

XÂY DỰNG VÀ THỰC NGHIỆM PHẦN MỀM HỖ TRỢ ĐÁNH GIÁ CHẤT LƯỢNG ĐỀ THI THEO HƯỚNG ĐỊNH LƯỢNG TẠI TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP QUẢNG NINH

Nguyễn Thu Hiền^{1,*}, Nguyễn Thị Phương¹,
Trương Thị Khánh Ly¹, Phạm Thị Hải¹,
¹Trường Đại học Công nghiệp Quảng Ninh
Email: hientn123@gmail.com

TÓM TẮT

Trong bối cảnh các cơ sở giáo dục đại học đang tăng cường đổi mới công tác kiểm tra, đánh giá theo chuẩn đầu ra, việc xây dựng công cụ hỗ trợ đánh giá chất lượng đề thi dựa trên các tiêu chí khoa học và khách quan trở nên cần thiết. Bài báo này trình bày quá trình xây dựng và thực nghiệm phần mềm hỗ trợ đánh giá chất lượng đề thi học phần tại Trường Đại học Công nghiệp Quảng Ninh dựa trên lý thuyết kiểm tra cổ điển (Classical Test Theory – CTT). Phần mềm sử dụng các chỉ số độ khó, độ phân biệt và độ tin cậy Cronbach's Alpha; trong đó, bách phân vị được sử dụng để xác định nhóm người học có kết quả cao và nhóm người học có kết quả thấp khi tính độ phân biệt.

Phần mềm được xây dựng bằng ngôn ngữ Python, sử dụng thư viện Pandas và NumPy để xử lý dữ liệu, Streamlit để xây dựng giao diện người dùng. Phần mềm hỗ trợ kiểm tra và chuẩn hoá dữ liệu đầu vào, tính toán các chỉ số, phân loại câu hỏi, cảnh báo các trường hợp cần rà soát và xuất báo cáo kết quả. Qua đó, phần mềm góp phần nâng cao hiệu quả công tác khảo thí và cung cấp căn cứ phục vụ đánh giá mức độ đạt chuẩn đầu ra học phần tại Trường Đại học Công nghiệp Quảng Ninh. Phần mềm được thử nghiệm trên 41 đề thi thuộc 29 học phần của các chương trình đào tạo khác nhau trong học kỳ I năm học 2025-2026, gồm 04 đề trắc nghiệm, 19 đề tự luận và 18 đề hỗn hợp.

Kết quả thực nghiệm cho thấy phần mềm có khả năng xử lý cả 03 hình thức đề thi, tự động tính toán các chỉ số và cung cấp thông tin phục vụ việc nhận diện câu hỏi quá dễ, quá khó, có độ phân biệt thấp hoặc có dấu hiệu bất thường. Tuy nhiên, với lớp có quy mô nhỏ hơn 10 sinh viên, kết quả độ phân biệt chỉ nên được xem là chỉ báo tham khảo. Việc kết luận câu hỏi giữ lại, chỉnh sửa hay loại bỏ cần kết hợp các chỉ số về độ khó, độ phân biệt, độ tin cậy và nhận định của giảng viên và bộ môn.

Từ khóa: Lý thuyết kiểm tra cổ điển (CTT), độ phân biệt; độ khó; độ tin cậy Cronbach's Alpha; phân tích đề thi; Python; Streamlit.

1. ĐẶT VẤN ĐỀ

Trong thực tiễn giáo dục đại học, việc đánh giá chất lượng đề thi vẫn còn một số hạn chế, đặc biệt trong bối cảnh yêu cầu chuẩn hoá và kiểm định chất lượng ngày càng được chú trọng. Một trong những nguyên nhân chính là việc thiếu các công cụ hỗ trợ phân tích dữ liệu khảo thí một cách hệ thống, khiến quá trình đánh giá đề thi chủ yếu dựa trên kinh nghiệm cá nhân và các phương pháp xử lý thủ công. Điều này không chỉ làm giảm độ chính xác của kết quả đánh giá mà còn hạn chế khả năng khai

thác dữ liệu để cải tiến đề thi và nâng cao chất lượng đào tạo.

Tại Trường Đại học Công nghiệp Quảng Ninh, mặc dù công tác khảo thí được triển khai ổn định, song việc phân tích chất lượng đề thi chưa được thực hiện thường xuyên và đồng bộ. Dữ liệu điểm thi chủ yếu được lưu trữ ở dạng tổng hợp, thiếu thông tin chi tiết cho từng câu hỏi, gây khó khăn cho việc áp dụng các phương pháp phân tích định lượng. Thêm vào đó, việc sử dụng các công cụ như Excel, cùng với công cụ hỗ trợ AI theo xu hướng hiện nay để xử lý dữ

liệu vẫn mang tính rời rạc, chưa đáp ứng yêu cầu tự động hoá. Xuất phát từ thực trạng trên, nghiên cứu đã tiến hành xây dựng và thực nghiệm phần mềm hỗ trợ đánh giá chất lượng đề thi dựa trên các chỉ số đo lường khoa học, qua đó cung cấp căn cứ định lượng phục vụ rà soát, điều chỉnh và hoàn thiện đề thi, ngân hàng câu hỏi.

2. DỮ LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Cơ sở lý luận

Trong lĩnh vực đo lường và đánh giá giáo dục, chất lượng đề thi thường được phân tích dựa trên các lý thuyết và mô hình thống kê nhằm nâng cao tính khách quan và độ tin cậy của kết quả đánh giá. Trong đó, Lý thuyết kiểm tra cổ điển (Classical Test Theory – CTT) là một cách tiếp cận phổ biến và được ứng dụng rộng rãi trong thực tiễn giáo dục [1], [4]. Theo CTT điểm số quan sát của người học được cấu thành từ điểm thực và sai số đo lường [4]. Trên cơ sở đó, có thể xây dựng các chỉ số phục vụ phân tích chất lượng câu hỏi và đánh giá chất lượng đề thi.

Trong nghiên cứu này, các chỉ số phân tích chính gồm độ khó P, phản ánh mức độ dễ hoặc khó của câu hỏi; độ phân biệt D, phản ánh khả năng phân loại người học; và độ tin cậy Cronbach's Alpha, phản ánh mức độ nhất quán nội tại của toàn bộ đề thi. Bách phân vị được sử dụng như công cụ hỗ trợ xác định nhóm người học có kết quả cao và nhóm người học có kết quả thấp khi tính độ phân biệt. Các nghiên cứu gần đây tiếp tục sử dụng các chỉ số của lý thuyết kiểm tra cổ điển để đánh giá và so sánh đặc tính của các dạng đề thi, bao gồm cả đề thi được xây dựng với sự hỗ trợ của trí tuệ nhân tạo [9].

Trên cơ sở lý thuyết và các chỉ số nêu trên, nghiên cứu lựa chọn cách tiếp cận định lượng để xây dựng phần mềm đánh giá chất lượng đề thi, làm nền tảng cho việc thiết kế thuật toán và các chức năng phân tích dữ liệu khảo thí [2].

2.2. Dữ liệu

Dữ liệu sử dụng trong nghiên cứu được thu thập từ kết quả của sinh viên tại Trường Đại học Công nghiệp Quảng Ninh. Nguồn dữ liệu bao gồm các tệp điểm do giảng viên cung cấp và dữ liệu trích xuất từ hệ thống quản lý đào tạo của Nhà trường. Bộ dữ liệu nghiên cứu gồm các đề

thi thuộc nhiều học phần khác nhau, với tổng số 41 đề thi của 29 học phần thuộc các chương trình đào tạo khác nhau của học kỳ I năm học 2025 - 2026. Dữ liệu bao gồm điểm từng câu đối với đề trắc nghiệm, tự luận và hỗn hợp và điểm tối đa tương ứng của từng câu hỏi. Các đề thi được lựa chọn bảo đảm tính đa dạng về hình thức, bao gồm 04 trắc nghiệm, 19 tự luận và 18 đề thi hỗn hợp.

Trước khi tiến hành phân tích, dữ liệu được xử lý và chuẩn hoá nhằm bảo đảm tính chính xác và đồng nhất. Các bước xử lý bao gồm: loại bỏ các bản ghi thiếu thông tin, chuẩn hoá định dạng dữ liệu, mã hoá kết quả làm bài. Sau quá trình xử lý, dữ liệu được đưa vào phần mềm để thực hiện phân tích và đánh giá chất lượng đề thi.

2.3. Phương pháp nghiên cứu

Nghiên cứu sử dụng phương pháp định lượng dựa trên lý thuyết kiểm tra cổ điển (Classical Test Theory – CTT), một trong những cách tiếp cận phổ biến trong lĩnh vực đo lường và đánh giá giáo dục [1], [4]. CTT được sử dụng rộng rãi để phân tích chất lượng đề thi thông qua các chỉ số thống kê cơ bản, cho phép đánh giá mức độ phù hợp, khả năng phân loại của từng câu hỏi và mức độ nhất quán nội tại của toàn bộ đề thi. Trên cơ sở đó, nghiên cứu áp dụng các phương pháp cụ thể sau:

(1) Phương pháp phân tích độ khó P: Độ khó của câu hỏi được xác định thông qua tỷ lệ người học trả lời đúng trên tổng số người tham gia. Chỉ số này phản ánh mức độ dễ hay khó của câu hỏi trong đề thi. Tùy vào hình thức thi, chỉ số độ khó được xác định như sau:

- *Đối với hình thức thi trắc nghiệm*: Độ khó của câu hỏi/đề thi được xác định thông qua tỷ lệ người học trả lời đúng trên tổng số người tham gia. Chỉ số độ khó được tính theo công thức như sau:

$$P = \frac{n}{N} \cdot 100\% \quad (1)$$

Trong đó: n là số người học trả lời đúng, N là tổng số người tham gia trả lời.

- *Đối với hình thức thi tự luận hoặc hình thức thi khác*: Độ khó của đề thi được xác định thông qua tỷ lệ giữa điểm trung bình của người học và

điểm tối đa có thể đạt được. Công thức được xác định như sau:

$$P = \frac{\bar{x}}{M} \cdot 100\% \quad (2)$$

Trong đó: $\bar{x}$ là điểm trung bình, M là điểm tối đa. Chỉ số P được sử dụng để phân loại câu hỏi theo mức độ từ dễ đến rất khó, làm cơ sở để điều chỉnh cấu trúc đề thi.

(2) Phương pháp phân tích độ phân biệt D: Độ phân biệt được sử dụng để đánh giá khả năng của câu hỏi trong việc phân loại người học theo trình độ. Phương pháp này dựa trên việc chia nhóm người học thành 02 nhóm: nhóm điểm cao và nhóm điểm thấp (lấy 27% trên và 27% dưới theo bách phân vị) [3], [4]. Việc lựa chọn nhóm 27% trên và 27% dưới được kế thừa từ các nghiên cứu kinh điển trong lý thuyết kiểm tra cổ điển, nhằm tối ưu hoá khả năng phân biệt giữa các nhóm người học có trình độ khác nhau. Phương pháp phân nhóm 27% người học có kết quả cao và 27% người học có kết quả thấp vẫn tiếp tục được sử dụng trong các nghiên cứu gần đây về phân tích câu hỏi nhằm xác định độ khó, độ phân biệt và hiệu quả phương án nhiều của câu hỏi trắc nghiệm [8]. Tùy vào hình thức thi, chỉ số độ phân biệt được xác định như sau:

- *Đối với hình thức thi trắc nghiệm*: Độ phân biệt là tỉ số giữa hiệu số người học nhóm có kết quả cao trả lời đúng và người học nhóm có kết quả thấp trả lời đúng với tổng số người học mỗi nhóm. Công thức tính như sau:

$$D = \frac{g_c - g_t}{g} \cdot 100\% \quad (3)$$

Trong đó:

g_c là số người học trả lời đúng thuộc nhóm điểm cao; g_t là số người học trả lời đúng thuộc nhóm điểm thấp; g là số người học trong mỗi nhóm. Chỉ số D càng cao cho thấy câu hỏi có khả năng phân loại tốt.

- *Đối với hình thức thi tự luận hoặc hình thức thi khác*: Độ phân biệt là tỉ số giữa hiệu số điểm trung bình của người học thuộc nhóm có kết quả cao và người học thuộc nhóm có kết quả

thấp với điểm số tối đa của câu hỏi/ đề thi đó.

$$D = \frac{\bar{x}_c - \bar{x}_t}{M} \cdot 100\% \quad (4)$$

Trong đó: $\bar{x}_c$ là điểm trung bình của người học thuộc nhóm có kết quả cao; $\bar{x}_t$ điểm trung bình của người học thuộc nhóm có kết quả thấp; M điểm tối đa của câu hỏi. Chỉ số D càng lớn cho thấy khả năng phân loại của câu hỏi càng cao; ngược lại, giá trị D thấp hoặc âm phản ánh câu hỏi chưa phù hợp và cần được xem xét điều chỉnh.

Bảng 1. Tiêu chuẩn đánh giá độ khó và độ phân biệt của từng câu trong đề thi

Chỉ số	Giá trị	Mức đánh giá
Độ khó P		
$P \geq 80$	Dễ	Cần xem xét giảm
$60 \leq P < 80$	Trung bình	Đạt yêu cầu chuẩn
$20 \leq P < 60$	Khó	Cần điều chỉnh
$P < 20$	Rất khó	Cần xem xét giảm
Độ phân biệt D		
$D \geq 0,4$	Rất tốt	Giữ lại
$0,3 \leq D < 0,4$	Tốt	Giữ lại
$0,2 \leq D < 0,3$	Trung bình	Cần nhắc
$D < 0,2$	Kém	Cần xem xét loại bỏ

Các ngưỡng phân loại trên được sử dụng phổ biến trong nghiên cứu đo lường giáo dục nhằm đánh giá mức độ phù hợp của câu hỏi trong đề thi [3], [4]. Việc kết hợp đồng thời hai chỉ số P và D cho phép đánh giá toàn diện cả mức độ khó và khả năng phân loại của câu hỏi.

(3) Phương pháp đánh giá độ tin cậy của toàn bộ đề thi: Độ tin cậy của toàn bộ đề thi được đánh giá thông qua hệ số Cronbach's Alpha, phản ánh mức độ nhất quán nội tại giữa các câu hỏi hoặc thành phần điểm trong cùng một đề thi [1], [4].

+ *Độ tin cậy toàn đề*:

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum_{i=1}^k \sigma_i^2}{\sigma_X^2} \right) \quad (5)$$

Trong đó: α là độ tin cậy Cronbach's Alpha.

k là số câu hỏi trong đề thi

σ_i^2 là phương sai của câu hỏi thứ i .

σ_x^2 là phương sai tổng điểm toàn bài.

Giá trị của hệ số này dao động từ 0 đến 1; giá trị càng cao cho thấy đề thi càng đáng tin cậy. Thông thường, hệ số Cronbach's Alpha $\geq 0,7$ được xem là chấp nhận được trong nghiên cứu giáo dục. Việc đánh giá độ tin cậy giúp đảm bảo rằng kết quả phân tích đề thi có ý nghĩa và có thể sử dụng làm cơ sở cho các quyết định cải tiến chất lượng đào tạo.

Bảng 2. Tiêu chuẩn đánh giá độ tin cậy của toàn bộ đề thi

Hệ số Cronbach's Alpha	Mức đánh giá	Ý nghĩa
$\alpha \geq 0,9$	Rất cao	Đề thi có độ tin cậy rất tốt, các câu hỏi có mức nhất quán cao
$0,8 \leq \alpha < 0,9$	Tốt	Đề thi có độ tin cậy tốt, có thể sử dụng trong đánh giá
$0,7 \leq \alpha < 0,8$	Chấp nhận được	Đề thi đạt yêu cầu về độ tin cậy trong phân tích giáo dục
$0,6 \leq \alpha < 0,7$	Tạm chấp nhận	Có thể xem xét sử dụng nhưng cần rà soát thêm câu hỏi
$\alpha < 0,6$	Thấp	Đề thi có độ tin cậy chưa đạt, cần điều chỉnh hoặc loại bỏ câu hỏi chưa phù hợp

(4) Phương pháp xây dựng và triển khai phần mềm: Để tự động hoá quá trình phân tích, nghiên cứu sử dụng ngôn ngữ lập trình Python kết hợp với thư viện Pandas và nền tảng Streamlit nhằm xử lý và phân tích dữ liệu khảo thí. Phần mềm được thiết kế theo quy trình khép kín gồm các bước: nhập dữ liệu; kiểm tra, xử lý và chuẩn hóa dữ liệu; xác định nhóm cao – thấp theo bách phân vị; tính độ khó P , độ phân biệt D , độ tin cậy Cronbach's Alpha; và xuất báo cáo kết quả.

Phương pháp này cho phép giảm thiểu sai sót trong quá trình tính toán thủ công, đồng thời nâng cao hiệu quả và tính nhất quán trong đánh giá chất lượng đề thi.

(5) Phương pháp thực nghiệm và so sánh: Phần mềm được thử nghiệm trên dữ liệu thực tế tại Trường Đại học Công nghiệp Quảng Ninh, sau đó so sánh kết quả với phương pháp xử lý thủ công truyền thống. Qua đó, đánh giá mức độ chính xác, tính khả thi và hiệu quả của phần mềm trong thực tiễn. Để đánh giá hiệu quả của phần mềm, nghiên cứu sử dụng các tiêu chí so sánh bao gồm: (i) thời gian xử lý dữ liệu; (ii) mức độ sai lệch kết quả do phần mềm tạo ra và phương pháp tính toán thủ công; (iii) khả năng phát hiện các câu hỏi không đạt yêu cầu theo các ngưỡng đánh giá. Các tiêu chí này được sử dụng nhằm bảo đảm việc đánh giá phần mềm có cơ sở định lượng rõ ràng và khách quan.

3. KẾT QUẢ NGHIÊN CỨU VÀ THẢO LUẬN

3.1. Kết quả xây dựng và triển khai phần mềm phân tích đề thi

Kết quả nghiên cứu cho thấy nhóm tác giả đã xây dựng phần mềm hỗ trợ phân tích chất lượng đề thi dựa trên ba chỉ số cơ bản là độ khó (P), độ phân biệt (D) và độ tin cậy Cronbach's Alpha [3]. Phần mềm có khả năng xử lý linh hoạt 03 dạng đề thi phổ biến gồm trắc nghiệm, tự luận và hỗn hợp thông qua một quy trình tự động hoá từ xử lý số liệu đến hiển thị kết quả.

Bảng 3: Phân bố đề thi thực nghiệm theo khoa và loại đề

STT	Khoa	Số đề đánh giá	Loại đề (TN/TL/hỗn hợp)		
			TN	TL	HH
1	Kinh tế	12	0	0	12
2	CKĐL	03	0	03	0
3	Điện	13	0	13	0
4	KHCB	13	04	03	06
	Tổng	41	04	19	18

Về mặt kỹ thuật, phần mềm được phát triển trên nền tảng Python kết hợp với thư viện Streamlit, cho phép xây dựng giao diện tương

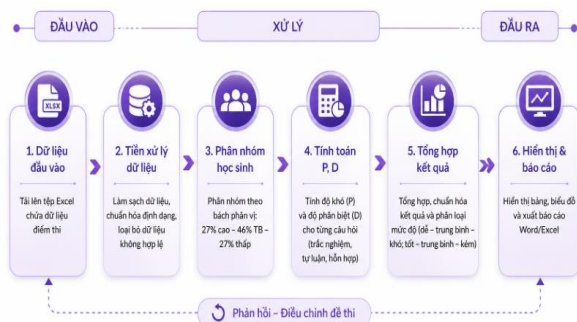
tác trực quan và dễ sử dụng. Người dùng chỉ cần thực hiện các thao tác cơ bản như lựa chọn loại đề thi và tải tệp dữ liệu đầu vào, phần mềm sẽ tự động thực hiện toàn bộ quá trình xử lý và trả về kết quả phân tích chi tiết.

So với phương pháp xử lý thủ công truyền thống, phần mềm cho thấy ưu thế rõ rệt về tốc độ xử lý, độ chính xác và tính nhất quán.

+ Thời gian xử lý: Với tệp dữ liệu 40 câu hỏi và 100 sinh viên, phương pháp thủ công trên Excel mất trung bình 15-20 phút, trong khi phần mềm chỉ mất dưới 02 giây để xuất báo cáo đầy đủ.

+ Độ chính xác: Kết quả do phần mềm tạo ra được đối chiếu với phép tính độc lập trên các bộ dữ liệu kiểm thử và trùng khớp trong phạm vi làm tròn đã lựa chọn.

Việc tự động hoá không chỉ giúp giảm thiểu sai sót trong tính toán mà còn tạo điều kiện chuẩn hoá quy trình phân tích dữ liệu khảo thí. Kết quả này phù hợp với xu hướng ứng dụng công nghệ thông tin trong đánh giá giáo dục, khi các công cụ phân tích định lượng ngày càng đóng vai trò quan trọng trong bảo đảm chất lượng đào tạo [1], [5]. Quy trình của phần mềm được thiết kế theo hướng tự động hoá, bao gồm các bước từ nhập dữ liệu, tiền xử lý, phân tích, đến xuất kết quả. Toàn bộ quy trình được mô tả trong Hình 3.1.



Hình 3.1. Sơ đồ quy trình phân tích chất lượng đề thi

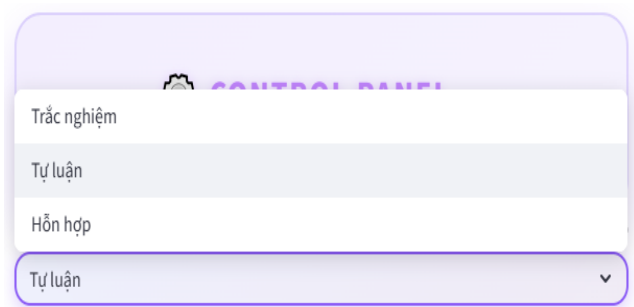
3.2. Kết quả thiết kế giao diện và chức năng

3.2.1. Kết quả thiết kế giao diện người dùng

Phần mềm được thiết kế theo định hướng tối giản hoá thao tác và tối đa hoá khả năng tiếp cận đối với người dùng không chuyên về công nghệ thông tin. Giao diện người dùng được tổ chức theo cấu trúc logic, phản ánh đúng quy

trình nghiệp vụ thực tế trong phân tích đề thi.

Cụ thể, các chức năng chính như lựa chọn loại đề thi, tải dữ liệu và hiển thị kết quả được bố trí rõ ràng và trực quan. Phần mềm hỗ trợ người dùng lựa chọn linh hoạt giữa các dạng đề thi khác nhau thông qua menu thả xuống, đồng thời cho phép nhập dữ liệu dưới dạng tệp Excel – định dạng phổ biến trong thực tiễn giảng dạy. Một điểm đáng chú ý là phần mềm tích hợp cơ chế kiểm tra và chuẩn hoá dữ liệu đầu vào giúp phát hiện và thông báo các lỗi như thiếu cột hoặc sai định dạng, góp phần nâng cao khả năng sử dụng của phần mềm [6], [7].



Hình 3.2. Giao diện lựa chọn loại đề thi

3.2.2. Kết quả hiển thị và trực quan hoá dữ liệu phân tích

Sau khi xử lý, phần mềm trả về kết quả dưới dạng bảng thống kê chi tiết cho từng câu hỏi, bao gồm các chỉ số định lượng như độ khó (P), độ phân biệt (D), điểm trung bình và mức phân loại theo các ngưỡng đánh giá đã thiết lập. Cách tiếp cận này cho phép biểu diễn đồng thời nhiều khía cạnh của chất lượng câu hỏi, hỗ trợ người dùng trong việc đánh giá một cách toàn diện.

Kết quả minh họa trên một đề thi cụ thể (Hình 3.3) cho thấy sự phân hoá rõ rệt giữa các câu hỏi về cả mức độ khó và khả năng phân loại. Một số câu hỏi đạt mức độ khó phù hợp nhưng có độ phân biệt thấp, trong khi một số câu hỏi có khả năng phân loại tốt lại chưa đảm bảo sự cân đối về độ khó [8].

Phát hiện này hàm ý rằng việc thiết kế đề thi không thể dựa trên một chỉ số đơn lẻ, mà cần xem xét đồng thời nhiều tiêu chí nhằm bảo đảm cả tính phù hợp về mức độ khó và hiệu quả phân loại người học. Cách tiếp cận này phù hợp với các nghiên cứu gần đây khi sử dụng đồng thời các chỉ số của CTT để đánh giá và so sánh

chất lượng các dạng đề thi, kể cả đề thi được xây dựng với sự hỗ trợ của AI [9].

Ở góc độ định lượng, kết quả phân tích cũng cho thấy tồn tại một tỷ lệ đáng kể các câu hỏi có độ phân biệt thấp ($D < 0,2$), phản ánh khả năng phân loại hạn chế. Đồng thời, sự phân bố chưa đồng đều của chỉ số độ khó cho thấy cấu trúc đề thi chưa đạt trạng thái tối ưu theo các tiêu chuẩn của CTT. Những bất cập này thường khó nhận diện nếu chỉ dựa trên đánh giá cảm tính hoặc kinh nghiệm giảng dạy. Thông qua phần mềm, các đặc điểm này được thể hiện một cách trực quan và có hệ thống. Nhờ đó giảng viên nhanh chóng xác định các câu hỏi chưa đạt yêu cầu để điều chỉnh hoặc thay thế.

BÁO CÁO ĐÁNH GIÁ P, D ĐỀ THI TỰ LUẬN

1. Kết quả tính độ khó từng câu

STT	Câu hỏi	Tổng số SV	Điểm TB	Điểm tối đa	Điểm cao nhất (thực tế)	Điểm thấp nhất	Độ lệch chuẩn	Độ khó (P)	Mức độ	Điểm TB - Nhóm cao	Điểm TB - Nhóm thấp	Độ phân biệt (D)	Mức độ phân biệt
1	Câu 1	17	2.47	3.0	3.0	2.0	0.51	82.35	Đễ	2.5	2.0	0.17	Kém
2	Câu 2	17	1.74	3.0	3.0	1.0	0.83	57.84	Khó	2.5	1.12	0.46	Rất tốt
3	Câu 3	17	1.68	4.0	4.0	0.5	1.01	41.91	Khó	3.25	0.88	0.59	Rất tốt

2. Đánh giá tổng quan đề thi

Cơ cấu độ khó so với mục tiêu

Nhóm	Số câu	Tỷ lệ thực tế	Tỷ lệ mục tiêu	Khoảng chấp nhận	Đạt nhóm?
Đễ	1	0.3333	0.5	[0.45, 0.55]	False
Trung bình	0	0.0	0.3	[0.25, 0.35]	False
Khó	2	0.6667	0.2	[0.15, 0.25]	False

3. Kết luận

✗ Không đạt chuẩn cơ cấu độ khó ✗ Đạt tiêu chí độ phân biệt

📄 Độ tin cậy đề thi - Cronbach's Alpha

0.7375

Chấp nhận được

✅ Đạt

Hình 3.3. Minh họa kết quả đánh giá tổng quát một đề thi hỗn hợp

Như vậy, phần mềm không chỉ đóng vai trò là công cụ tính toán mà còn là một phương tiện hỗ trợ ra quyết định dựa trên dữ liệu. Kết quả nghiên cứu này khẳng định rằng việc ứng dụng phân tích định lượng trong đánh giá đề thi không chỉ giúp nâng cao tính khách quan mà còn cung cấp những phát hiện có ý nghĩa thực tiễn trong việc cải tiến cấu trúc đề thi và phát triển ngân hàng câu hỏi theo hướng chuẩn hoá.

Kết quả tổng hợp từ 41 đề thi cho thấy sự phân bố các chỉ số đánh giá chưa đồng đều giữa các khối học phần. Về độ phân biệt, tỷ lệ câu hỏi có độ phân biệt kém ($D < 0,2$) chiếm khoảng 27,5% ở khối Kinh tế và lên tới 44,4% ở

khối Cơ khí Động lực. Trong khi đó, tỷ lệ câu hỏi đạt mức tốt ($D \geq 0,3$) có sự chênh lệch lớn giữa các khối, chỉ đạt khoảng 11,1% ở khối Cơ khí Động lực nhưng lên tới 91,1% ở khối Khoa học cơ bản.

Về độ khó, tỷ lệ câu hỏi ở mức trung bình ($60 \leq P < 80$) dao động từ khoảng 19% ở khối Kinh tế đến 25,7% ở khối Điện. Tuy nhiên, vẫn tồn tại tỷ lệ đáng kể các câu hỏi quá dễ (ví dụ: khối Khoa học cơ bản chiếm khoảng 60%) hoặc quá khó (ví dụ: khối Điện chiếm khoảng 42,9%). Điều này cho thấy cấu trúc đề thi chưa thực sự cân đối theo cơ cấu mục tiêu do nhóm nghiên cứu thiết lập trong phạm vi thử nghiệm (50% dễ – 30% trung bình – 20% khó).

Sự khác biệt về độ khó và độ phân biệt giữa các khối học phần có thể xuất phát từ đặc điểm không đồng đều về mục tiêu, nội dung kiến thức, hình thức thi, số lượng câu hỏi và quy mô lớp học phần. Đề tự luận thường có số câu ít và thang điểm khác với đề trắc nghiệm hoặc hỗn hợp, trong khi quy mô mẫu nhỏ ở một số lớp cũng có thể làm chỉ số độ phân biệt biến động.

Bên cạnh đó, một số câu hỏi có thể chưa bám sát ma trận đề, chuẩn đầu ra học phần và mức độ nhận thức cần đánh giá. Vì vậy, các bộ môn cần tăng cường xây dựng, rà soát và chuẩn hóa câu hỏi theo chuẩn đầu ra, bảo đảm sự phù hợp giữa nội dung, mức độ yêu cầu, đáp án và thang điểm; đồng thời theo dõi kết quả phân tích qua nhiều lớp học phần và nhiều đợt thi để tiếp tục cải tiến chất lượng đề thi. Kết quả mô tả cho thấy có sự khác biệt về phân bố các chỉ số giữa các nhóm dữ liệu. Tuy nhiên, do số lượng đề, hình thức thi, quy mô lớp và cấu trúc câu hỏi giữa các khoa chưa đồng đều, kết quả chưa đủ để kết luận về sự khác biệt chất lượng đề thi giữa các lĩnh vực đào tạo.

4. KẾT LUẬN

Nghiên cứu đã xây dựng và triển khai thành công phần mềm hỗ trợ phân tích chất lượng đề thi dựa trên các chỉ số định lượng cơ bản của lý thuyết kiểm tra cổ điển. Phần mềm cho phép xử lý linh hoạt nhiều hình thức thi khác nhau, đồng thời tự động hoá toàn bộ quy trình từ nhập dữ liệu, tiền xử lý đến tính toán và hiển thị kết quả phân tích. Phần mềm không chỉ là công cụ tính

toán mà còn là công cụ hỗ trợ chuẩn hoá và sàng lọc chất lượng câu hỏi.

Kết quả thực nghiệm cho thấy phần mềm hoạt động ổn định, cung cấp các chỉ số rõ ràng và có ý nghĩa trong việc đánh giá chất lượng câu hỏi. Thông qua phân tích minh họa, nghiên cứu chỉ ra rằng độ khó phù hợp chưa đảm bảo khả năng phân loại tốt, qua đó nhấn mạnh sự cần thiết của việc kết hợp nhiều chỉ số trong đánh giá đề thi.

Về mặt thực tiễn, phần mềm góp phần giảm đáng kể khối lượng công việc thủ công, nâng cao tính khách quan và nhất quán trong công

tác khảo thí. Đồng thời, nghiên cứu cung cấp một công cụ hữu ích hỗ trợ giảng viên trong việc cải tiến đề thi và phát triển ngân hàng câu hỏi theo hướng chuẩn hoá.

Trong thời gian tới, phần mềm cần được tiếp tục hoàn thiện theo hướng xây dựng cơ sở dữ liệu tập trung, lưu trữ lịch sử phân tích qua nhiều học kỳ và liên kết câu hỏi với chuẩn đầu ra học phần. Bên cạnh đó, nghiên cứu có thể mở rộng theo hướng phân tích phương án nhiễu, mức độ nhận thức của câu hỏi và theo dõi sự thay đổi chất lượng đề thi qua nhiều năm học.

TÀI LIỆU THAM KHẢO

- [1]. Nguyễn Công Khanh (2013), *Đo lường và đánh giá trong giáo dục*, Nhà xuất bản Giáo dục Việt Nam, Hà Nội.
- [2]. Trần Kiều (2011), *Đánh giá kết quả học tập của học sinh*, Nhà xuất bản Đại học Sư phạm, Hà Nội.
- [3]. Haladyna, T.M. (2004), *Developing and Validating Multiple-Choice Test Items*, Lawrence Erlbaum.
- [4]. Ebel, R. L. and Frisbie, D. A. (1991), *Essentials of Educational Measurement*, Prentice Hall, New Jersey.
- [5]. Few, S. (2009), *Now You See It: Simple Visualization Techniques for Quantitative Analysis*, Analytics Press, Oakland.
- [6]. Nielsen, J. (1993), *Usability Engineering*, Academic Press.
- [7]. Shneiderman, B. (2010), *Designing the User Interface*, Pearson.
- [8]. Rezigalla, A. A., Eleragi, A. M. E. S. A., Elhussein, A. B., et al. (2024), *Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items*, BMC Medical Education, Vol. 24, Article 445. DOI: 10.1186/s12909-024-05433-y.
- [9]. O, K. M. (2024), *A comparative study of AI-human-made and human-made test forms for a university TESOL theory course*, Language Testing in Asia, Vol. 14, Article 19. DOI: 10.1186/s40468-024-00291-3.

Thông tin của tác giả:

Th.s Nguyễn Thu Hiền

Phòng Thanh tra và KĐCL, Trường Đại học Công nghiệp Quảng Ninh

Điện thoại: +(84).862078861 - Email: hientn123@gmail.com

Th.s Nguyễn Thị Phương

Trưởng phòng Thanh tra &ĐBCL, Trường Đại học Công nghiệp Quảng Ninh

Điện thoại: +(84).904544205 - Email: phuongnt@qui.edu.vn

Th.s Trương Thị Khánh Ly

Phó phòng Thanh tra &ĐBCL, Trường Đại học Công nghiệp Quảng Ninh

Điện thoại: +(84).337730694 - Email: khanhly6381@gmail.com

Th.s Phạm Ngọc Hải

Bộ môn toán, KHCB, Trường Đại học Công nghiệp Quảng Ninh

Điện thoại: +(84).389153242 - Email: ngochaiqn87@gmail.com

DEVELOPMENT AND EXPERIMENTAL IMPLEMENTATION OF SOFTWARE SUPPORTING QUANTITATIVE EXAM QUALITY ASSESSMENT AT QUANG NINH UNIVERSITY OF INDUSTRY**Information about authors:****Nguyen Thu Hien**, Master, Quang Ninh University of Industry. Email: hientn123@gmail.com**Nguyen Thi Phuong**, Master, Quang Ninh University of Industry. Email: phuongnt@qui.edu.vn**Trương Thị Khánh Ly**, Master, Quang Ninh University of Industry. Email: khanhly6381@gmail.com**Phạm Ngọc Hải**, Master, Quang Ninh University of Industry. Email: ngochaiqn87@gmail.com**ABSTRACT:**

In the context of higher education institutions increasingly reforming assessment practices in alignment with learning outcomes, developing a tool to support the evaluation of examination quality based on scientific and objective criteria has become necessary. This paper presents the development and experimental implementation of software designed to support the evaluation of course examination quality at Quang Ninh University of Industry based on Classical Test Theory (CTT). The software employs difficulty, discrimination, and Cronbach's Alpha reliability indices; percentile ranking is used to identify high-performing and low-performing groups when calculating the discrimination index.

The software was developed in Python, using Pandas and NumPy for data processing and Streamlit for the user interface. The software supports the validation and standardization of input data, calculation of indices, classification of test items, identification of cases requiring review, and generation of analytical reports. Through these functions, the software contributes to improving the effectiveness of assessment activities and provides evidence to support the evaluation of course learning outcomes at Quang Ninh University of Industry. The software was tested on 41 examinations from 29 courses across different academic programs in the first semester of the 2025–2026 academic year, including 4 multiple-choice examinations, 19 essay examinations, and 18 mixed-format examinations.

The experimental results show that the software can process all three examination formats, automatically calculate the relevant indices, and provide information for identifying items that are overly easy, overly difficult, have low discrimination, or show abnormal patterns. However, for classes with fewer than 10 students, the discrimination index should be regarded only as a reference indicator. Decisions on whether to retain, revise, or remove an item should be based on a combined consideration of difficulty, discrimination, reliability, and the professional judgment of lecturers and academic departments.

Keywords: Classical Test Theory (CTT); difficulty index; discrimination index; Cronbach's Alpha reliability; examination analysis; Python; Streamlit.

REFERENCES

- [1]. Nguyen Cong Khanh (2013), *Measurement and Evaluation in Education*, Vietnam Education Publishing House, Hanoi.
- [2]. Tran Kieu (2011), *Assessment of Students' Learning Outcomes*, Hanoi National University of Education Publishing House, Hanoi.
- [3]. Haladyna, T.M. (2004), *Developing and Validating Multiple-Choice Test Items*, Lawrence

Erlbaum.

- [4]. Ebel, R. L. and Frisbie, D. A. (1991), *Essentials of Educational Measurement*, Prentice Hall, New Jersey.
- [5]. Few, S. (2009), *Now You See It: Simple Visualization Techniques for Quantitative Analysis*, Analytics Press, Oakland.
- [6]. Nielsen, J. (1993), *Usability Engineering*, Academic Press.
- [7]. Shneiderman, B. (2010), *Designing the User Interface*, Pearson.
- [8]. Rezigalla, A. A., Eleragi, A. M. E. S. A., Elhussein, A. B., et al. (2024), *Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items*, BMC Medical Education, Vol. 24, Article 445. DOI: 10.1186/s12909-024-05433-y.
- [9]. O, K. M. (2024), *A comparative study of AI-human-made and human-made test forms for a university TESOL theory course*, Language Testing in Asia, Vol. 14, Article 19. DOI: 10.1186/s40468-024-00291-3.

Ngày nhận bài: 26/06/2026;

Ngày nhận bài sửa: 14/07/2026;

Ngày chấp nhận đăng: 14/07/2026.