TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
Tập 17, Số 10 (2020): 1804-1818
ISSN: 1859-3100
HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
Vol. 17, No. 10 (2020): 1804-1818
Website: *http://journal.hcmue.edu.vn*

Bài báo nghiên cứu

PHÂN TÍCH VÀ LỰA CHỌN CÂU HỎI TRẮC NGHIỆM KHÁCH QUAN DỰA TRÊN LÍ THUYẾT TRẮC NGHIỆM CỔ ĐIỂN VÀ LÍ THUYẾT ỨNG ĐÁP CÂU HỎI

Nguyễn Văn Cảnh[1*], Nguyễn Phước Hải[2]
[1]Trường Đại học Đồng Tháp, Việt Nam
[2]Trường Cao đẳng Sư phạm Kiên Giang, Việt Nam
[*]Tác giả liên hệ: Nguyễn Văn Cảnh – Email: nvcanh@dthu.edu.vn
Ngày nhận bài: 28-8-2020; ngày nhận bài sửa: 18-9-2020, ngày chấp nhận đăng: 19-10-2020

TÓM TẮT

Nghiên cứu này trình bày kết quả phân tích và lựa chọn 50 câu hỏi thi trắc nghiệm khách quan (TNKQ) học phần Tiếng Anh 1 của 798 sinh viên trong năm học 2019-2020 ở Trường Đại học Đồng Tháp dựa trên phần mềm IATA. Bằng cách kết hợp phương pháp phân tích dựa trên lí thuyết trắc nghiệm cổ điển và lí thuyết ứng đáp câu hỏi, những câu hỏi tốt sẽ được phát hiện và đưa vào ngân hàng đề thi dùng để kiểm tra, đánh giá kết quả học tập của người học. Những câu hỏi không đạt yêu cầu sẽ được phát hiện để điều chỉnh hoặc loại bỏ. Kết quả nghiên cứu cho thấy phần mềm IATA có nhiều ưu điểm như dễ sử dụng, tiết kiệm thời gian, cho kết quả chính xác, trực quan, rõ ràng để phân tích, đánh giá và lựa chọn các câu hỏi TNKQ. Nghiên cứu này không chỉ có thể dùng để phân tích, lựa chọn các câu hỏi TNKQ mà còn có thể nâng cao chất lượng các bài thi, để từ đó xây dựng các đề thi đáp ứng yêu cầu đổi mới trong giáo dục.

Từ khóa: câu hỏi trắc nghiệm khách quan; phần mềm IATA; lí thuyết trắc nghiệm cổ điển; lí thuyết ứng đáp câu hỏi

1. Đặt vấn đề

Để việc đánh giá kết quả học tập của người học được công bằng, khách quan, đồng thời đánh giá chính xác năng lực của người học, người dạy thường sử dụng kết hợp nhiều hình thức đánh giá khác nhau, trong đó có hình thức TNKQ. Ưu điểm nổi bật của hình thức đánh giá này là bao quát được nhiều nội dung trong một đề kiểm tra, đồng thời việc chấm điểm được thực hiện rất nhanh và có thể đảm bảo tính khách quan. Tuy nhiên, hình thức đánh giá này cũng tồn tại một vài hạn chế, đặc biệt là trong quá trình biên soạn. Trên thực tế, phần lớn các câu hỏi được sử dụng trong các đề kiểm tra đều do người dạy tự biên soạn và chưa qua giai đoạn phân tích thử nghiệm và đánh giá. Điều này đã dẫn đến tình trạng trong đề kiểm tra xuất hiện những câu hỏi kém chất lượng. Cụ thể, đối với những câu hỏi quá dễ (có độ khó thấp hơn năng lực

Cite this article as: Nguyen Van Canh, & Nguyen Phuoc Hai (2020). Analyzing and selecting multiple-choice test items based on classical test theory and item response theory. *Ho Chi Minh City University of Education Journal of Science, 17*(10), 1804-1818.

của toàn bộ thí sinh dự thi) sẽ dẫn đến mọi thí sinh dự thi đều trả lời đúng, và những câu hỏi quá khó (độ khó cao hơn năng lực của toàn bộ thí sinh) thì hầu hết thí sinh đều không làm được, một số ít thí sinh làm được có thể do yếu tố đoán mò. Điều này ảnh hưởng đến việc đánh giá kết quả học tập và năng lực của thí sinh dự thi. Do đó, để việc kiểm tra đánh giá bằng hình thức TNKQ mang lại hiệu quả, các trường cần phải triển khai xây dựng các ngân hàng đề thi, trong đó các câu hỏi cần phải được phân tích đánh giá và điều chỉnh trước khi đưa vào sử dụng. Việc phân tích, đánh giá đề thi TNKQ hiện nay thường được thực hiện dựa trên lí thuyết trắc nghiệm cổ điển và lí thuyết ứng đáp câu hỏi (hay còn gọi là lí thuyết trắc nghiệm hiện đại) thông qua các phần mềm chuyên dụng. Trong thời gian gần đây, ở Việt Nam, vấn đề này đã được một số nhà nghiên cứu quan tâm và thực hiện qua một số nghiên cứu với những phương pháp khác nhau, như: sử dụng phương pháp PROX (Nguyen, & Nguyen, 2006), sử dụng các phần mềm Quest/ConQuest (Nguyen, 2008; Bui, 2017; Nguyen, & Nguyen, 2020), sử dụng phương pháp lấy mẫu GIBB (Le et al., 2017), sử dụng phần mềm R (Doan et al., 2016), sử dụng bảng SP/GSP và phương pháp ROC thông qua phần mềm MATLAB (Nguyen, & Du, 2015; Nguyen, 2017). Mỗi phần mềm được sử dụng trong các nghiên cứu đều có ưu điểm khác nhau và cùng nhận diện những câu hỏi thực sự có chất lượng, đồng thời chỉ ra những câu hỏi chưa thực sự tốt cần phải được cải tiến trước khi đưa vào sử dụng. Trong số những phần mềm chuyên dụng có chức năng phân tích câu hỏi TNKQ hiện nay, chúng tôi nhận thấy phần mềm IATA rất hữu ích và dễ sử dụng. Điểm nổi bật của phần mềm này là chức năng phân tích câu hỏi dựa trên lí thuyết trắc nghiệm cổ điển và lí thuyết ứng đáp câu hỏi. Việc ứng dụng phần mềm này đã được thực hiện trong một số nghiên cứu (Bui, & Bui, 2018; Pham, & Bui, 2019). Tuy nhiên, các nghiên cứu trên chủ yếu sử dụng các tham số của câu hỏi theo lí thuyết trắc nghiệm cổ điển vào quá trình phân tích đánh giá và chưa đưa ra được đề xuất cụ thể để lựa chọn câu hỏi TNKQ. Nghiên cứu này tiếp tục ứng dụng phần mềm IATA vào phân tích, đánh giá đề thi TNKQ dựa trên sự kết hợp của lí thuyết trắc nghiệm cổ điển và lí thuyết ứng đáp câu hỏi. Bên cạnh đó, chúng tôi sẽ đề xuất cách lựa chọn câu hỏi dựa trên các tham số được phân tích từ phần mềm IATA. Kết quả nghiên cứu sẽ giúp người biên soạn đề thi lựa chọn được những câu hỏi thực sự có chất lượng để đưa vào ngân hàng câu hỏi, đồng thời phát hiện những câu hỏi chưa đạt yêu cầu, cần phải được xem xét lại trước khi sử dụng hoặc loại bỏ.

2. Cơ sở lí thuyết và phương pháp nghiên cứu

2.1. Giới thiệu sơ lược về lí thuyết trắc nghiệm cổ điển

Lí thuyết trắc nghiệm cổ điển (Classical Test Theory – CTT) ra đời vào khoảng cuối thế kỉ XIX và hoàn thiện vào những năm 60 của thế kỉ XX. Lí thuyết này được xây dựng dựa trên khoa học thống kê và được ứng dụng chủ yếu trong việc phân tích, đánh giá câu hỏi TNKQ. Việc đánh giá câu hỏi TNKQ theo CTT chủ yếu dựa trên các tham số độ khó, độ phân biệt và hệ số tương quan của câu hỏi với đề thi sau khi có kết quả phản hồi của thí sinh đối với các câu hỏi trong đề kiểm tra.

2.1.1. Độ khó của câu hỏi

Độ khó (*P*) của câu hỏi là tỉ lệ thí sinh trả lời đúng câu hỏi đó trên tổng số thí sinh dự thi. Theo cách định nghĩa như trên, giá trị *P* càng bé thì độ khó của câu hỏi càng cao và ngược lại. Thông thường độ khó của một câu hỏi có thể chấp nhận được khi giá trị *P* đạt giá trị từ 0,25 đến 0,75 tương ứng với số lượng thí sinh trả lời đúng đạt từ 25% đến 75%. Những câu hỏi quá dễ khi giá trị độ khó P > 0.75 (số lượng thí sinh trả lời đúng trên 75%) và những câu hỏi quá khó giá trị P < 0.25 (số lượng thí sinh trả lời đúng thấp hơn 25%). Với các câu hỏi TNKQ có *n* phương án lựa chọn, độ khó của câu hỏi ở mức trung bình là $P = \frac{1}{2}\left(1+\frac{1}{n}\right)$. Như vậy, những câu hỏi dạng Đúng/Sai có độ khó ở mức trung bình P = 0.75 (tương ứng 75% thí sinh trả lời đúng), những câu hỏi với 4 phương án lựa chọn có độ khó ở mức trung bình P = 0.65 (tương ứng 65% thí sinh trả lời đúng), những câu hỏi với 5 phương án lựa chọn có độ khó ở mức trung bình P = 0.6 (tương ứng 60% thí sinh trả lời đúng). Ngoài ra, khi chọn lựa các câu hỏi TNKQ theo độ khó, người ta thường phải loại các câu quá khó (có rất ít thí sinh trả lời đúng) hoặc quá dễ (có quá nhiều thí sinh làm đúng). Một đề trắc nghiệm tốt thường có nhiều câu hỏi có độ khó ở mức trung bình (Lam, 2011, p.60).

2.1.2. Độ phân biệt của câu hỏi

Độ phân biệt của câu hỏi TNKQ là khả năng câu hỏi đó thực hiện sự phân biệt giữa nhóm những thí sinh có năng lực cao và nhóm những thí sinh năng lực thấp trong việc trả lời câu hỏi đó. Câu hỏi có độ phân biệt tốt là khi trả lời câu hỏi đó, nhóm thí sinh có năng lực cao phải có tỉ lệ làm đúng câu hỏi cao nhóm những thí sinh có năng lực thấp.

Việc phân loại nhóm thí sinh có năng lực cao và nhóm thí sinh có năng lực thấp theo CTT là dựa trên tổng điểm thô của những thí sinh thực hiện đề thi đó. Cụ thể, nhóm thí sinh có năng lực cao bằng 27% tổng số thí sinh đạt điểm cao tính từ trên xuống; nhóm thí sinh có năng lực thấp bằng 27% tổng số thí sinh đạt điểm thấp tính từ dưới lên (Lam, 2011, p.61). Độ phân biệt (*D*) của câu hỏi được xác định theo công thức sau:

$$D = \frac{N_c - N_t}{N} \quad (1)$$

trong đó, N_c là số thí sinh thuộc nhóm có năng lực cao làm đúng câu hỏi, N_t là số thí sinh thuộc nhóm có năng lực thấp làm đúng câu hỏi, N là 27% tổng số thí sinh dự thi.

Độ phân biệt của câu hỏi theo CTT được chia thành các mức như sau: mức *rất tốt* khi $D \geq 0.4$, mức *khá tốt* khi $0.30 \leq D \leq 0.39$, mức *tạm được* khi $0.20 \leq D \leq 0.29$ và mức *kém* khi $D \leq 0.19$ (Duong, 2005, p.159). Những câu hỏi TNKQ được sử dụng trong các đề thi nên có giá trị độ phân biệt từ 0.2 trở lên (Lam, 2011, p.62).

2.1.3. Hệ số tương quan giữa điểm của câu hỏi với điểm của cả bài trắc nghiệm (hệ số tương quan Point Biserial)

Điểm số của các câu hỏi trong đề thi cần có mối tương quan với điểm số của cả bài

trắc nghiệm. Giá trị hệ số tương quan này được xác định theo công thức sau:

$$r = \frac{(\overline{x}_i - \overline{x}_c)}{\sigma}\sqrt{\frac{p_i}{1-p_i}} \quad (2)$$

trong đó: $\overline{x}_i$ là điểm trung bình cộng của những người trả lời đúng câu hỏi thứ *i* đang xem xét mối tương quan với bài trắc nghiệm; $\overline{x}_c$ là điểm trung bình của toàn bài trắc nghiệm; p_i là độ khó của câu hỏi thứ *i* đang xem xét mối tương quan với bài trắc nghiệm; σ là độ lệch chuẩn của điểm cả bài trắc nghiệm và được xác định theo công thức:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1}} \quad (3)$$

với x_i là điểm của thí sinh thứ *i* làm bài trắc nghiệm, $\overline{x}$ là điểm trung bình của toàn bài trắc nghiệm, *n* là số thí sinh làm bài trắc nghiệm.

Hệ số tương quan của câu hỏi TNKQ có giá trị từ -1 đến 1. Khi những thí sinh làm đúng câu hỏi có điểm cao (câu hỏi có nhiều thí sinh trả lời đúng) đồng thời điểm toàn bài thi của những thí sinh này cũng cao thì hệ số tương quan của các câu hỏi có giá trị gần bằng 1. Hệ số tương quan của câu hỏi có giá trị gần bằng -1 khi những thí sinh làm đúng câu hỏi có điểm cao nhưng điểm của cả đề trắc nghiệm lại thấp, và ngược lại. Hệ số tương quan của câu hỏi bằng 0 nếu điểm của câu hỏi và điểm của cả đề trắc nghiệm không có mối liên hệ chặt chẽ và ổn định với nhau (Lam, 2011, p.61). Do đó, những câu hỏi này cần phải được loại bỏ khỏi đề thi.

Mặc dù đạt được một số thành tựu và được xem là khởi đầu cho sự phát triển của khoa học đo lường trong giáo dục, nhưng CTT vẫn tồn tại một số hạn chế; trong đó, hạn chế cơ bản nhất là không tách biệt được năng lực của các thí sinh dự thi với các tham số của câu hỏi trong đề thi TNKQ, đặc trưng này chỉ có thể được giải thích bởi đặc trưng kia và ngược lại. Do đó, rất khó so sánh năng lực của thí sinh khi họ thực hiện trên các bài trắc nghiệm khác nhau (Lam, 2011, p.76). Những hạn chế này đã được khắc phục với sự ra đời của lí thuyết ứng đáp câu hỏi.

2.2. Giới thiệu sơ lược về lí thuyết ứng đáp câu hỏi

Lí thuyết ứng đáp câu hỏi là một lí thuyết của khoa học về đo lường trong giáo dục, ra đời từ nửa sau của thế kỉ XX và phát triển mạnh mẽ cho đến nay. Lí thuyết này được xây dựng dựa trên các mô hình toán học nhằm nghiên cứu sự tương tác giữa "thí sinh – câu hỏi" khi triển khai một TNKQ. Mỗi người học đứng trước một câu hỏi sẽ ứng đáp như thế nào, điều đó phụ thuộc vào năng lực tiềm ẩn của người học và các đặc trưng của câu hỏi (Lam, 2011, p.82). Lí thuyết ứng đáp câu hỏi thường gồm ba mô hình toán học phổ biến tương ứng với số lượng các tham số của câu hỏi được sử dụng trong mô hình.

Mô hình đơn giản nhất là mô hình 1 tham số hay còn gọi là mô hình Rasch, mô hình này dựa vào giả thuyết như sau:

> Nếu một thí sinh có năng lực cao hơn một thí sinh khác thì xác suất để thí sinh đó trả lời đúng một câu hỏi bất kì phải lớn hơn xác suất tương ứng của người kia; tương tự như vậy, nếu một câu hỏi khó hơn một câu hỏi khác thì xác suất để một thí sinh bất kì trả lời đúng câu hỏi đó phải nhỏ hơn xác suất để người đó trả lời đúng câu hỏi kia (Rasch, 1960, p.117).

Trong mô hình này, để xem xét mối quan hệ giữa thí sinh - câu hỏi, Rasch chọn tham số năng lực đối với mỗi người học và tham số độ khó đối với mỗi câu hỏi. Công thức toán học của mô hình này có dạng như sau:

$$P(\theta) = \frac{e^{\theta - b}}{1 + e^{\theta - b}} \tag{4}$$

trong đó: θ là năng lực của thí sinh, b là độ khó của câu hỏi và $P(\theta)$.

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. Trên lí thuyết, tham số độ khó *b* của câu hỏi có thể đạt giá trị từ $-\infty$ đến $+\infty$. Tuy nhiên, 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 dự thi; do đó, những câu hỏi TNKQ được sử dụng trong các đề thi nên có giá trị từ -3.0 đến 3.0 (Baker, 2001, p.168). Những câu hỏi có giá trị tham số độ khó nằm ngoài khoảng trên cần phải được xem xét lại trước khi đưa vào sử dụng.

Trên cơ sở mô hình 1 tham số, mô hình 2 tham số được đề xuất bằng cách mở rộng thêm tham số phân biệt *a* nhằm thể hiện khả năng phân loại năng lực của thí sinh dự thi (Birnbaum, 1968). Công thức toán học của mô hình này có dạng như sau:

$$P(\theta) = \frac{e^{a(\theta - b)}}{1 + e^{a(\theta - b)}} \tag{5}$$

Độ phân biệt của câu hỏi 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. Trên lí thuyết, tham số phân biệt *a* của câu hỏi TNKQ có thể đạt giá trị từ $-\infty$ đến $+\infty$. Tuy nhiên, những câu hỏi có tham số phân biệt quá thấp hoặc quá cao sẽ không có ý nghĩa trong việc đo lường năng lực và phân loại thí sinh; do đó, những câu hỏi được sử dụng trong các đề thi nên có giá trị tham số phân biệt *a* đạt từ 0.5 đến dưới 2.0 (Baker, 2001, p.168). Những câu hỏi có giá trị phân biệt *a* nằm ngoài khoảng giá trị trên cần được xem xét trước khi đưa vào sử dụng.

Trên thực tế, trong quá trình làm bài trắc nghiệm khách quan, một số thí sinh có thể trả lời đúng câu hỏi dựa trên sự đoán mò. Vì vậy, mô hình 3 tham số được đề xuất với việc bổ sung tham số đoán mò *c* vào mô hình 2 tham số (Birnbaum, 1968). Công thức toán học của mô hình này có dạng như sau:

$$P(\theta) = c + (1 - c)\frac{e^{a(\theta - b)}}{1 + e^{a(\theta - b)}} \tag{6}$$

Tham số đoán mò của câu hỏi có thể đạt giá trị từ 0 đến 1. Tuy nhiên, những câu hỏi có giá trị tham số đoán mò quá cao cho thấy việc trả lời đúng câu hỏi chịu ảnh hưởng bởi yếu tố may rủi, không hoàn toàn do năng lực của thí sinh dự thi.

2.3. Giới thiệu phần mềm IATA

IATA (Item and Test Analysis) là phần mềm được dùng để phân tích các câu hỏi TNKQ (Cartwright, 2007). So với các phần mềm khác có cùng chức năng, phần mềm IATA có các ưu điểm như sau:

(1) Đưa ra chỉ dẫn đề xuất lựa chọn câu hỏi TNKQ theo biểu tượng hình ảnh của câu hỏi trong kết quả phân tích. Trong đó, những câu hỏi có biểu tượng *hình tròn màu xanh* (câu hỏi không có vấn đề lớn và có thể sử dụng ngay), *hình thoi màu vàng* (câu hỏi tương đối tốt nhưng cần kiểm tra lại trước khi sử dụng) và hình *tam giác màu đỏ* (câu hỏi không nên sử dụng hoặc xem xét thật kĩ và cải tiến trước khi sử dụng).

(2) Đưa ra tỉ lệ thí sinh lựa chọn các phương án của câu hỏi TNKQ theo các nhóm năng lực của thí sinh. Điều này giúp cho việc đánh giá từng phương án trả lời trong câu hỏi được thuận lợi hơn, giúp người biên soạn dễ dàng điều chỉnh nhằm nâng cao chất lượng câu hỏi.

(3) Việc tiếp cận phần mềm IATA đơn giản hơn rất nhiều so với các phần mềm khác có cùng chức năng phân tích câu hỏi TNKQ. Người dùng dễ dàng tải phần mềm miễn phí từ địa chỉ https://polymetrika.com/Downloads/IATA và cài đặt vào máy tính để sử dụng.

(4) Đối với người dùng chưa quen sử dụng ngôn ngữ tiếng Anh có thể chọn ngôn ngữ tiếng Việt trong quá trình sử dụng (Bui, & Bui, 2018). Để sử dụng phần mềm vào việc phân tích dữ liệu, người dùng cần chuẩn bị một tệp Excel chứa dữ liệu trả lời các câu hỏi của thí sinh và tệp dữ liệu có chứa đáp án của các câu hỏi. Ngoài ra, sau khi cài đặt phần mềm IATA vào máy tính, một thư mục có chứa các tệp dữ liệu mẫu sẽ được tạo tự động trên màn hình của máy tính. Người dùng có thể dựa vào đó để tạo thành các tệp dữ liệu dùng cho việc phân tích. Chỉ với vài thao tác, phần mềm sẽ cho ra kết quả phân tích từng câu hỏi trắc nghiệm quan trong tệp dữ liệu.

(5) Có thể phân tích câu hỏi TNKQ theo lí thuyết trắc nghiệm cổ điển kết hợp với lí thuyết ứng đáp câu hỏi, giúp việc phân tích và lựa chọn câu hỏi được chính xác hơn. Trên cơ sở đó, người biên soạn đề thi có cơ sở đề điều chỉnh, cải tiến câu hỏi và xây dựng đề thi có chất lượng tốt hơn, có thể đánh giá được năng lực của người học.

3. Kết quả nghiên cứu và thảo luận

3.1. Dữ liệu nghiên cứu

Dữ liệu được sử dụng trong bài viết này là kết quả trả lời của 798 sinh viên đối với 50 câu hỏi TNKQ trong đề thi Tiếng Anh 1 được sử dụng tại Trường Đại học Đồng Tháp năm học 2019 – 2020. Dữ liệu được trình bày trong tệp Excel (định dạng dữ liệu dùng để phân tích bằng phần mềm IATA) như Bảng 1 sau đây:

Bảng 1. *Trích một phần dữ liệu*

TT	Cau 01	Cau 02	Cau 03	Cau 04	…	Cau 47	Cau 48	Cau 49	Cau 50
1	D	B	A	B	…	D	C	D	D
2	A	B	C	A	…	B	C	D	B
3	B	B	D	D	…	D	C	A	A
4	C	B	A	B	…	D	B	D	A
5	A	C	C	D	…	A	D	D	A
6	D	C	B	D	…	A	D	D	A
7	D	B	A	B	…	D	B	D	D
8	A	C	A	B	…	D	B	D	D
9	B	D	A	B	…	D	B	D	A
…	…	…	…	…	…	…	…	…	…
791	B	B	B	D	…	C	C	D	A
792	B	C	D	B	…	D	C	D	D
793	C	B	B	B	…	B	C	D	D
794	C	B	B	B	…	B	C	C	A
795	B	D	B	A	…	D	D	D	A
796	B	B	A	B	…	D	C	D	D
797	C	B	A	A	…	D	D	D	D
798	A	C	A	B	…	D	D	D	C

Ngoài ra, các câu hỏi trong dữ liệu này có đáp án (từ câu hỏi 1 đến câu hỏi 50) lần lượt là BBBDCBBDCABDDBBDCABCDDCBAACCABACDCAABCDABCDABBBCDA.

3.2. Độ tin cậy của đề thi

Trước khi sử dụng phần mềm IATA để phân tích, đánh giá các câu hỏi TNKQ trong đề thi Tiếng Anh 1 đã được sử dụng, chúng tôi đã tiến hành đánh giá độ tin cậy của dữ liệu thông qua hệ số Cronbach Alpha. Kết quả tính toán giá trị Cronbach Alpha bằng 0.807. Điều này cho thấy dữ liệu có độ tin cậy ở mức cao, phù hợp để tiến hành các phân tích, đánh giá tiếp theo.

3.3. Kết quả nghiên cứu và thảo luận

Kết quả phân tích 50 câu hỏi TNKQ trong đề thi Tiếng Anh 1 bằng phần mềm IATA được thể hiện ở Hình 1 dưới đây:

Use	O	Name	Discr	PVal	PBis	a	b	Use	O	Name	Discr	PVal	PBis	a	b
☑	●	Cau 01	0.48	0.45	0.38	0.34	-0.07	☑	◆	Cau 26	0.24	0.89	0.30	0.55	-2.35
☑	◆	Cau 02	0.33	0.68	0.28	0.34	-1.63	☑	●	Cau 27	0.58	0.49	0.45	0.54	0.16
☑	◆	Cau 03	0.29	0.36	0.28	0.39	1.07	☑	●	Cau 28	0.52	0.78	0.48	0.63	-1.26
☑	▲	Cau 04	0.02	0.23	0.06	0.08	8.29	☑	◆	Cau 29	0.25	0.38	0.21	0.28	1.22
☑	◆	Cau 05	0.22	0.37	0.21	0.38	1.03	☑	◆	Cau 30	0.24	0.32	0.26	0.35	1.37
☑	▲	Cau 06	0.48	0.36	0.40	0.08	-8.64	☑	◆	Cau 31	0.31	0.39	0.28	0.36	0.70
☑	◆	Cau 07	0.27	0.30	0.24	0.32	1.45	☑	▲	Cau 32	0.09	0.33	0.09	0.05	1.52
☑	▲	Cau 08	0.05	0.26	0.05	0.13	7.20	☑	●	Cau 33	0.59	0.54	0.46	0.69	-0.20
☑	●	Cau 09	0.42	0.31	0.36	0.47	1.24	☑	▲	Cau 34	0.43	0.54	0.37	0.08	-7.41
☑	●	Cau 10	0.42	0.54	0.35	0.46	-0.31	☑	◆	Cau 35	0.26	0.23	0.26	0.37	1.96
☑	●	Cau 11	0.51	0.56	0.39	0.52	-0.33	☑	◆	Cau 36	0.31	0.85	0.36	0.61	-1.78
☑	●	Cau 12	0.60	0.55	0.46	0.47	-0.33	☑	●	Cau 37	0.46	0.78	0.42	0.58	-1.36
☑	●	Cau 13	0.77	0.40	0.61	0.76	0.42	☑	▲	Cau 38	0.03	0.33	0.02	0.11	7.20
☑	●	Cau 14	0.55	0.60	0.44	0.49	-0.62	☑	▲	Cau 39	0.23	0.91	0.33	0.14	-7.55
☑	●	Cau 15	0.65	0.68	0.52	0.65	-0.72	☑	▲	Cau 40	0.14	0.17	0.16	0.05	22.77
☑	●	Cau 16	0.61	0.48	0.50	0.52	0.00	☐	◆	Cau 41	-0.20	0.41	-0.14	-1.00	-999.00
☑	●	Cau 17	0.57	0.39	0.45	0.45	0.55	☑	●	Cau 42	0.71	0.50	0.55	0.77	-0.09
☑	●	Cau 18	0.44	0.60	0.35	0.46	-0.63	☑	●	Cau 43	0.39	0.66	0.35	0.44	-0.88
☑	●	Cau 19	0.39	0.72	0.34	0.46	-1.47	☑	◆	Cau 44	0.21	0.81	0.19	0.37	-1.76
☑	▲	Cau 20	0.55	0.54	0.46	0.09	-5.31	☑	◆	Cau 45	0.29	0.38	0.26	0.34	0.93
☑	●	Cau 21	0.43	0.75	0.37	0.52	-1.31	☑	▲	Cau 46	0.18	0.37	0.20	0.04	-9.99
☑	●	Cau 22	0.47	0.63	0.37	0.42	-0.95	☑	●	Cau 47	0.36	0.25	0.34	0.43	1.80
☑	◆	Cau 23	0.31	0.33	0.30	0.41	1.20	☑	◆	Cau 48	0.23	0.84	0.26	0.49	-2.15
☑	◆	Cau 24	0.24	0.91	0.29	0.57	-2.51	☑	◆	Cau 49	0.16	0.93	0.24	0.55	-2.93
☑	●	Cau 25	0.38	0.42	0.31	0.37	0.51	☑	●	Cau 50	0.37	0.37	0.34	0.41	0.70

Hình 1. *Kết quả phân tích 50 câu hỏi TNKQ bằng phần mềm IATA*

Kết quả phân tích trong Hình 1 cho biết tham số của 50 câu hỏi TNKQ được sử dụng trong đề thi Tiếng Anh 1 dựa trên CTT gồm *Discr* (độ phân biệt), *PVal* (độ khó), *PBis* (hệ số tương quan) và dựa trên IRT gồm *a* (độ phân biệt), *b* (độ khó). Ngoài ra, các câu hỏi trong dữ liệu đã được chia thành ba nhóm có biểu tượng hình ảnh khác nhau. Cụ thể, nhóm *hình tròn màu xanh* gồm các câu hỏi không có vấn đề nghiêm trọng và có thể sử dụng được ngay, nhóm *hình thoi màu vàng* gồm các câu hỏi ít tối ưu hơn so với các câu hỏi có hình tròn màu xanh và cần phải được xem xét lại trước khi đưa vào sử dụng và nhóm *hình tam giác màu đỏ* gồm những câu hỏi 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 (Cartwright, 2007, p.24). Các câu hỏi theo từng nhóm được phân tích bằng phần mềm IATA thể hiện qua Bảng 2.

Bảng 2. *Các nhóm câu hỏi được phân loại từ phần mềm IATA*

TT	Nhóm câu hỏi	Số lượng	Các câu hỏi
1	Hình tròn màu xanh	23	1, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 22, 25, 27, 28, 33, 37, 42, 43, 47, 50
2	Hình thoi màu vàng	17	2, 3, 5, 7, 23, 24, 26, 29, 30, 31, 35, 36, 41, 44, 45, 48, 49
3	Hình tam giác màu đỏ	10	4, 6, 8, 20, 32, 34, 38, 39, 40, 46

Trong đề thi này, số lượng câu hỏi trong mỗi nhóm lần lượt là 23 câu hỏi (*hình tròn màu xanh*), 17 câu hỏi (*hình thoi màu vàng*) và 10 câu hỏi (*hình tam giác màu đỏ*). Điều này có nghĩa trong số 50 câu hỏi được sử dụng trong đề thi này có 23 câu hỏi đạt yêu cầu và có thể đưa vào ngân hàng câu hỏi, 17 câu hỏi cần được xem xét thêm trước khi đưa vào sử dụng, 10 câu hỏi kém cần phải được loại bỏ hoặc phải xem xét thật kĩ trước khi đưa vào sử dụng.

Để đảm bảo việc lựa chọn câu hỏi được chính xác hơn, chúng tôi tiến hành xem xét, phân tích các câu hỏi trong từng nhóm dựa trên CTT và IRT.

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

Tham số của những câu hỏi thuộc nhóm có biểu tượng *hình tròn màu xanh* dựa trên CTT được phân tích bằng phần mềm IATA thể hiện qua Bảng 3.

Bảng 3. *Tham số của các câu hỏi có hình tròn màu xanh được phân tích bằng phần mềm IATA theo CTT*

Câu hỏi	**Độ phân biệt** (*Discr*)	**Độ khó** (*PVal*)	**Hệ số tương quan** (*PBis*)	**Câu hỏi**	**Độ phân biệt** (*Discr*)	**Độ khó** (*PVal*)	**Hệ số tương quan** (*PBis*)
01	0.48	0.45	0.38	21	0.43	0.75	0.37
09	0.42	0.31	0.36	22	0.47	0.63	0.37
10	0.42	0.54	0.35	25	0.38	0.42	0.31
11	0.51	0.56	0.39	27	0.58	0.49	0.45
12	0.60	0.55	0.46	28	0.52	0.78	0.48
13	0.77	0.40	0.61	33	0.59	0.54	0.46
14	0.55	0.60	0.44	37	0.46	0.78	0.42
15	0.65	0.68	0.52	42	0.71	0.50	0.55
16	0.61	0.48	0.50	43	0.39	0.66	0.35
17	0.57	0.39	0.45	47	0.36	0.25	0.34
18	0.44	0.60	0.35	50	0.37	0.37	0.34
19	0.39	0.72	0.34				

Kết quả thống kê trong Bảng 3 cho thấy trong nhóm *hình tròn màu xanh* có 2 câu hỏi (28 và 37) có vấn đề cần phải được xem xét trước khi đưa vào sử dụng. Đây đều là những câu hỏi quá dễ do có giá độ khó *PVal* = 0.78 tương ứng với 78% thí sinh trả lời đúng. Những câu hỏi còn lại đều có giá trị độ khó, độ phân biệt, hệ số tương quan đều trong khoảng chấp nhận được $0.25 \le PVal \le 0.75$, $Discr \ge 0.2$, $PBis > 0$. Những câu hỏi trong nhóm này tiếp tục được phân tích, đánh giá dựa trên IRT. Kết quả tính toán các tham số của từng câu hỏi được thể hiện ở Bảng 4 sau đây:

Bảng 4. *Tham số của các câu hỏi có biểu tượng hình tròn màu xanh được phân tích bằng phần mềm IATA dựa trên IRT*

Câu hỏi	**Độ phân biệt** (*a*)	**Độ khó** (*b*)	**Câu hỏi**	**Đô phân biệt** (*a*)	**Độ khó** (*b*)
01	0.33	-0.07	21	0.50	-1.35
09	0.46	1.27	22	0.41	-0.98
10	0.45	-0.32	25	0.36	0.53
11	0.51	-0.34	27	0.53	0.16
12	0.46	-0.34	28	0.63	-1.26
13	0.73	0.43	33	0.68	-0.20
14	0.48	-0.64	37	0.57	-1.37
15	0.64	-0.74	42	0.77	-0.09
16	0.51	0.00	43	0.44	-0.89
17	0.44	0.56	47	0.43	1.81
18	0.45	-0.65	50	0.41	0.71
19	0.45	-1.51			

Bảng 4 cho thấy trong các câu hỏi có biểu tượng *hình tròn màu xanh* đều có tham số độ khó *b* trong khoảng chấp nhận được $-3.0 \le b \le 3.0$. Tuy nhiên, trong nhóm này có đến 13 câu hỏi (1, 9, 10, 12, 14, 17, 18, 19, 22, 25, 43, 47, 50) có tham số độ phân biệt chưa đạt yêu cầu ($a < 0.5$), những câu hỏi này cần phải được xem xét lại trước khi đưa vào ngân hàng câu hỏi.

Những câu hỏi còn lại trong nhóm này (10 câu hỏi) đều có giá trị các tham số độ khó b, độ phân biệt a trong khoảng chấp nhận được với $-3.0 \le b \le 3.0$ và $0 \le a < 2.0$. Tuy nhiên, chúng tôi đề xuất chỉ nên lựa chọn 8 câu hỏi (11, 13, 15, 16, 21, 27, 33, 42), còn lại 2 câu hỏi 28 và 37 cần được xem xét thêm trước khi sử dụng hoặc đưa vào ngân hàng câu hỏi vì có tỉ lệ sinh viên làm đúng ở mức cao với 78% (do giá trị độ khó *PVal* = 0.78).

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

Tham số của những câu hỏi có biểu tượng *hình thoi màu vàng* dựa trên CTT được phân tích bằng phần mềm IATA thể hiện qua Bảng 5.

Bảng 5. *Tham số của các câu hỏi có hình thoi màu vàng được phân tích bằng phần mềm IATA theo CTT*

Câu hỏi	**Độ phân biệt** (*Discr*)	**Độ khó** (*PVal*)	**Hệ số tương quan** (*PBis*)	**Câu hỏi**	**Độ phân biệt** (*Discr*)	**Độ khó** (*PVal*)	**Hệ số tương quan** (*PBis*)
02	0.33	0.68	0.28	31	0.31	0.39	0.28
03	0.29	0.36	0.28	35	0.26	0.23	0.26
05	0.22	0.37	0.21	36	0.31	0.85	0.36

07	0.27	0.30	0.24	41	-0.20	0.41	-0.14
23	0.31	0.33	0.30	44	0.21	0.81	0.19
24	0.24	0.91	0.29	45	0.29	0.38	0.26
26	0.24	0.89	0.30	48	0.23	0.84	0.26
29	0.25	0.38	0.21	49	0.16	0.93	0.24
30	0.24	0.32	0.26				

Kết quả thống kê trong Bảng 5 cho thấy trong nhóm *hình thoi màu vàng* có 8 câu hỏi có vấn đề cần phải được xem xét thêm trước khi đưa vào sử dụng. Cụ thể, các câu hỏi 24, 26, 36, 44, 48, 49 là các câu hỏi quá dễ do có giá trị tham số độ khó *PVal* lần lượt đạt từ 0.81 đến 0.93 tương ứng số lượng sinh viên trả lời đúng đạt từ 81% đến 93% và câu 35 là câu hỏi quá khó do giá trị *PVal* = 0.23 tương ứng với 23% sinh viên trả lời đúng. Ngoài ra, trong nhóm này chứa 2 câu hỏi có giá trị tham số độ phân biệt thấp là câu hỏi 41 với $Discr = -0.2 < 0.2$ và câu hỏi 49 với $Discr = 0.16 < 0.2$. Bên cạnh đó, câu hỏi 41 còn có hệ số tương quan kém với cả đề thi do giá trị $Pt\,Bis = -0.14 < 0$. Những câu hỏi còn lại trong nhóm này (2, 3, 5, 7, 23, 29, 30, 31, 45) đều có các tham số trong các khoảng chấp nhận được theo CTT. Chúng tôi tiếp tục phân tích những câu hỏi trong nhóm này theo IRT. Kết quả tính toán các tham số của từng câu hỏi trong nhóm được thể hiện ở Bảng 6 sau đây:

Bảng 6. *Tham số của các câu hỏi có biểu tượng hình thoi màu vàng được phân tích bằng phần mềm IATA dựa trên IRT*

Câu hỏi	**Độ phân biệt** (a)	**Độ khó** (b)	**Câu hỏi**	**Độ phân biệt** (a)	**Độ khó** (b)
02	0.33	-1.67	31	0.36	0.71
03	0.38	1.10	35	0.37	1.98
05	0.37	1.06	36	0.61	-1.79
07	0.31	1.49	41	-1.00	-999.00
23	0.39	1.24	44	0.37	-1.77
24	0.55	-2.58	45	0.34	0.94
26	0.55	-2.37	48	0.49	-2.16
29	0.27	1.23	49	0.55	-2.95
30	0.34	1.38			

Kết quả thống kê trong Bảng 6 cho thấy nhóm này có 13 câu hỏi (2, 3, 5, 7, 23, 29, 30, 31, 35, 41, 44, 45, 48) có giá trị độ phân biệt thấp với $a < 0.5$ và 4 câu hỏi (24, 26, 36 và 49) có giá trị độ phân biệt trong khoảng chấp nhận được $0.5 \le a < 2.0$. Ngoài ra, khi xét về mức độ khó, phần lớn câu hỏi trong nhóm này đều có giá trị độ khó b trong khoảng chấp nhận được $-3.0 \le b \le 3.0$, ngoại trừ câu hỏi 41 có giá trị độ khó b rất thấp. Như vậy,

nếu chỉ dựa vào kết quả phân tích theo IRT thì nhóm này có 4 câu hỏi có thể được xem xét lựa chọn gồm câu 24, 26, 36, và 49. Tuy nhiên, khi xem xét các tham số của các câu hỏi theo lí thuyết trắc nghiệm cổ điển, những câu hỏi này đều có tỉ lệ sinh viên trả lời đúng ở mức cao, lần lượt đạt 85%, 89%, 91% và 93%. Do đó, những câu hỏi này cần phải được xem xét thêm trước khi đưa vào sử dụng.

Đối với các câu hỏi được gợi ý lựa chọn dựa theo CTT như câu 2, 3, 5, 7, 23, 29, 30, 31, 45 khi được phân tích theo IRT đều cho thấy độ phân biệt a chưa đạt yêu cầu. Vì vậy, các câu hỏi này cần phải được xem xét thêm trước khi sử dụng.

Như vậy, kết quả phân tích cho thấy tất cả câu hỏi trong nhóm *hình thoi màu vàng* đều chưa thực sự tối ưu để lựa chọn và đưa vào ngân hàng câu hỏi. Muốn sử dụng những câu hỏi này, người biên soạn cần phải xem xét thêm về nội dung và kĩ thuật soạn thảo. Điều này trùng khớp với gợi ý lựa chọn câu hỏi từ phần mềm IATA.

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

Kết quả phân tích dữ liệu bằng phần mềm IATA dựa trên CTT đối với những câu hỏi có biểu tượng *hình tam giác màu đỏ* được thể hiện ở Bảng 7 sau đây:

Bảng 7. *Tham số của các câu hỏi có hình tam giác màu đỏ được phân tích bằng phần mềm IATA theo lí thuyết trắc nghiệm cổ điển*

Câu hỏi	**Độ phân biệt** (*Discr*)	**Độ khó** (*PVal*)	**Hệ số tương quan** (*PBis*)	**Câu hỏi**	**Độ phân biệt** (*Discr*)	**Độ khó** (*PVal*)	**Hệ số tương quan** (*PBis*)
04	0.02	0.23	0.06	34	0.43	0.54	0.37
06	0.48	0.36	0.40	38	0.03	0.33	0.02
08	0.05	0.26	0.05	39	0.23	0.91	0.33
20	0.55	0.54	0.46	40	0.14	0.17	0.16
32	0.09	0.33	0.09	46	0.18	0.37	0.20

Kết quả thống kê trong Bảng 7 cho thấy trong nhóm *hình tam giác màu đỏ* có 6 câu hỏi (4, 8, 32, 38, 40, 46) có vấn đề cần phải được xem xét thêm trước khi đưa vào sử dụng. Cụ thể, cả 6 câu hỏi đều có tham số độ phân biệt thấp với giá trị $Discr < 0.2$ và lần lượt đạt từ 0.02 đến 0.18, giá trị hệ số tương quan của các câu hỏi này đều dương, tuy nhiên các giá trị này đều tiệm cận giá trị 0 cho thấy các câu hỏi này đều có tương quan chưa cao với cả đề thi. Ngoài ra câu 4 và 40 là những câu hỏi quá khó với giá trị $PVal$ lần lượt đạt 0.23 và 0.17 tương ứng với 23% và 17% sinh viên trả lời đúng. Như vậy, nếu chỉ dựa vào CTT để phân tích và lựa chọn câu hỏi, nhóm này có 4 câu hỏi có thể được xem xét lựa chọn gồm câu 6, 20, 34 và câu 39. Tuy nhiên, để việc lựa chọn câu hỏi được chính xác hơn, chúng tôi tiếp tục phân tích những câu hỏi trong nhóm này theo IRT. Kết quả tính toán các tham số của câu hỏi trong nhóm theo IRT được thể hiện ở Bảng 8 sau đây:

Bảng 8. *Tham số của các câu hỏi có biểu tượng hình tam giác màu đỏ được phân tích bằng phần mềm IATA dựa trên IRT*

Câu hỏi	**Độ phân biệt** (a)	**Độ khó** (b)	**Câu hỏi**	**Độ phân biệt** (a)	**Độ khó** (b)
04	0.08	8.53	34	0.07	-7.47
06	0.07	-8.89	38	0.10	7.26
08	0.13	7.40	39	0.14	-7.60
20	0.09	-5.46	40	0.05	22.94
32	0.05	1.53	46	0.04	-10.07

Bảng 8 cho thấy các câu hỏi trong nhóm có biểu *tượng hình tam giác màu đỏ* đều có giá trị tham số phân biệt thấp với $a < 0.5$ và có giá trị từ 0.04 đến 0.14. Ngoài ra, khi xét đến mức độ khó, nhóm này có 9 câu hỏi có giá trị tham số độ khó chưa đạt yêu cầu. Trong đó, các câu hỏi 6, 20, 34, 39, 46 đều có giá trị tham số độ khó quá thấp với $b < -3.0$ và 4 câu hỏi 4, 8, 38, 40 có giá trị tham số độ khó quá cao với $b > 3.0$. Mặc dù nhóm này có 4 câu hỏi (6, 20, 34, 39) có thể xem xét lựa chọn theo CTT, tuy nhiên khi xem xét theo IRT, các tham số của những câu hỏi này đều chưa đạt yêu cầu. Như vậy, tất cả câu hỏi thuộc nhóm có biểu tượng *hình tam giác màu đỏ* trong đề thi này này đều chưa đạt yêu cầu và không nên đưa vào ngân hàng câu hỏi. Điều này trùng khớp với gợi ý không nên lựa chọn câu hỏi của phần mềm IATA.

Như vậy, để lựa chọn được những câu hỏi TNKQ thực sự có chất lượng, chúng tôi đề xuất lựa chọn những câu hỏi chứa các tham số theo CTT và theo IRT đều nằm trong các khoảng chấp nhận được. Cụ thể, khi xét theo CTT, các tham số của câu hỏi như độ khó, độ phân biệt, hệ số tương quan phải đồng thời thỏa mãn các khoảng giá trị $0.25 \le PVal \le 0.75,\ Discr \ge 0.2,\ PBis > 0$ và các tham số của câu hỏi theo IRT cũng đồng thời thỏa mãn các khoảng giá trị $-3.0 \le b \le 3.0,\ 0.5 \le a < 2.0$. Với cách lựa chọn câu hỏi TNKQ theo đề xuất như trên, số lượng câu hỏi trong đề thi này được chọn đưa vào ngân hàng câu hỏi để sử dụng vào việc đánh giá kết quả học tập và năng lực của người học là 8 câu hỏi (11, 13, 15, 16, 21, 27, 33, 42). Ngoài ra, các câu hỏi còn lại cần phải được xem xét thêm về nội dung và kĩ thuật thiết kế trước khi đưa vào sử dụng.

4. Kết luận

Việc ứng dụng phần mềm IATA vào phân tích, đánh giá câu hỏi TNKQ dựa trên lí thuyết trắc nghiệm cổ điển kết hợp với lí thuyết ứng đáp câu hỏi đã chỉ ra được những câu hỏi thực sự tốt để đưa vào ngân hàng câu hỏi và những câu hỏi chưa đạt yêu cầu cần loại bỏ hoặc phải được xem xét thêm trước khi sử dụng. Kết quả nghiên cứu và thảo luận trong bài viết này đã cho thấy việc ứng dụng phần mềm IATA vào phân tích, đánh giá và lựa chọn câu hỏi TNKQ là một phương pháp rất hữu ích, giúp cải tiến và nâng cao chất lượng đề thi, đặc biệt là ứng dụng vào việc xây dựng ngân hàng câu hỏi TNKQ. Trên cơ sở ngân

hàng câu hỏi thi đã được xây dựng, người biên soạn đề thi có thể chủ động lựa chọn những câu hỏi có giá trị các tham số độ khó, độ phân biệt phù hợp để đưa vào các đề thi, giúp đánh giá chính xác năng lực của người học, góp phần nâng cao chất lượng đào tạo của nhà trường.

❖ ***Tuyên bố về quyền lợi:*** *Tác giả xác nhận hoàn toàn không có xung đột về quyền lợi.*

TÀI LIỆU THAM KHẢO

Baker, F. B. (2001). *The basics of item response theory*. For full text: http://ericae. net/irt/baker

Birnbaum, A. L. (1968). *Some latent trait models and their use in inferring an examinee's ability*. Statistical theories of mental test scores.

Bui, N. Q. (2017). Danh gia chat luong ngan hang de thi trac nghiem khach quan mon Nhan hoc dai cuong bang mo hinh Rasch va phan mem Quest [Evaluation of the quality of multiple choice test bank for the module of Introduction to Anthropology by using the RASCH model and QUEST software]. *Science of Technology Development*, 20(X3), 42-54.

Bui, A. K., & Bui, N. P. (2018). Su dung phan mem IATA de phan tich, danh gia va nang cao chat luong cau hoi trac nghiem khach quan trong chuong trinh ham so luy thua, ham so mu, ham so logarit [Using IATA to analyze, evaluate and improve the quality of the multiple-choice questions in chapter power functions, exponential functions and logarithmic functions]. *Can Tho University Journal of Science*, 54 (9C), 81-93.

Cartwright, F. (2007). IATA 3.0 Item and Test Analysis: a software tutorial and theoretical introduction.

Doan, H. C., Le, A. V., & Pham, H. U. (2016). Ap dung mo hinh IRT 3 tham so vao do luong va phan tich do kho, do phan biet va muc do du doan cua cac cau hoi trong de thi trac nghiem khach quan [Applying 3-parameter logistic model in validating the level of difficulty, discrimination and guessing of items in a multiple choice test]. *Ho Chi Minh City University of Education Journal of Science, 7*(85), 174-184.

Duong, T. T. (2005). *Trac nghiem va do luong thanh qua hoc tap* [*Test and measure academic achievement*]. Hanoi: Social Sciences Publishing House.

Lam, Q. T. (2011). *Do luong trong giao duc li thuyet va ung dung [Measurement in Education - Theory and Application]*. Hanoi: Vietnam National University Publishing House.

Le, A. V., Pham, H. U., Doan, H. C., Le, T. H. (2017). Ap dung lay mau GIBBS vao do luong va danh gia do kho cau hoi trong mo hinh Rasch [Using Gibbs Sampler to evaluate item difficulty in Rasch model]. *Ho Chi Minh city University of Education Journal of Science, 14*(4), 119-130.

Nguyen, B. H. T. (2008). Using Quest software to analyze objective test questions [Su dung phan mem Quest de phan tich cau hoi trac nghiem khach quan]. *Journal of Science and Technology, Da Nang University*, 2, 119-126.

Nguyen, P. H. (2017). Su dung bang GSP va phuong phap ROC de phan tich va lua chon cau hoi trac nghiem khach quan [Using GSP chart and ROC method to analyze and select mutiple choice items]. *Dong Thap University Journal of Science, 24*(2), 11-17.

Nguyen, P. H., & Du, T. N. (2015). Phan tich va lua chon cau hoi trac nghiem khach quan dua tren bang S-P, phan tich quan he xam va duong cong ROC [The analysis and selection of objective test items based on S-P chart, Grey Relational Analysis, and ROC curve]. *Ho Chi Minh city University of Education Journal of Science, 6*(72), 163-173.

Nguyen, T. H. M., & Nguyen, D. T. (2006). Do luong danh gia trong de thi trac nghiem khach quan: Do kho cau hoi va nang luc cua thi sinh [Measurement Assessment in the objective test: Question difficulty and Examinees' ability]. *Vietnam National University Journal of Science*, 4, 34-47.

Nguyen, V. C., & Nguyen, Q. T. (2020). Ung dung phan mem ConQuest voi mo hinh IRT hai tham so vao viec danh gia chat luong de thi trac nghiem khach quan [Applying ConQuest software with the two-parameter IRT model to evaluate the quality of multiple-choice test]. *HNUE Journal of Science*, *65*(7), 230-242.

Pham, T. M., & Bui, D. N. (2019). Ung dung phan mem IATA de phan tich, danh gia cau hoi trac nghiem khach quan o truong Dai hoc Thu do Ha Noi [The IATA software for analyzing, evaluation of multiple-choice questions at Ha Noi Metropolitan University]. *Scientific Journal of Ha Noi Metropolitan University*, 20, 97-108.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

ANALYZING AND SELECTING MULTIPLE-CHOICE TEST ITEMS BASED ON CLASSICAL TEST THEORY AND ITEM RESPONSE THEORY

Nguyen Van Canh[1*], Nguyen Phuoc Hai[2]

[1]Dong Thap University, Vietnam

[2]Kien Giang Teachers Training College, Vietnam

[*]Corresponding author: Nguyen Van Canh – Email: nvcanh@dthu.edu.vn

Received: August 28, 2020; Revised: September 18, 2020; Accepted: October 19, 2020

ABSTRACT

This study presents the results of analysis and selection of 50 multiple-choice items of English 1 course for the final test of 798 students at Dong Thap University in the academic year 2019-2020 based on IATA software. Using a combination of analytical methodology based on Classical Test Theory (CTT) and Item Response Theory (IRT), good items will be discovered and put into a test bank to assess the student learning outcomes. Unsatisfactory items will be discovered for adjustment or removal. The study results have shown that IATA software has many advantages such as: easy to use, time-saving, accurate and clearly visual output to analyse, assess, and select multiple-choice items. This study 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 for an exam in order to meet the demands of radical and comprehensive innovation in education and training.

Keywords: multiple-choice items; IATA software; Classical Test Theory; Item Response Theory