TẠP CHÍ KHOA HỌC TRƯỜNG ĐẠI HỌC HOA LƯ
ISSN 2615-9538
Website: http://hluv.edu.vn/vi/tckh

SỬ DỤNG PHẦN MỀM IATA ĐỂ PHÂN TÍCH CÂU HỎI TRẮC NGHIỆM KHÁCH QUAN DỰA TRÊN NỀN TẢNG LÝ THUYẾT KHẢO THÍ CỔ ĐIỂN VÀ LÝ THUYẾT KHẢO THÍ HIỆN ĐẠI

Trần Thị Hà Tâm[1], Nguyễn Hữu Tiến[1]

Ngày nhận bài: 24/4/2023 *Ngày chấp nhận đăng: 25/5/2023*

***Tóm tắt**: Bài viết giới thiệu và ứng dụng phần mềm IATA, phần mềm phân tích đánh giá chất lượng câu hỏi trắc nghiệm khách quan (TNKQ) với các tham số đặc trưng cơ bản dựa trên 2 nền tảng là Lý thuyết khảo thí cổ điển (CTT) và Lý thuyết khảo thí hiện đại (IRT). Bài viết đưa ra kết quả phân tích 40 câu hỏi TNKQ, mã đề 03 học phần Kinh tế chính trị Mác-Lênin với 106 sinh viên khóa D14 của trường Đại học Hoa Lư năm học 2021-2022. Phần mềm IATA sẽ phân tích các câu hỏi TNKQ dựa trên cả 2 nền tảng CTT và IRT với các tham số như độ khó, độ phân biệt, hệ số tương quan, ước lượng năng lực của người học…Dựa trên các kết quả đó đề xuất lựa chọn, điều chỉnh hoặc loại bỏ những câu hỏi TNKQ để xây dựng các câu hỏi phù hợp với năng lực người học và mục đích của kỳ thi.*

***Từ khóa:** phần mềm IATA, lý thuyết khảo thí cổ điển, lý thuyết khảo thí hiện đại, câu hỏi trắc nghiệm khách quan*

USING IATA SOFTWARE TO ANALYZE MULTIPLE CHOICE QUESTIONS BASED ON CLASSICAL TEST THEORY AND ITEM RESPONSE THEORY

***Abstract:** The objective of this article presents and applies IATA sofware used to analyze and evaluate multiple choice questions based on CTT and IRT with characteristic parameters such as the difficulty, the discrimination and correlation coefficients of questions. The article also gives the results of analysis of 40 multiple choice questions of "Marxist Political Economy" for the final test of 106 students at Hoa Lu University in the academic year 2021-2022 and discovers good items and unsatisfactory items to adjust or remove. These results can be used not only to analyze and select multiple choice items, but also to improve the quality of multiple choice test items to build a test suitable for ability of student and for the exam.*

***Keywords:** IATA software, Classical Test Theory, Item Response Theory, Multiple choice questions*

1. Giới thiệu

Hiện nay có thể sử dụng đa dạng nhiều phương pháp đánh giá kết quả học tập của người học như là tự luận, vấn đáp, TNKQ, thực hành trên máy, bài tập lớn…với mỗi phương pháp có những thuận lợi và khó khăn khác nhau. Trong đó xu hướng sử dụng câu hỏi TNKQ cũng được sử dụng rộng rãi bởi vì câu hỏi TNKQ với số lượng câu hỏi lớn sẽ bao quát được các kiến thức của chương trình, thời gian chấm bài nhanh, khách quan và hạn chế tiêu cực trong quá trình chấm thi. Sử dụng câu hỏi TNKQ đảm bảo yêu cầu sẽ có tính định lượng cao, cung cấp số liệu chính xác, ổn định và có thể áp dụng công nghệ đo lường hiện đại trong việc phân tích xử lý và nâng cao chất lượng câu hỏi thi [1]. Tuy nhiên vấn đề đặt ra là làm thế nào chúng ta biết có thể biên soạn và thẩm định chất lượng một câu hỏi, một đề thi TNKQ đảm bảo chất lượng và phù hợp với các mục tiêu kiểm tra đánh giá?

Vấn đề này từ rất lâu đã được các nhà giáo dục nghiên cứu và tìm hiểu giúp chúng ta có cái nhìn tổng quan về đo lường và đánh giá trong giáo dục, các phương pháp đánh giá kết quả học tập bằng

[1] Phòng Quản lý chất lượng, Trường Đại học Hoa Lư; Email: tthtam@hluv.edu.vn

TNKQ và ứng dụng khoa học đo lường vào thực tiễn. Ví dụ như công trình nghiên cứu "Evaluation to improve learning" của Benjamin S.Bloom, George F.Madaus và Thomas J.Hasting nhằm hỗ trợ, tư vấn người dạy các kỹ thuật để đánh giá kết quả học tập của người học từ đó người dạy sẽ sử dụng việc đánh giá như công cụ để cải tiến quá trình dạy và học [2]. Lâm Quang Thiệp đã nghiên cứu "Những cơ sở của kỹ thuật trắc nghiệm" và "Đo lường trong giáo dục. Lý thuyết và Ứng dụng" [1],[3] và đây là các nghiên cứu về TNKQ và các nội dung liên quan đến đo lường và đánh giá trong giáo dục. Lý thuyết khảo thí cổ điển (Classical Test Theory- CTT) ra đời vào cuối thế kỷ XIX và những năm đầu của thế kỷ XX được ứng dụng chủ yếu trong phân tích, đánh giá câu hỏi TNKQ dựa trên các tham số như độ khó, độ phân biệt và hệ số tương quan của câu hỏi với đề thi sau khi có phản hồi của các thí sinh với các câu hỏi. Tuy nhiên hạn chế của CTT là không tách biệt được các đặc trưng của thí sinh độc lập với đặc trưng của đề thi. Do vậy Lý thuyết khảo thí hiện đại (Item Response Theory- IRT) ra đời là kết quả kế thừa và phát triển CTT và được xây dựng trên mô hình toán học mô tả xác suất làm đúng câu hỏi phụ thuộc vào năng lực người dự thi và các tham số đặc trưng câu hỏi. Lý thuyết này đòi hỏi nhiều tính toán nhưng nhờ sự phát triển của máy tính vào những năm 60 của thế kỷ XX nên lý thuyết phát triển nhanh chóng và đạt những thành tựu quan trọng giúp ngành khoa học đo lường và đánh giá phát triển mạnh mẽ [4]

Hiện nay ở các trường đại học trên cả nước nói chung và Đại học Hoa Lư nói riêng hầu như chưa áp dụng công cụ để xử lý, đánh giá quá trình biên soạn và phân tích các đề thi tự luận hay TNKQ một cách khoa học và chính xác. Nếu một câu hỏi TNKQ không được biên soạn tốt, không được phân tích đánh giá thì sẽ xảy ra khả năng một số câu hỏi kém chất lượng vẫn được sử dụng trong các bài thi sau sẽ không đánh giá chính xác năng lực của người học và có thể ảnh hưởng đến thành tích học tập của thí sinh. Vì vậy cần thiết phải có công cụ đánh giá câu hỏi thi, đề thi để xây dựng một đề thi có các câu hỏi thi có chất lượng, đảm bảo độ tin cậy và đo lường được năng lực người học. Hiện nay có rất nhiều phần mềm được sử dụng để phân tích chất lượng đề thi và đánh giá chất lượng câu hỏi thi được sử dụng rộng rãi như là: Phần mềm Quest (2000), Conquest (2020) của Úc, phần mềm Winsteps, Parscale của Mỹ, phần mềm Vitesta của Việt Nam… và các nhiều nghiên cứu sử dụng các phần mềm này để phân tích đánh giá chất lượng câu hỏi thi TNKQ như là bài viết "Sử dụng phần mềm Quest để phân tích câu hỏi TNKQ" của Nguyễn Hoàng Bảo Thanh [5], bài viết "Phân tích và lựa chọn câu hỏi TNKQ dựa trên lý thuyết trắc nghiệm cổ điển và lí thuyết ứng đáp câu hỏi" của Nguyễn Văn Cảnh và Nguyễn Phước Hải [6], bài viết "Tính toán và so sánh độ khó của câu hỏi theo các lí thuyết khảo thí cổ điển-hiện đại bằng các phần mềm CETA/R" của Vũ Đỗ Long, Nguyễn Văn Dũng, Vũ Thị Thảo, Nguyễn Thị Mỹ Linh [7]….Các nghiên cứu này sử dụng các phần mềm để phân tích, đánh giá đề thi, câu hỏi thi dưới nhiều góc độ khác nhau tuy nhiên các bài viết chưa kết hợp phân tích các tham số đặc trưng của câu hỏi theo cả hai lý thuyết khảo thí và phân tích đường cong đặc trưng của câu hỏi để đánh giá chất lượng từng câu hỏi. Do đó bài viết này giới thiệu và áp dụng phần mềm IATA để phân tích chất lượng câu hỏi TNKQ nhằm giúp mọi người có góc nhìn tổng quan về các tham số đặc trưng của câu hỏi TNKQ dựa trên cả hai nền tảng CTT và IRT đồng thời thông qua các biểu đồ minh họa có thể phân tích chính xác hơn từng câu hỏi. Kết quả nghiên cứu sẽ giúp người biên soạn đề thi lựa chọn được những câu hỏi thực sự có chất lượng và phát hiện những câu hỏi chưa đạt yêu cầu, cần phải xem xét lại trước khi sử dụng hoặc loại bỏ đồng thời các giảng viên có thể áp dụng phần mềm này trong quá trình biên soạn đề thi và đánh giá kết quả học tập của người học để nâng cao chất lượng của các câu hỏi thi TNKQ.

2. Nội dung

2.1. Giới thiệu phần mềm IATA

Phần mềm IATA (Item and Test Analysis) là phần mềm phân tích các câu hỏi TNKQ dựa trên nền tảng CTT và IRT để tìm ra các câu hỏi TNKQ có chất lượng phù hợp với năng lực người học và mục đích kiểm tra đánh giá. Nghiên cứu sử dụng phần mềm này bởi IATA có nhiều ưu điểm hơn so với các phần mềm khác có cùng chức năng, cụ thể như sau:

Phần mềm IATA đưa ra kết quả phân tích cụ thể cho từng câu hỏi riêng lẻ với từng bước phân tích các thông số như là kết quả trả lời của thí sinh theo từng nhóm năng lực, từng phương án trả lời của mỗi câu hỏi và hiển thị lựa chọn câu hỏi tối ưu, giải thích kết quả của từng câu hỏi và đề xuất các nội dung cần điều chỉnh để câu hỏi tối ưu hơn [8]

Phần mềm IATA được cung cấp miễn phí, người dùng có thể tải phần mềm từ địa chỉ https: polymetrika.com/Downloads/IATA và tải phần mềm về máy tính để sử dụng. Phần mềm có hiển thị ngôn ngữ bằng nhiều thứ tiếng trong đó có Tiếng Việt và các file dữ liệu đầu vào có thể được nhập

bằng file excel, Access, SPSS nên thuận tiện dễ dàng sử dụng. Việc tiếp cận phần mềm cũng đơn giản hơn các phần mềm khác.

Phần mềm có giao diện trực quan, sử dụng bằng bảng chọn và chuột nên thao tác cũng đơn giản (Phần mềm Quest yêu cầu người dùng viết câu lệnh sẽ khó hơn)

Phần mềm đưa ra các chỉ dẫn đề xuất lựa chọn câu hỏi TNKQ theo biểu tượng hình ảnh của câu hỏi trong kết quả phân tích. Cụ thể: câu hỏi có biểu tượng *hình tròn màu xanh* (câu hỏi không có vấn đề gì lớn và có thể sử dụng ngay), *hình thoi màu vàng* (câu hỏi ít tối ưu hơn, đề xuất cần sửa đổi một chút kỹ thuật hay nội dung, tuy nhiên câu hỏi cũng không ảnh hưởng đáng kể trong kết quả phân tích) và *hình tam giác màu đỏ* (câu hỏi có vấn đề về dữ liệu, thông số kỹ thuật hay nội dung, đề xuất loại bỏ câu hỏi hoặc kiểm tra, điều chỉnh thật kỹ trước khi sử dụng) [3]

Một điểm khác biệt lớn của phần mềm IATA so với các phần mềm khác là phần mềm có thể phân tích câu hỏi TNKQ dựa trên nền tảng CTT kết hợp IRT với đầy đủ tính năng cần thiết của một phần mềm thống kê phân tích câu hỏi thi TNKQ như là độ khó, độ phân biệt, hệ số tương quan, ước lượng năng lực thực sự của thí sinh…, điều này giúp việc phân tích và lựa chọn câu hỏi TNKQ đầy đủ và chính xác hơn. Sau đây bài viết giới thiệu tổng quan về các tham số đặc trưng của câu hỏi có liên quan khi sử dụng phần mềm IATA

2.2. Các tham số đặc trưng câu hỏi theo lý thuyết khảo thí cổ điển

2.2.1. Độ khó

Độ khó của câu hỏi (p) là tỉ lệ thí sinh trả lời đúng câu hỏi đó trên tổng số thí sinh dự thi. Ta có công thức tính độ khó:

$$p = \frac{\text{Tổng thí sinh làm đúng}}{\text{Tổng thí sinh tham gia trả lời}}$$

Giá trị p nằm trong khoảng từ 0 đến 1, giá trị p càng bé thì độ khó của câu hỏi càng cao và ngược lại. Thông thường độ khó của câu hỏi có thể chấp nhận được khi giá trị p đạt giá trị trong khoảng 0.25 đến 0.75 tương ứng với số thí sinh trả lời đúng đạt từ 25% đến 75%. Câu hỏi có độ khó lớn hơn 0.75 (tương ứng trên 75% thí sinh trả lời đúng) là quá dễ, có độ khó nhỏ hơn 0.25(tương ứng dưới 25% thí sinh trả lời đúng) là quá khó [1].

2.2.2. Độ phân biệt

Độ phân biệt của câu hỏi TNKQ (D) là khả năng của câu hỏi trắc nghiệm thực hiện được sự phân biệt giữa các nhóm thí sinh có năng năng lực khác nhau. Độ phân biệt của một câu hỏi liên quan đến độ khó của câu hỏi bởi vì nếu một câu hỏi quá khó hay quá dễ thì phản ứng của thí sinh có năng lực tốt hay kém đều giống nhau, hoặc là trả lời đúng hết hoặc trả lời sai hết, như vậy thì câu hỏi đó không phân biệt được thí sinh. Một câu hỏi TNKQ được coi là có độ phân biệt tốt thì câu hỏi có độ khó ở mức trung bình, khi thí sinh trả lời câu hỏi đó thì nhóm thí sinh có năng lực cao phải có tỉ lệ làm đúng câu hỏi cao hơn những nhóm thí sinh có năng lực thấp [1]. Độ phân biệt được chia ra các mức sau: $D \geq 0.4$: Câu hỏi có độ phân biệt rất tốt; $0.30 \leq D \leq 0.39$: Câu hỏi có độ phân biệt tốt, nếu thay đổi một chút càng tốt; $0.20 \leq D \leq 0.29$: Câu hỏi có độ phân biệt có thể chấp nhận được nhưng cần thay đổi một chút; $D \leq 0.19$: Câu hỏi có độ phân biệt chưa đạt yêu cầu, phải loại bỏ hoặc thay đổi để nâng cao độ phân biệt [4]

2.2.3. Hệ số tương quan giữa điểm của câu hỏi với điểm cả bài trắc nghiệm

Hệ số tương quan (*Pbis)* là một đại lượng để đo mối quan hệ tuyến tính giữa hai biến ngẫu nhiên là điểm câu hỏi TNKQ và điểm toàn bài thi. Hệ số tương quan của câu hỏi TNKQ có giá trị từ -1 đến 1 [2]. Hệ số tương quan giữa 2 biến định lượng được chia theo các mức sau: $0.8 \leq$ Pbis ≤ 1: Tương quan cao đáng tin cậy; $0.6 \leq$ Pbis ≤ 0.79: Tương quan vừa phải; $0.4 \leq$ Pbis ≤ 0.59: Tương quan tạm được; $0.2 \leq$ Pbis ≤ 0.39: Tương quan ít; $0 \leq$ Pbis ≤ 0.19: Tương quan không đáng kể; Pbis < 0: Tương quan nghịch [9]. Do vậy cần giữ lại những câu hỏi có mối tương quan chặt chẽ, xem xét điều chỉnh những câu hỏi có mối tương quan quá thấp hoặc loại bỏ nếu tương quan nghịch để tăng độ tin cậy của đề thi

2.3. Các tham số đặc trưng câu hỏi theo lý thuyết khảo thí hiện đại

Với IRT sử dụng mô hình toán học để dự đoán xác xuất trả lời đúng một câu hỏi, dựa trên chỉ số năng lực của người trả lời và độ khó của câu hỏi. Xác suất trả lời đúng một câu hỏi của các đối tượng trả lời khác nhau được biểu diễn dưới một đường cong gọi là đường cong đặc tính câu hỏi (Item

Characteristic Curve -ICC) [10]. Câu hỏi được đặc trưng bởi 3 tham số là độ khó, độ phân biệt và khả năng đoán mò của thí sinh, tương ứng với các tham số đó là các mô hình ứng đáp câu hỏi. Có hai mô hình IRT thường được sử dụng là mô hình ứng đáp một tham số (Mô hình Rash) và mô hình ứng đáp hai tham số

2.3.1 Mô hình ứng đáp một tham số

Trong mô hình một tham số, độ khó của câu hỏi là đại lượng đặc trưng cho khả năng trả lời đúng câu hỏi của thí sinh. Câu hỏi có độ khó càng cao thì xác suất trả lời đúng câu hỏi đó của thí sinh càng thấp và ngược lại. Do đó xác suất trả lời đúng của câu hỏi được tính theo công thức sau: $P(\theta) = \frac{e^{(\theta-b)}}{1+e^{(\theta-b)}}$ (trong đó: θ là năng lực của thí sinh, *b* là độ khó của câu hỏi). Tham số độ khó *b* của câu hỏi có thể đạt từ giá trị từ -∞ đến +∞. Với những câu hỏi có giá trị tham số *b* quá thấp hoặc quá cao thường không có ý nghĩa trong việc đo lường năng lực của thí sinh nên những câu hỏi TNKQ có giá trị độ khó từ -3,0 đến 3,0 nên được đưa vào sử dụng còn những câu hỏi có giá trị tham số độ khó nằm ngoài khoảng trên cần phải loại bỏ hoặc điều chỉnh, xem xét kỹ trước khi đưa vào sử dụng [11]

2.3.2 Mô hình ứng đáp hai tham số

Theo IRT không chỉ có độ khó và năng lực của thí sinh mà độ phân biệt cũng ảnh hưởng đến xác suất trả lời đúng của thí sinh, câu hỏi TNKQ có độ phân biệt càng lớn thì sự chênh lệch về xác suất trả lời đúng giữa các thí sinh có năng lực cao và năng lực thấp càng cao. Xác suất trả lời đúng của câu hỏi được tính theo công thức sau: $P(\theta) = \frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}}$ (trong đó: *a* là độ phân biệt, *b* là độ khó của câu hỏi). Tham số độ phân biệt *a* của câu hỏi có thể đạt từ giá trị từ -∞ đến +∞. Với những câu hỏi có giá trị tham số *a* quá thấp hoặc quá cao thường không có ý nghĩa trong việc phân loại năng lực của thí sinh nên những câu hỏi TNKQ có giá trị độ phân biệt từ 0,5 đến 2 nên được sử dụng còn những câu hỏi có giá trị tham số độ phân biệt nằm ngoài khoảng trên cần phải loại bỏ hoặc điều chỉnh, xem xét kỹ trước khi đưa vào sử dụng [11].

2.4. Phân tích 40 câu hỏi TNKQ của mã đề 03 học phần Kinh tế chính trị Mac-Lênin

Kết quả phân tích câu trả lời của 106 sinh viên đối với 40 câu hỏi TNKQ của mã đề 3 học phần "Kinh tế chính trị Mác - Lênin" trong kỳ thi kết thúc học phần của trường Đại học Hoa Lư năm học 2021-2022 bằng phần mềm IATA được thể hiện ở Hình 1 dưới đây:

Use	O	Name	Discr	PVal	PBis	a	b
☑	▲	C1	0,08	0,94	0,26	0,93	-1,89
☑	◆	C2	0,32	0,56	0,20	0,34	-0,38
☑	▲	C3	0,17	0,91	0,24	0,15	-12,77
☑	◆	C4	0,27	0,72	0,24	0,25	-1,95
☑	▲	C5	0,13	0,85	0,14	0,09	-26,51
☑	●	C6	0,63	0,50	0,43	0,47	0,15
☑	▲	C7	0,13	0,88	0,22	0,12	-32,83
☑	▲	C8	0,16	0,95	0,18	0,23	-9,36
☑	▲	C9	0,34	0,78	0,34	0,12	-12,46
☑	●	C10	0,39	0,61	0,38	0,53	-0,69
☑	◆	C11	0,35	0,69	0,29	0,43	-0,94
☑	●	C12	0,56	0,75	0,44	0,57	-1,27
☑	●	C13	0,53	0,70	0,44	0,58	-0,96
☑	◆	C14	0,28	0,90	0,32	0,81	-1,69
☑	▲	C15	0,25	0,81	0,26	0,15	-4,91
☑	◆	C16	0,16	0,93	0,29	0,81	-2,01
☑	◆	C17	0,11	0,77	0,12	0,41	-2,02
☑	▲	C18	0,07	0,76	0,12	0,06	-87,08
☑	●	C19	0,61	0,61	0,55	0,63	-0,49
☑	●	C20	0,74	0,49	0,55	0,82	-0,01

Use	O	Name	Discr	PVal	PBis	a	b
☑	◆	C21	0,27	0,75	0,23	0,35	-1,76
☑	▲	C22	0,34	0,76	0,30	0,13	-12,20
☑	◆	C23	0,17	0,86	0,26	0,62	-1,81
☑	●	C24	0,55	0,57	0,42	0,52	-0,35
☑	▲	C25	0,08	0,97	0,20	1,09	-2,19
☑	▲	C26	0,05	0,41	0,08	0,12	4,93
☑	●	C27	0,63	0,42	0,51	0,62	0,39
☑	▲	C28	0,69	0,64	0,50	0,11	-2,45
☑	●	C29	0,45	0,70	0,39	0,51	-1,18
☑	●	C30	0,46	0,60	0,41	0,50	-0,54
☑	●	C31	0,44	0,45	0,40	0,40	0,56
☑	▲	C32	0,38	0,67	0,22	0,09	-3,53
☑	▲	C33	0,62	0,57	0,47	0,12	-9,19
☑	●	C34	0,82	0,41	0,64	0,78	0,44
☑	●	C35	0,42	0,67	0,37	0,52	-0,91
☑	▲	C36	0,28	0,66	0,22	0,06	-28,36
☑	◆	C37	0,25	0,88	0,27	0,63	-1,89
☑	◆	C38	0,33	0,58	0,32	0,45	-0,51
☑	▲	C39	0,65	0,57	0,57	0,14	-7,21
☑	▲	C40	0,26	0,75	0,21	0,04	-39,30

Hình 1. *Kết quả phân tích 40 câu hỏi TNKQ bằng phần mềm IATA*

Trong Hình 1 cho biết kết quả phân tích bằng IATA của 40 câu hỏi TNKQ như sau:

- Cột thứ 1 (*Use*): có các dấu "v" cho phép lựa chọn hoặc loại bỏ bất kỳ câu hỏi nào ra hay vào bảng phân tích kết quả các câu hỏi

- Cột thứ 2 (*0*) là các biểu tượng chất lượng của các câu hỏi theo 3 nhóm:

+ Nhóm biểu tượng hình tròn màu xanh gồm những câu hỏi tốt, không có vấn đề nghiêm trọng và có thể sử dụng được ngay (gồm 13 câu hỏi là 6,10,12,13,19,20,24,27,29,30,31,34,35)

+ Nhóm biểu tượng hình thoi màu vàng gồm các câu hỏi ít tối ưu hơn so với câu hỏi hình tròn màu xanh và cần kiểm tra lại một chút trước khi sử dụng (gồm 10 câu hỏi là 2, 4, 11, 14, 16, 17, 21, 23, 37, 38)

+ Nhóm biểu tượng hình tam giác màu đỏ gồm những câu hỏi chưa đạt yêu cầu, có khả năng xảy ra vấn đề trong quá trình thiết kế cần loại bỏ hoặc phải được xem xét thật kĩ trước khi sử dụng (gồm 17 câu hỏi là 1,3,5,7,8,9,15,18,22,25,26,28,32,33,36,39,40)

- Cột thứ 3 *(Name)* là tên của 40 câu hỏi trong đề thi
- Cột thứ 4 (*Discr*) và thứ 5 (*Pval)* là độ phân biệt và độ khó cuả câu hỏi được phân tích theo CTT
- Cột thứ 6 (*PBis*) là hệ số tương quan của các câu hỏi TNKQ với bài thi: Kết quả tất cả các *PBis* > 0 chứng tỏ điểm của các câu hỏi và điểm của bài thi có mối tương quan
- Cột thứ 7 (*a*) và thứ 8 (*b)* là độ phân biệt và độ khó cuả câu hỏi được phân tích theo IRT

2.4.1 Kết quả phân tích câu hỏi có biểu tượng hình tròn màu xanh

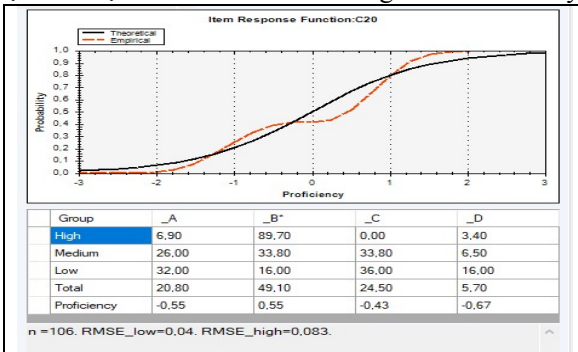
Kết quả các câu hỏi có biểu tượng hình tròn màu xanh được thể hiện qua Bảng 1

stt	Câu hỏi	Theo CTT		Pbis	Theo IRT	
		Discr (Độ phân biệt)	Pval (Độ khó)		a (Độ phân biệt)	b (Độ khó)
1	**6**	0.63	0.50	0.43	0.47	0.15
2	**10**	0.39	0.61	0.38	0.53	-0.69
3	**12**	0.56	0.75	0,44	0.55	-1.27
4	**13**	0.53	0.70	0.44	0.58	-0.96
5	**19**	0.61	0.61	0.55	0.63	-0.49
6	**20**	0.74	0.49	0.55	0.82	-0.01
7	**24**	0.55	0.57	0.42	0.52	-0.35
8	**27**	0.63	0.42	0.51	0.62	0.39
9	**29**	0.45	0.70	0.39	0.51	-1.18
10	**30**	0.46	0.60	0.41	0.50	-0.54
11	**31**	0.44	0.45	0.40	0.40	0.56
12	**34**	0.82	0.41	0.64	0.78	0.44
13	**35**	0.42	0.67	0.37	0.52	-0.91

Bảng 1. *Tham số các câu hỏi có biểu tượng hình tròn màu xanh*

Kết quả thống kê các câu hỏi trong bảng 1 đều có giá trị độ phân biệt, độ khó, hệ số tương quan trong khoảng chấp nhận được ($Discr \geq 0.2;\ 0.25 \leq Pval \leq 0.75$; $0.5 \leq a \leq 2.0$; $-3 \leq b \leq 3$ và Pbis >0), các câu hỏi có độ khó đạt yêu cầu, có độ phân biệt tốt trở lên, hệ số tương quan chặt chẽ cho nên nên tất cả các câu hỏi trong bảng 1 tốt và có khả năng phân loại thí sinh.

Quan sát đồ thị mô tả mối quan hệ tương quan giữa kết quả bài làm theo lý thuyết (đường liên tục) và kết quả bài làm theo thực tế (đường gián đoạn) của 13 câu hỏi trên ta thấy các đường biểu diễn tương quan giữa năng lực thí sinh và độ khó câu hỏi có độ dốc đều và gần với đường kỳ vọng. Trong bảng tổng hợp cho thấy các phương án trả lời (đáp án, phương án nhiễu) tương ứng với từng nhóm năng lực của thí sinh (nhóm năng lực cao, trung bình và thấp) đều đạt yêu cầu, phân biệt tốt từng nhóm năng lực thí sinh, do vậy các câu hỏi này đạt yêu cầu và có thể sử dụng ngay được. Sau đây là ví dụ minh họa của câu hỏi số 20 trong Hình 2 dưới đây:



Câu 20: Trong 8 giờ, công nhân sản xuất được 16 sản phẩm có tổng giá trị là 80 USD. Khi cường độ lao động tăng lên 1.5 lần thì giá trị của một sản phẩm làm ra trong ngày là bao nhiêu?

a.2.5 USD
b. 5 USD
c.7.5 USD
d. 10 USD

Hình 2. *Kết quả phân tích và câu hỏi số 20*

Kết quả phân tích câu hỏi số 20: Ta có Pval=0.49/ b= - 0.01: Đây là câu hỏi có độ khó trung bình. Disr=0.74/a=0.82: Câu hỏi có độ phân biệt rất tốt. Các phương án nhiễu A, C, D và đáp án B tốt phân loại tốt các nhóm năng lực thí sinh. Đồ thị: Hoàn toàn tương hợp giữa lý thuyết và thực tế ở các khoảng năng lực

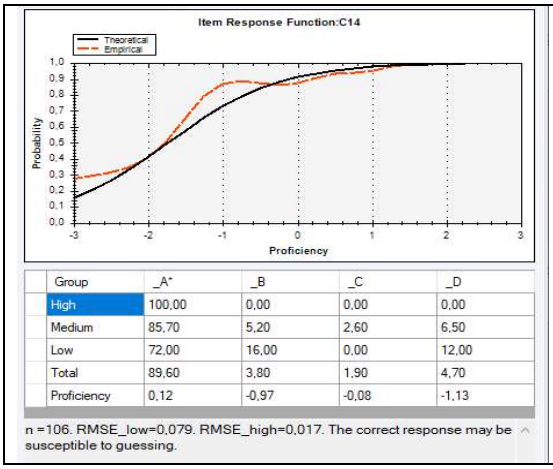
2.4.2 Kết quả phân tích câu hỏi có biểu tượng hình thoi màu vàng

Kết quả các câu hỏi có biểu tượng hình thoi màu vàng được thể hiện qua Bảng 2

stt	Câu hỏi	Theo CTT		Pbis	Theo IRT	
		Discr (Độ phân biệt)	Pval (Độ khó)		a (Độ phân biệt)	b (Độ khó)
1	**2**	0.32	0.56	0.20	**0.34**	-0.38
2	**4**	0.27	0.72	0.24	**0.25**	-1.95
3	**11**	0.35	0.69	0.29	**0.43**	-0.94
4	**14**	0.28	**0.90**	0.32	0.81	-1.69
5	**16**	**0.16**	**0.93**	0.29	0.81	-2.01
6	**17**	0.11	**0.77**	0.12	**0.41**	-2.02
7	**21**	0.27	0.75	0.23	**0.35**	-1.76
8	**23**	**0.17**	**0.86**	0.26	0.62	-1.81
9	**37**	0.25	**0.88**	0.27	0.63	-1.89
10	**38**	0.33	0.58	0.72	**0.45**	-0.51

Bảng 2. *Tham số các câu hỏi có biểu tượng hình thoi màu vàng*

Kết quả thống kê trong Bảng 2 cho thấy tất cả các câu hỏi có Pbis > 0 nhưng mỗi câu hỏi có tham số độ phân biệt *Disr* hoặc *a* hoặc tham số độ khó *Pval* nằm ngoài khoảng chấp nhận được ($Discr \geq 0.2;\ 0.25 \leq Pval \leq 0.75;\ 0.5 \leq a \leq 2.0;\ -3 \leq b \leq 3$) vì vậy các câu hỏi này chưa được tối ưu, cần điều chỉnh một chút trước khi đưa vào sử dụng. Đường biểu diễn tương quan giữa năng lực thí sinh và độ khó câu hỏi có độ dốc khá đều và khá gần với đường kỳ vọng. Chủ yếu các câu hỏi dễ và có độ khó trung bình, các phương án nhiễu lộ liễu nên thí sinh có thể đoán mò câu trả lời, phân loại chưa tốt năng lực người học. Vì vậy các câu hỏi này cũng cần điều chỉnh các phương án nhiễu và đáp án cho phù hợp. Sau đây là ví dụ minh họa của câu hỏi số 14 trong Hình 3 dưới đây:



Group	_A*	_B	_C	_D
High	100,00	0,00	0,00	0,00
Medium	85,70	5,20	2,60	6,50
Low	72,00	16,00	0,00	12,00
Total	89,60	3,80	1,90	4,70
Proficiency	0,12	-0,97	-0,08	-1,13

n =106. RMSE_low=0,079. RMSE_high=0,017. The correct response may be susceptible to guessing.

Câu 14:Lựa chọn phương án đúng nhất để trả lời cho câu hỏi: Lượng giá trị của hàng hóa được tính bằng?

a. Thời gian lao động xã hội cần thiết
b. Năng suất lao động
c. Cường độ lao động
d. Lao động giản đơn và lao động phức tạp

Hình 3. *Kết quả phân tích và câu hỏi số 14*

Kết quả phân tích câu hỏi số 14: Ta có Pval=0.90/ b= - 1.69. Câu hỏi quá dễ. Disr = 0.28/a=0.81 Câu hỏi có độ phân biệt ở mức tạm được. Cả 3 phương án nhiễu chưa đạt yêu cầu vì tỉ lệ 3 nhóm năng lực chọn rất thấp (Phương án B: 3.8%, C:1.9%, D: 4.7% thí sinh chọn), có thể các phương án nhiễu quá lộ liễu, thí sinh có thể đoán mò đáp án. Đáp án A không phân biệt năng lực thí sinh (100% TS năng lực cao, 85.7% TS năng lực TB và 72% TS năng lực thấp chọn đúng). Đồ thị: Cơ bản có sự tương hợp giữa lý thuyết và thực tế ở các khoảng năng lực. Vì thể câu hỏi này cần điều chỉnh cả 3 phương án nhiễu và đáp án để tăng độ khó và độ phân biệt cho câu hỏi.

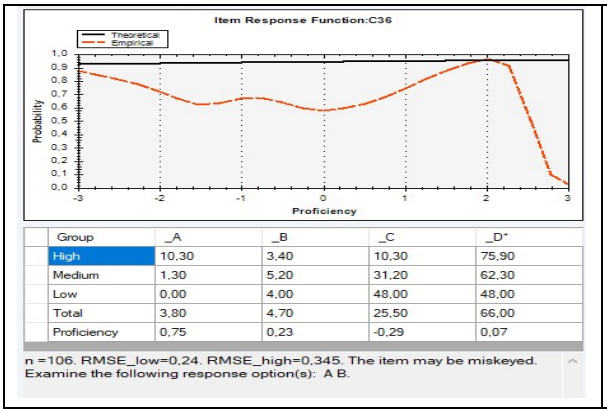
2.4.3 Kết quả phân tích câu hỏi có biểu tượng hình tam giác màu đỏ

Kết quả các câu hỏi có biểu tượng hình tam giác màu đỏ được thể hiện qua Bảng 3

stt	Câu hỏi	Theo CTT		Pbis	Theo IRT	
		Discr (Độ phân biệt)	Pval (Độ khó)		a (Độ phân biệt)	b (Độ khó)
1	**1**	**0.08**	**0.94**	0.26	0.93	-1.89
2	**3**	**0.17**	**0.91**	0.24	**0.15**	**-12.77**
3	**5**	**0.13**	**0.85**	0.14	**0.09**	**-26.51**
4	**7**	**0.13**	**0.88**	0.22	**0.12**	**-32.83**
5	**8**	**0.16**	**0.95**	0.18	**0.23**	**-9.36**
6	**9**	0.34	**0.78**	0.34	**0.12**	**-12.46**
7	**15**	0.25	**0.81**	0.26	**0.15**	**-4.91**
8	**18**	**0.07**	**0.76**	0.12	**0.06**	**-87.08**
9	**22**	0.34	**0.76**	0.30	**0.13**	**-12.20**
10	**25**	**0.08**	**0.97**	0.20	1.09	-2.19
11	**26**	**0,05**	0,41	0.08	**0.12**	**4.93**
12	**28**	0.69	0.64	0.50	**0.11**	-2.45
13	**32**	0.38	0.67	0.22	**0.09**	**-3.53**
14	**33**	0.62	0.57	0.47	**0.12**	**-9.19**
15	**36**	0.28	0.66	0.22	**0.06**	**-28.36**
16	**39**	0.65	0.57	0.57	**0.14**	**-7.21**
17	**40**	0.26	0.75	0.21	**0.04**	**-39.30**

Bảng 3. *Tham số các câu hỏi có biểu tượng hình tam giác màu đỏ*

Kết quả thống kê trong Bảng 3 cho thấy tất cả các câu hỏi có Pbis > 0 nhưng hầu hết các câu hỏi có giá trị Pbis gần tiệm cận giá trị 0. Mỗi câu hỏi đều có tham số độ phân biệt *Disr* hoặc *a* hoặc tham số độ khó *Pval hoặc b* nằm ngoài và xa khoảng chấp nhận được ($Discr \geq 0.2;\ 0.25 \leq Pval \leq 0.75;\ 0.5 \leq a \leq 2.0;\ -3 \leq b \leq 3$). Đường biểu diễn tương quan giữa năng lực thí sinh và độ khó câu hỏi có độ dốc không đều, cách xa đường kỳ vọng, cho thấy không có sự tương hợp giữa lý thuyết và thực tế ở các nhóm năng lực. Do đó 17 câu hỏi này chưa đạt yêu cầu, có độ phân biệt thấp, mối tương quan giữa câu hỏi với đề thi ít không đáng kể nên các câu hỏi này cần loại bỏ hoặc xem xét điều chỉnh về nội dung và kỹ thuật trước khi đưa vào sử dụng. Sau đây là ví dụ minh họa của câu hỏi số 36 trong Hình 4 dưới đây:



Group	_A	_B	_C	_D*
High	10,30	3,40	10,30	75,90
Medium	1,30	5,20	31,20	62,30
Low	0,00	4,00	48,00	48,00
Total	3,80	4,70	25,50	66,00
Proficiency	0,75	0,23	-0,29	0,07

n =106. RMSE_low=0,24. RMSE_high=0,345. The item may be miskeyed.
Examine the following response option(s): A B.

Câu 36: Lựa chọn nhận định đúng nhất?
A. Hàng hóa sức lao động tồn tại trong cơ thể của một người đang sống
B. Hàng hóa sức lao động mang yếu tố tinh thần và lịch sử
C. Giá trị sử dụng của hàng hóa sức lao động có khả năng tạo ra giá trị mới
D. Tất cả các phương án trên

Hình 4. *Kết quả phân tích và câu hỏi số 36*

Kết quả phân tích câu hỏi số 36: Ta có Pval=0.66/ b=-28.36, Disr = 0.28/a= 0.06 theo CTT câu hỏi có độ khó, độ phân biệt đạt yêu cầu nhưng theo IRT chưa đạt yêu cầu. Phương án nhiễu A, B chưa tốt (thí sinh năng lực trung bình chọn nhiều hơn năng lực thấp) và đáp án D chưa đạt và không phân biệt rõ các nhóm năng lực. Đồ thị: Không hoàn toàn tương hợp giữa lý thuyết và thực tế ở khoảng năng lực. Vì vậy nên loại bỏ hoặc điều chỉnh phương án nhiễu A, B, đáp án D và kiểm tra lại câu hỏi có khả năng câu hỏi bị lỗi.

Như vậy qua phân tích như trên bài viết đề xuất để lựa chọn được các câu hỏi TNKQ có chất lượng cần lựa chọn các câu hỏi chứa tham số theo cả CTT và IRT nằm trong khoảng chấp nhận được và đường cong đặc trưng câu hỏi có sự tương hợp giữa lý thuyết và thực tế ở các khoảng năng lực bao

gồm 13 câu hỏi (câu hỏi số 6,10,12,13,19,20,24,27,29,30,31,34,35), với những câu hỏi còn lại cần xem xét điều chỉnh hoặc loại bỏ trước khi đưa vào sử dụng.

3. Kết luận

Trong khuôn khổ bài viết này giới thiệu và ứng dụng phần mềm IATA vào phân tích các câu hỏi TNKQ dựa trên nền tảng CTT và IRT. Kết quả phân tích cho thấy trong đề thi có nhiều câu hỏi dễ, có tính phân loại không cao tuy nhiên xét theo tính chất đây là môn chung, áp dụng cho toàn trường thì khả thi, chấp nhận được còn nếu áp dụng cho sinh viên chuyên ngành hoặc kỳ thi có mục đích phân loại năng lực của thí sinh thì chưa phù hợp, khó phân loại được các nhóm thí sinh khá, giỏi. Bằng cách phân tích số liệu kết hợp với sử dụng biểu đồ trực quan cũng chỉ ra đề thi này có nhiều câu hỏi tương đối tốt phù hợp với năng lực thí sinh, độ phân biệt chấp nhận được tuy nhiên vẫn còn một số câu hỏi chưa đạt yêu cầu, gặp vấn đề về nội dung và kỹ thuật, độ phân biệt thấp đồng thời chẩn đoán, thăm dò nguyên nhân sai sót của các câu hỏi kém chất lượng để từ đó có biện pháp điều chỉnh, bổ sung cho phù hợp với các kỳ thi tiếp theo hoặc đưa vào ngân hàng đề thi. Việc ứng dụng các phần mềm đánh giá, phân tích đề thi là thao tác cần thiết và rất quan trọng trong quá trình biên soạn và đánh giá chất lượng đề thi, xây dựng ngân hàng câu hỏi thi để chỉ ra kịp thời các vấn đề cần bổ sung, điều chỉnh, cải tiến giúp cho người biên soạn được các câu hỏi thi chất lượng phù hợp với năng lực thí sinh, có sự phân biệt giữa các nhóm năng lực cao và thấp. Do vậy nhà trường tạo điều kiện cho giảng viên nâng cao trình độ chuyên môn nghiệp vụ về biên soạn đề thi, kiểm tra đánh giá kết quả học tập và ứng dụng các phần mềm chuyên dụng phân tích đánh giá chất lượng đề thi, ngân hàng câu hỏi thi nhằm đánh giá chính xác năng lực người học và nâng cao chất lượng đào tạo của nhà trường.

TÀI LIỆU THAM KHẢO

[1] Lâm Quang Thiệp (2010*). Đo lường trong giáo dục. Lý thuyết và Ứng dụng*. NXB Đại học Quốc gia Hà Nội.

[2] Sái Công Hồng, Lê Thái Hưng, Lê Thị Hồng Hà, Lê Đức Ngọc (2017). *Giáo trình Kiểm tra đánh giá trong giáo dục*. NXB Đại học Quốc gia Hà Nội.

[3] Cartwright, F. (2007). *IATA 3.0 Item and Test Analysis: a software tutorial and theoretical introduction*, p139, 141

[4] Wu, M & Adams, R (2007). *Applying the Rasch Model to Psycho-social Measurement: A practical Approach.* Tài liệu tập huấn Thiết kế công cụ đánh giá Do Ngân hàng thế giới phối hợp với ACER tổ chức năm 2007-2008 tại Việt Nam

[5] Baker, F, B (2001). *The basics of item response theory*. http://ericae.net/irt/baker. p168

[6] Dương Thiệu Tống (2000), *Thống kê ứng dụng trong nghiên cứu khoa học giáo dục*, NXB Đại học Quốc gia Hà Nội