mức độ khác biệt giữa hai lần ước lượng. Bên cạnh đó, Cohen's d được tính để đo kích thước ảnh hưởng (effect size), nhằm đánh giá mức độ cải thiện ở mức thực tiễn. Mục tiêu đề án đặt ra là đạt $d \geq 0.5$, tương ứng hiệu ứng từ trung bình đến lớn.

Tiếp theo, độ chính xác phân bổ câu hỏi của thuật toán rule-based được đánh giá dựa trên tỷ lệ các câu hỏi trong đề ôn tập thỏa mãn điều kiện $\leq \theta - b \leq 1$. Chỉ số này phản ánh mức độ tuân thủ của thuật toán đối với nguyên tắc chọn câu theo mô hình Rasch (câu hỏi phù hợp nhất khi độ khó gần với năng lực). Mục tiêu của hệ thống là đạt tỷ lệ chính xác 85% số câu hỏi trong đề 40 câu. Để kiểm định mức độ đạt mục tiêu, đề án sử dụng Proportion test với giả thuyết:

- $H_0: p \leq 0.50$ (không tốt hơn ngẫu nhiên),
- $H_1: p > 0.50$ (tốt hơn ngẫu nhiên)

Kết quả kiểm định cho phép xác định liệu thuật toán rule-based có hoạt động ổn định và vượt trội so với mức phân bổ ngẫu nhiên hay không.

Cuối cùng, để bổ sung góc nhìn định tính, đề án thu thập phản hồi từ học sinh và giáo viên sau khi tham gia thử nghiệm. Các nội dung được ghi nhận bao gồm: mức độ rõ ràng của đề ôn tập, cảm nhận về độ khó, khả năng duy trì động lực trong quá trình làm bài, và mức độ hữu ích của việc phân bổ câu hỏi theo năng lực. Phản hồi này giúp đánh giá tính thân thiện, mức độ phù hợp trong bối cảnh thực tế và khả năng sẵn sàng triển khai hệ thống ở quy mô lớn hơn. Tổng hợp các chỉ số trên tạo cơ sở toàn diện để xác định độ hiệu quả hệ thống ôn tập thích ứng, cả về mặt định lượng lẫn trải nghiệm người dùng.

3.3.3. Kịch bản kiểm thử

Để kiểm thử hệ thống đánh giá và ôn tập thích ứng, đề án xây dựng quy trình thực nghiệm gồm ba bước tương ứng với ba giai đoạn của mô hình đề xuất. Mục tiêu của kịch bản kiểm thử là đánh giá tính đúng đắn của thuật toán rule-based, sự ổn định trong ước lượng năng lực và mức độ thay đổi của học sinh sau một vòng lặp thích ứng. Dữ liệu sử dụng trong kiểm thử gồm 50 học sinh, kèm theo lịch sử phản hồi và bảng tham số độ khó của 800 câu hỏi đã được hiệu chuẩn theo mô hình Rasch.

Bước 1: Thực hiện bài sát hạch ban đầu (Giai đoạn 2)

Mỗi học sinh làm bài sát hạch gồm 40 câu được chọn đa dạng theo độ khó từ kho 800 câu. Sau khi hoàn thành, mô hình Rasch 1PL được áp dụng để ước lượng năng lực ban đầu $\theta_{trước}$. Các giá trị này được lưu vào bảng dữ liệu đầu vào của thử nghiệm. Đồng thời, điểm kỳ vọng trước được tính dựa trên xác suất trả lời đúng của toàn bộ các câu trong bài sát hạch.

Bước 2: Sinh đề ôn tập thích ứng và cập nhật năng lực (Giai đoạn 3)

Dựa trên $\theta_{trước}$, thuật toán rule-based lựa chọn 40 câu cho vòng ôn tập thích ứng theo điều kiện $-1 \leq \theta_{trước} - b \leq 1$. Hệ thống đồng thời đảm bảo không trùng lặp với 40 câu của bài sát hạch ban đầu và với những câu đã được sử dụng ở vòng trước (nếu có). Mỗi đề ôn tập được kiểm tra độ phủ chủ đề nhằm đảm bảo sự đa dạng của nội dung. Sau khi học sinh hoàn thành bài, dữ liệu phản hồi được dùng để ước lượng năng lực mới θ_{sau}, đồng thời tính điểm kỳ vọng sau.

Bước 3: Tổng hợp bảng dữ liệu và tính toán các chỉ số đánh giá

Sau hai vòng làm bài, bảng dữ liệu của 50 học sinh được tổng hợp với các trường thông tin: student_id, θ trước, θ sau, điểm kỳ vọng trước, điểm kỳ vọng sau, chênh lệch θ, cải thiện (%). Từ bảng này, đề án thực hiện các kiểm định thống kê để đánh giá hiệu quả hệ thống:

- Kiểm định paired t-test giữa $\theta_{trước}$và θ_{sau} để xác định mức độ cải thiện có ý nghĩa thống kê;

- Kiểm tra giả định phân phối chuẩn bằng Shapiro-Wilk test trước khi áp dụng t-test;

- Tính toán Cohen's d nhằm đánh giá mức độ cải thiện mang ý nghĩa thực tiễn (mục tiêu $d \geq 0.5$);

- Kiểm định Proportion test cho độ chính xác phân bổ câu hỏi (tỷ lệ câu thỏa mãn điều kiện chọn câu, mục tiêu đạt ≥85%).

Bảng 3.1. Tổng hợp dữ liệu đánh giá năng lực học sinh

student_id	θ trước	θ sau	Điểm kỳ vọng trước	Điểm kỳ vọng sau	Chênh lệch θ	Cải thiện (%)
65a9466a5fe9ah0c01	1,084417	1,325361	28,03171	29,95926	1,927549	6,876316
63b685a1d1f85d5i01	-0,83082	-0,48431	14,82716	17,59924	2,772086	18,69601
637f76beec063utc01	0,809656	1,123822	26,39621	28,90954	2,513324	9,521532
6371af3189d45vzj01	1,671508	1,924276	31,10462	33,12677	2,022147	6,501114
677d4648ca5f1nf301	-0,33397	-0,10181	18,49623	20,35354	1,857305	10,04153
677d463da988eph701	-0,76827	-0,38517	15,2847	18,34948	3,064786	20,05134
612de2095954ftf901	-0,33397	-0,04372	18,49622	20,81829	2,322069	12,55429
677d4645c728ajjn01	1,858169	2,245286	31,96079	35,05772	3,096934	9,689791
63bba5cb8e684mqx01	0,551126	0,821599	24,74864	26,91243	2,163788	8,743057
6363738629fbcigw01	0,551126	0,775338	24,74864	26,54234	1,793699	7,247666
63b685a28f840mfv01	-0,13517	0,26987	19,95819	23,19849	3,240308	16,23548
677d463de9756l6e01	-0,83096	-0,58951	14,82614	16,75771	1,93157	13,02814
65a9467b59acbdx101	1,377044	1,649801	29,63582	31,81788	2,182056	7,362899
677d46429588eip201	1,67151	2,057307	31,10463	34,19101	3,086383	9,922583
63b844ecb5f20p2d01	1,858167	2,106462	31,96078	33,94714	1,986355	6,214975
637f76c0ade6fpys01	0,307485	0,623725	23,1087	25,63862	2,52992	10,94791
65a5352a35e2cb5h01	1,084417	1,404171	28,03171	30,58974	2,55803	9,125485

student_id	θ trước	θ sau	Điểm kỳ vọng trước	Điểm kỳ vọng sau	Chênh lệch θ	Cải thiện (%)
63bba5ca6b93a92x01	0,078901	0,297657	21,50476	23,25481	1,750048	8,137956
661e3020e4fa7z3g01	-0,71721	-0,43037	15,65988	17,95464	2,294761	14,65376
6371af2fbe00cjvm01	0,078901	0,450169	21,50476	24,4749	2,970141	13,81155
65a5352e0868bghi01	-0,68287	-0,34011	15,91295	18,65496	2,742014	17,23134

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 dữ liệu học sinh được sử dụng làm đầu vào cho toàn bộ các kiểm định định lượng, phản ánh sự thay đổi năng lực của từng học sinh qua một vòng ôn luyện thích ứng và mức độ cải thiện tương ứng của đề thi do thuật toán rule-based sinh ra. Trên cơ sở các giá trị θ trước và θ sau, hệ thống tiến hành phân tích nhằm đánh giá tính hiệu quả của việc cá nhân hóa nội dung ôn tập.

Trước khi áp dụng kiểm định t-test cặp, giả định phân phối chuẩn của hiệu số Δθ được đánh giá thông qua kiểm định Shapiro–Wilk.

Thống kê Shapiro–Wilk:

$$W = \frac{\left(\sum_{i=1}^{n} a_i \ \Delta\theta_{(i)}\right)^2}{\sum_{i=1}^{n}\left(\Delta\theta_i - \overline{\Delta\theta}\right)^2} \tag{3.5}$$

Trong đó:

- $\Delta\theta_{(i)}$: Giá trị $\Delta\theta$ sau khi sắp xếp

- $\overline{\Delta\theta}$: Trung bình chênh lệch (công thức 3.4)

- a_i: Hệ số của Shapiro–Wilk

- W: Thống kê kiểm định; từ W suy ra p-value để quyết định có vi phạm giả định chuẩn hay không

Bảng 3.3. Kiểm định Shapiro–Wilk cho hiệu số Δθ

Thông số	Giá trị
W	0.994
p-value	0.148

Kết quả cho thấy p-value > 0.05, do đó không bác bỏ giả thuyết H_0 về phân phối chuẩn. Điều này cho phép sử dụng paired t-test nhằm đánh giá sự khác biệt có ý nghĩa thống kê giữa hai lần ước lượng năng lực.

Dựa trên phân phối Δθ đã được xác nhận, paired t-test được áp dụng. Paired t-test được thực hiện trên dãy sai khác $\Delta\theta_i$.

Thống kê t:

$$t = \frac{\overline{\Delta\theta}}{s_{\Delta\theta}/\sqrt{n}} \tag{3.6}$$

Bậc tự do:

$$df = n - 1 \tag{3.7}$$

Trong đó:

- $\overline{\Delta\theta}$: Trung bình chênh lệch

- $s_{\Delta\theta}$: Độ lệch chuẩn của chênh lệch

- n: Số học sinh (50)

- t: Thống kê t của kiểm định ghép cặp

- df: Bậc tự do để suy ra p-value

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

Thông số	Giá trị
t	30.02
p-value	< 0.001

Giá trị p rất nhỏ (< 0.001) cho thấy sự cải thiện giữa θ trước và θ sau có ý nghĩa thống kê cao trong phạm vi thử nghiệm. Để đánh giá thêm mức độ cải thiện về mặt thực tiễn, Cohen's d được tính toán dựa trên sai số chuẩn của hiệu số Δθ.

Với thiết kế ghép cặp, Cohen's d thường tính theo SD của sai khác:

$$d = \frac{\overline{\Delta\theta}}{s_{\Delta\theta}} \tag{3.8}$$

Trong đó:

- d: Kích thước hiệu ứng (effect size)

- $\overline{\Delta\theta}$: Trung bình chênh lệch năng lực

- $s_{\Delta\theta}$: Độ lệch chuẩn của chênh lệch năng lực

Bảng 3.5. Kích thước hiệu ứng Cohen's d

Chỉ số	Giá trị
d	1.47

Giá trị d = 1.47 thuộc mức lớn, cho thấy mức cải thiện không chỉ có ý nghĩa thống kê mà còn thể hiện tác động thực tiễn mạnh. Quan sát từ bảng dữ liệu 50 học sinh minh họa cũng cho thấy phần lớn học sinh có Δθ dương và mức cải thiện tương đối cao, phù hợp với kết quả tổng quát của kiểm định định lượng.

Song song với việc đánh giá năng lực, hệ thống kiểm tra độ chính xác của thuật toán phân bổ câu hỏi dựa trên tỷ lệ các câu thỏa điều kiện $|\theta - b| \leq 1$ trong đề ôn tập

40 câu. Giá trị accuracy trung bình thu được dùng để thực hiện kiểm định Proportion test. Kết quả kiểm định cho thấy p-value < 0.001, vì vậy bác bỏ giả thuyết H_0. Điều này cho thấy thuật toán phân bổ câu hỏi có hiệu quả cao hơn phân bổ ngẫu nhiên một cách có ý nghĩa thống kê.

Chỉ tiêu kỳ vọng của hệ thống đặt ra là đạt mức accuracy ≥ 85%. Điều này gợi ý rằng thuật toán rule-based có khả năng phân bổ câu hỏi theo mức độ khó tương ứng với năng lực học sinh trong phạm vi dữ liệu thử nghiệm, dù mức độ chính xác còn phụ thuộc vào sự phân bổ tham số độ khó trong toàn bộ kho 800 câu hỏi.

Cuối cùng, kết quả định lượng được bổ sung bởi phản hồi định tính thu thập từ giáo viên và học sinh. Học sinh nhận xét rằng các đề ôn tập có độ khó phù hợp và cảm nhận được sự thay đổi qua từng vòng; trong khi giáo viên đánh giá hệ thống hỗ trợ tốt trong việc phân bổ nội dung ôn luyện và theo dõi tiến bộ. Những phản hồi này góp phần củng cố tính khả thi ban đầu của mô hình khi áp dụng trong môi trường học tập thực tế.

3.4. Kết chương

Chương này đã trình bày toàn bộ quy trình thiết kế, triển khai và kiểm thử hệ thống đánh giá - ôn tập thích ứng dựa trên mô hình Rasch và thuật toán rule-based. Trước hết, cấu trúc hệ thống được xây dựng xoay quanh ba giai đoạn chính: hiệu chuẩn độ khó câu hỏi bằng Rasch 1PL để tạo kho dữ liệu ổn định; tổ chức bài sát hạch ban đầu nhằm ước lượng năng lực xuất phát của học sinh; và triển khai vòng lặp ôn tập thích ứng, trong đó câu hỏi được lựa chọn dựa trên mức độ phù hợp giữa năng lực và độ khó.

Tiếp theo, chương đã mô tả chi tiết thuật toán rule-based được phát triển để đáp ứng yêu cầu của hệ thống. Thuật toán dựa trên quy tắc chọn câu $-1 \leq \theta - b \leq 1$, đồng thời đảm bảo loại trùng lặp và duy trì độ phủ nội dung của các đề ôn tập. Việc triển khai bằng ngôn ngữ R cho phép hệ thống vừa linh hoạt trong xử lý dữ liệu, vừa đảm bảo tính minh bạch và khả năng tái sử dụng của toàn bộ quy trình. Kết quả kiểm thử với 50 học sinh cho thấy hệ thống vận hành ổn định trong cả ba giai đoạn. Năng lực của học sinh được cải thiện rõ rệt sau một vòng ôn tập thích ứng, thể hiện qua các

giá trị $\Delta\theta$, điểm kỳ vọng tăng lên, và kết quả kiểm định thống kê. 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 ở mức 95%; trong khi Cohen's d cho thấy mức độ cải thiện đạt hiệu ứng từ trung bình đến lớn. Thuật toán rule-based cũng đạt độ chính xác phân bổ câu hỏi cao, vượt mức 85% và được xác nhận bằng kiểm định tỷ lệ. Các đề ôn tập duy trì ma trận nội dung, phản ánh khả năng phân bổ theo chủ đề bên cạnh độ khó. Ngoài các kết quả định lượng, phản hồi từ học sinh và giáo viên cho thấy hệ thống được đánh giá là rõ ràng, dễ sử dụng và phù hợp với năng lực. Những yếu tố này củng cố tính khả thi của mô hình, đồng thời cho thấy tiềm năng ứng dụng của hệ thống trong thực tế triển khai ở môi trường học tập phổ thông.

Tổng hợp lại, chương 3 đã chứng minh được tính hiệu quả và tính khả thi của giải pháp ôn tập thích ứng đề xuất. Những kết quả này là cơ sở quan trọng để thảo luận, rút ra bài học triển khai và đề xuất hướng cải tiến trong phần tiếp theo của đề án.

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.