HNUE JOURNAL OF SCIENCE DOI: 10.18173/2354-1075.2020-0093
Educational Sciences, 2020, Volume 65, Issue 7, pp. 230-242
This paper is available online at http://stdb.hnue.edu.vn

ỨNG DỤNG PHẦN MỀM CONQUEST VỚI MÔ HÌNH IRT HAI THAM SỐ VÀO VIỆC ĐÁNH GIÁ CHẤT LƯỢNG ĐỀ THI TRẮC NGHIỆM KHÁCH QUAN

Nguyễn Văn Cảnh[1] và Nguyễn Quốc Tuấn[2]
[1,2]Phòng Đảm bảo Chất lượng, Trường Đại học Đồng Tháp

Tóm tắt. Bài viết trình bày kết quả ứng dụng mô hình IRT hai tham số vào việc phân tích, đánh giá chất lượng câu hỏi trong đề thi trắc nghiệm khách quan thông qua việc sử dụng phần mềm ConQuest. Dữ liệu được sử dụng trong bài viết này là kết quả thi học phần *Tiếng Anh 1* của 798 sinh viên trong kì thi kết thúc học phần tại Trường Đại học Đồng Tháp. Kết quả nghiên cứu giúp giảng viên biên soạn đề thi phát hiện được những câu hỏi có chất lượng tốt để đưa vào ngân hàng câu hỏi thi, đồng thời phát hiện những câu hỏi kém chất lượng để điều chỉnh hoặc loại bỏ.

Từ khóa: ConQuest Software, đề thi, IRT, mô hình hai tham số.

1. Mở đầu

Đánh giá kết quả học tập là một khâu quan trọng không thể thiếu trong quá trình dạy học. Việc đánh giá kết quả học tập một cách chính xác, khách quan sẽ cung cấp cho giảng viên những thông tin hữu ích để đưa ra những quyết định kịp thời nhằm nâng cao hiệu quả hoạt động giảng dạy [1]. Hiện nay, cùng với các phương pháp kiểm tra khác, trắc nghiệm khách quan đang được sử dụng khá phổ biến trong các trường đại học. Mặc dù có nhiều ưu điểm trong đánh giá kết quả học tập, phương pháp này vẫn có một số hạn chế. Để hoạt động kiểm tra đánh giá bằng phương pháp trắc nghiệm khách quan đạt hiệu quả cao, các trường cần phải quan tâm đến việc xây dựng các ngân hàng câu hỏi thi có chất lượng tốt, có khả năng đo lường chính xác năng lực của người học. Vì vậy, việc đánh giá chất lượng đề thi sẽ giúp người biên soạn xác định được những câu hỏi có chất lượng tốt và đưa vào ngân hàng câu hỏi thi, đồng thời nhận ra được những câu hỏi kém chất lượng cần phải điều chỉnh hoặc loại bỏ.

Việc đánh giá chất lượng đề thi trắc nghiệm khách quan hiện nay thường được thực hiện dựa trên lí thuyết ứng đáp câu hỏi (Item Response Theory – IRT) bởi những ưu điểm của nó, trong đó, nổi bật nhất là việc khắc phục được những hạn chế của lí thuyết khảo thí cổ điển (Classical Test Theory – CTT) trong việc ước lượng các tham số của câu hỏi và đánh giá năng lực của thí sinh. Ở Việt Nam, việc vận dụng IRT vào đo lường, đánh giá chất lượng đề thi trắc nghiệm khách quan đã được thực hiện qua một số nghiên cứu, cụ thể như: nghiên cứu của nhóm tác giả Nguyễn Thị Hồng Minh và Nguyễn Đức Thiện (2006) với việc sử dụng phương pháp PROX [2], nghiên cứu của tác giả Lâm Quang Thiệp và các cộng sự (2007) với việc sử dụng phần mềm Vitesta [3], các nghiên cứu của các tác giả Nguyễn Bảo Hoàng Thanh (2008) [4], Nguyễn Thị Ngọc Xuân (2014) [5], Bùi Ngọc Quang (2017) [6] với việc sử dụng các phần mềm Quest/ConQuest, nghiên cứu của Bùi Anh Kiệt và cộng sự (2018) với việc sử dụng phần mềm

Ngày nhận bài: 11/5/2020. Ngày sửa bài: 15/7/2020. Ngày nhận đăng: 23/7/2020.
Tác giả liên hệ: Nguyễn Văn Cảnh. Địa chỉ e-mail: nvcanh@dthu.edu.vn

IATA [7], nghiên cứu của Đoàn Hồng Chương và các cộng sự (2016) với việc sử dụng gói ltm của phần mềm R [8], nghiên cứu của Lê Anh Vũ và các cộng sự (2017) với việc sử dụng phương pháp lấy mẫu GIBB [9].

Trong bài viết này, chúng tôi ứng dụng phần mềm ConQuest vào việc phân tích, đánh giá đề thi trắc nghiệm khách quan. Bên cạnh việc ước lượng các tham số của câu hỏi, phần mềm ConQuest còn hỗ trợ phân tích chất lượng các phương án nhiễu trong từng câu hỏi. Đây chính là ưu điểm của phần mềm ConQuest so với các phần mềm khác. Các nghiên cứu trước đây sử dụng phần mềm ConQuest vào việc đánh giá đề thi chỉ dừng lại với việc ứng dụng mô hình Rasch (mô hình IRT một tham số), việc đánh giá mức độ phân biệt của câu hỏi vẫn còn sử dụng theo CTT. Tuy nhiên, việc sử dụng độ phân biệt theo CTT có hạn chế là phụ thuộc vào năng lực thí sinh trả lời câu hỏi. Trong phạm vi bài viết này, chúng tôi sử dụng phần mềm ConQuest với mô hình IRT hai tham số vào việc đánh giá chất lượng đề thi. Trong đó, độ khó và độ phân biệt của câu hỏi đều được ước lượng theo IRT, không phụ thuộc vào năng lực của thí sinh làm bài thi. Ngoài ra, chất lượng của từng câu hỏi còn được chúng tôi đánh giá thông qua chất lượng của từng phương án nhiễu với sự hỗ trợ của phần mềm ConQuest.

2. Nội dung nghiên cứu

2.1. Mô hình IRT hai tham số

Lí thuyết Ứng đáp câu hỏi (IRT) được xây dựng dựa trên hai giả thiết: (1) Sự ứng đáp của một thí sinh đối với một câu hỏi có thể được tiên đoán bằng năng lực tiềm ẩn của thí sinh; (2) Quan hệ giữa sự ứng đáp câu hỏi của thí sinh và năng lực tiềm ẩn làm cơ sở cho sự ứng đáp đó có thể mô tả bằng một hàm đặc trưng câu hỏi đồng biến [10]. Điểm nổi bật của IRT là các tham số đặc trưng của câu hỏi độc lập với mẫu khảo sát [11].

Năm 1960, Georg Rasch - nhà toán học người Đan Mạch đã đưa ra một mô hình ứng đáp câu hỏi để mô tả mối tương tác giữa một thí sinh với một câu hỏi của đề thi trắc nghiệm. Để xem xét mối quan hệ giữa thí sinh và câu hỏi trong sự ứng đáp câu hỏi, đối với mỗi thí sinh Rasch chọn tham số năng lực, đồng thời đối với mỗi câu hỏi ông chỉ chọn một tham số liên quan là độ khó. Vì chỉ sử dụng một tham số liên quan đến câu hỏi nên mô hình này còn được gọi là mô hình ứng đáp câu hỏi một tham số. Mô hình này xuất phát từ giả thuyết như sau:

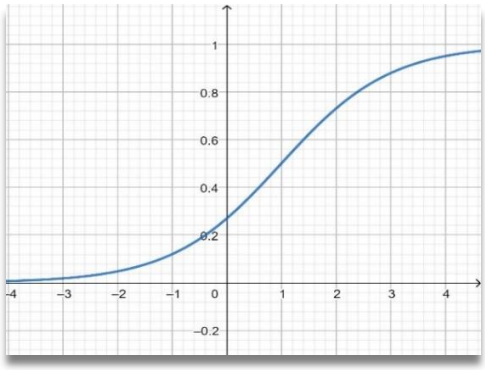
Nếu một thí sinh có năng lực cao hơn một thí sinh khác thì xác suất để thí sinh đó trả lời đúng một câu hỏi bất kì phải lớn hơn xác suất tương ứng của người kia; tương tự như vậy, nếu một câu hỏi khó hơn một câu hỏi khác thì xác suất để một thí sinh bất kì trả lời đúng câu hỏi đó phải nhỏ hơn xác suất để người đó trả lời đúng câu hỏi kia [12].

Dựa trên giả thuyết đó, Rasch đã xây dựng một mô hình toán học cho sự ứng đáp câu hỏi của mỗi thí sinh. Hàm đặc trưng của câu hỏi trong mô hình này có dạng như sau:

$$P(X_{ij} = 1 / \theta_i, b_j) = \frac{e^{\theta_i - b_j}}{1 + e^{\theta_i - b_j}} \quad (1)$$

Trong đó: θ_i là năng lực của thí sinh thứ i, b_j là độ khó của câu hỏi thứ j, X_{ij} là trả lời của thí sinh thứ i với câu hỏi thứ j. Giá trị $X_{ij} = 1$ nếu thí sinh trả lời đúng câu hỏi và $X_{ij} = 0$ khi thí sinh trả lời sai. 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.

Đường cong đặc trưng của câu hỏi trong mô hình Rasch có dạng như Hình 1.



Hình 1. Đường cong đặc trưng của câu hỏi theo mô hình Rasch

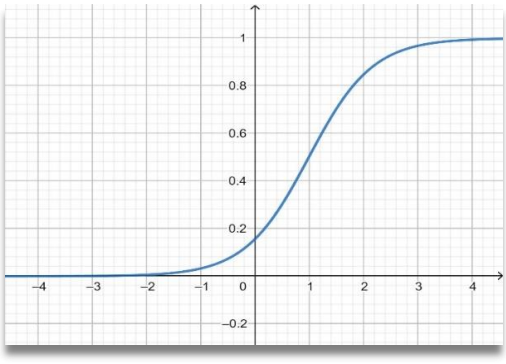
Đường cong đặc trưng của câu hỏi biểu thị xác suất trả lời đúng câu hỏi tương ứng với năng lực của thí sinh. Xác suất này sẽ tiến dần về 1 khi năng lực của thí sinh tiến đến $+\infty$. Trong mô hình Rasch, nếu $\theta_i = b_j$ thì khả năng trả lời đúng câu hỏi đó của thí sinh là 0,5. Mức năng lực này được gọi là ngưỡng của câu hỏi. Như vậy, độ khó của mỗi câu hỏi chính là ngưỡng mà với năng lực đó, xác suất trả lời đúng câu hỏi của thí sinh là 0,5. Baker (2001) cho rằng điểm nổi bật trong mô hình Rasch là nó mô tả được mối liên hệ giữa năng lực của mỗi thí sinh đối với các tham số đặc trưng của các câu hỏi thông qua sự ứng đáp của mỗi thí sinh khi trả lời các câu hỏi trong đề thi [13].

Với mỗi câu hỏi trong đề thi trắc nghiệm, ngoài tham số độ khó b_j, Birnbaum (1968) đã đề xuất mở rộng thêm một tham số nữa là độ phân biệt a_j [14]. Hàm đặc trưng của câu hỏi trong mô hình này có dạng như sau:

$$P\left(X_{ij} = 1 / \theta_i, a_j, b_j\right) = \frac{e^{a_j\left(\theta_i - b_j\right)}}{1 + e^{a_j\left(\theta_i - b_j\right)}} \quad (2)$$

Độ phân biệt của câu hỏi thể hiện khả năng phân loại thí sinh tham gia làm bài. Thông thường độ phân biệt của câu hỏi có giá trị dương. Trong trường hợp câu hỏi sai hoặc mắc lỗi thiết kế thì độ phân biệt có thể mang giá trị âm. Câu hỏi 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 lớn. Tuy nhiên, những câu hỏi có độ phân biệt quá thấp hoặc quá cao sẽ không có ý nghĩa trong việc đo lường năng lực và phân loại thí sinh.

Đường cong đặc trưng của câu hỏi theo mô hình IRT hai tham số có dạng như Hình 2.



Hình 2. Đường cong đặc trưng của câu hỏi theo mô hình IRT hai tham số

Độ dốc của đường cong đặc trưng cho biết mức độ phân biệt của câu hỏi, đường cong có độ dốc càng lớn thì câu hỏi có độ phân biệt càng cao. Khi cùng giá trị độ khó, đường cong đặc trưng câu hỏi trong mô hình IRT hai tham số có dốc lớn hơn so với đường cong đặc trưng câu hỏi trong mô hình Rasch khi giá trị độ phân biệt lớn hơn 1 và ngược lại khi giá trị độ phân biệt bé hơn 1.

2.2. Giới thiệu về dữ liệu phân tích

Bài viết này dựa trên kết quả phân tích dữ liệu đề thi trắc nghiệm khách quan học phần Tiếng Anh 1 trong kỳ thi kết thúc học phần học kì 1, năm học 2019 - 2020 tại Trường Đại học Đồng Tháp. Đề thi gồm 50 câu hỏi trắc nghiệm khách quan, mỗi câu hỏi đều có 04 phương án trả lời, trong đó có 01 phương án đúng và 03 phương án nhiễu. Số lượng thí sinh tham gia trả lời các câu hỏi trong đề thi là 798. Kết quả phản hồi của các thí sinh được lưu lại trong file dữ liệu TiengAnh1.dat (định dạng file dữ liệu bắt buộc để chạy phần phềm ConQuest).

2.3. Giới thiệu về phần mềm ConQuest và cách sử dụng với mô hình IRT 2 tham số

Phần mềm ConQuest đươc viết bởi Hội đồng nghiên cứu giáo dục Úc (Australian Council of Educational Research - ACER) nhằm ứng dụng IRT vào phân tích dữ liệu đề thi và đánh giá năng lực của thí sinh. Các phiên bản đầu tiên của phần mềm ConQuest chỉ thực hiện được việc phân tích đề thi với mô hình Rasch một tham số [15]. Phần mềm ConQuest với phiên bản 4.0 cho phép thực hiện việc phân tích đánh giá đề thi với mô hình IRT hai tham số [16].

Để sử dụng phần mềm ConQuest cần có hai file dữ liệu đầu vào gồm (1) file cấu hình có định dạng *.cqc và (2) file chứa kết quả trả lời của các thi sinh có định dạng *.dat. Việc sử dụng mô hình IRT nào phụ thuộc vào các lệnh điều khiển trong file cấu hình. Để sử dụng được mô hình IRT hai tham số cho dữ liệu của nghiên cứu này, file cấu hình có nội dung dưới đây.

Bảng 1. Nội dung file cấu hình để sử dụng phần mềm ConQuest với mô hình IRT hai tham số

```
Datafile TiengAnh1.dat;
Format responses 1-50;
set constraints=cases;
Key BBBDCBBDCABDDBBDCABCDDCBAACCABACDCAABCDABCDABBBCDA! 1;
Model item! scoresfree;
Estimate;
Show! filetype=xlsx >> TiengAnh1_show.xlsx;
Itanal! filetype=xlsx >> TiengAnh1_Itanal.xlsx;
Plot icc! filesave=yes;
Plot mcc! legend=yes, filesave=yes;
plot icc! gins=all, raw=no, overlay=yes, filesave=yes;
```

Kết quả phân tích được xuất ra từ phần mềm ConQuest bao gồm 03 phần và được tạo ra từ các lệnh Show!, Itanal!, Plot icc! và Plot mcc! trong file cấu hình. Phần 1 được thể hiện trong file TiengAnh1_show.xlsx chứa các bảng thống kê giá trị bình phương trung bình (Mean Square – MNSQ), độ khó, độ phân biệt. Phần 2 được thể hiện trong file TiengAnh1_Itanal.xlsx chứa kết quả phân tích từng câu hỏi. Phần 3 gồm các đường cong đặc trưng của câu hỏi và các đường biểu diễn xác suất phản hồi các phương án trong mỗi câu hỏi.

2.4. Đánh giá chất lượng đề thi trắc nghiệm khách quan dựa vào mô hình IRT hai tham số bằng phần mềm ConQuest

2.4.1. Sự phù hợp của câu hỏi với mô hình Rasch

Mức độ phù hợp của các câu hỏi trong đề thi với mô hình IRT được xác định dựa vào giá trị MNSQ. Câu hỏi được coi là phù hợp với mô hình nếu giá trị MNSQ của câu hỏi nằm trong khoảng

tin cậy (Confidence Interval - CI) tương ứng. Những câu hỏi có giá trị MNSQ trong cả 2 cột UNWEIGHTED FIT và WEIGHTED FIT đều nằm ngoài các khoảng CI tương ứng chứng tỏ có điều bất thường xảy ra đối với câu hỏi đó, cần phải được xem xét lại. Những bất thường xảy ra đối với câu hỏi có khả năng là đáp án bị sai, hay nội dung câu hỏi được thể hiện không rõ ràng gây ra sự hiểu nhầm cho thí sinh. Giá trị MNSQ của các câu hỏi trong đề thi này được thể hiện qua Bảng 2.

Bảng 2. Trích giá trị MNSQ và độ khó của các câu hỏi trong đề thi

VARIABLES				UNWEIGHTED FIT				WEIGHTED FIT				2PL SCALED
item		ESTIMATE	ERROR^	MNSQ	Confidence Interval		T	MNSQ	Confidence Interval		T	ESTIMATE
1	1	0.239	0.08	1.02	0.9	1.1	0.4	1	0.95	1.05	-0.1	0.322
2	2	-0.8	0.084	1.01	0.9	1.1	0.2	1	0.94	1.06	0	-1.279
3	3	0.631	0.078	1	0.9	1.1	0.1	1	0.95	1.05	-0.1	1.316
4	4	1.185	0.084	1	0.9	1.1	0	1	0.91	1.09	0	-27.978
5	5	0.554	0.075	1	0.9	1.1	0.1	1	0.96	1.04	0	1.834
6	6	0.673	0.086	1.04	0.9	1.1	0.8	0.99	0.94	1.06	-0.5	0.797
7	7	0.864	0.08	1.01	0.9	1.1	0.2	1	0.94	1.06	0	2.548
8	8	1.077	0.081	1	0.9	1.1	0	1	0.92	1.08	0	-18.322
9	9	0.876	0.085	1.02	0.9	1.1	0.4	0.99	0.93	1.07	-0.2	1.342
10	10	-0.154	0.079	0.99	0.9	1.1	-0.1	1.01	0.96	1.04	0.3	-0.218
11	11	-0.3	0.084	1	0.9	1.1	-0.1	1.01	0.95	1.05	0.2	-0.338
12	12	-0.257	0.088	1	0.9	1.1	0	1	0.94	1.06	0.2	-0.244

Kết quả thống kê trong Bảng 2 cho thấy không có câu hỏi nào có đồng thời các giá trị MNSQ trong cả hai cột UNWEIGHTED FIT và WEIGHTED FIT nằm ngoài các khoảng tin cậy CI tương ứng. Như vậy, các câu hỏi trong đề thi không có hiện tượng bất thường và đều phù hợp với mô hình IRT đang được xem xét.

2.4.2. Độ khó của câu hỏi

Baker (2001) cho rằng độ khó của câu hỏi theo IRT được chia thành 05 mức: *rất dễ* (nếu giá trị độ khó bé hơn -2,0); *dễ* (từ -2,0 đến dưới -0,5); *trung bình* (từ -0,5 đến dưới 0,5); *khó* (từ 0,5 đến dưới 2,0) và *rất khó* (từ 2,0 trở lên). Tuy nhiên, các câu hỏi trong đề thi nên có giá trị độ khó từ -3,0 đến 3,0 [13]. Những câu hỏi có giá trị độ khó quá thấp hoặc quá cao thường không có ý nghĩa trong việc đo lường năng lực của thí sinh.

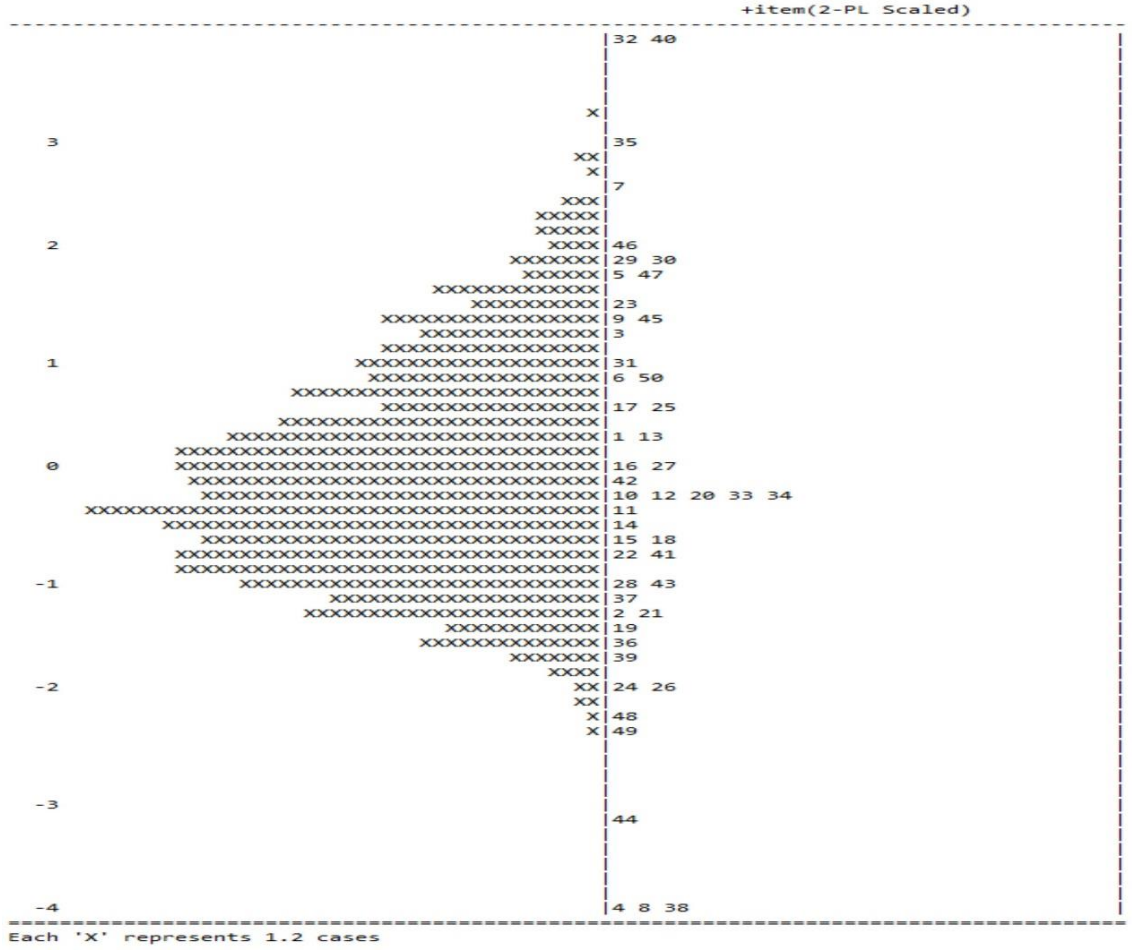
Độ khó của các câu hỏi trong đề thi này được ước lượng bằng phần mềm ConQuest theo mô hình IRT hai tham số thể hiện tại cột 2PL SCALE ESTIMATE trong Bảng 2. Theo cách phân loại trên, độ khó của các câu hỏi trong đề thi này được phân bố theo các mức độ như sau:

Bảng 3. Thống kê mức độ khó của các câu hỏi trong đề thi

Giá trị độ khó	Mức độ	Số lượng	Tỉ lệ %	Các câu hỏi
Dưới -2,0	Rất dễ	6	12,0	4, 8, 38, 44, 48, 49
Từ -2,0 đến dưới -0,5	Dễ	15	30,0	2, 14, 15, 18, 19, 21, 22, 24, 26, 28, 36, 37, 39, 41, 43
Từ -0,5 đến dưới 0,5	Trung bình	11	22,0	1, 10, 11, 12, 13, 16, 20, 27, 33, 34, 42
Từ 0,5 đến dưới 2,0	Khó	13	26,0	3, 5, 6, 9, 17, 23, 25, 29, 30, 31, 45, 47, 50
Từ 2,0 trở lên	Rất khó	5	10,0	7, 32, 35, 40, 46

Kết quả thống kê trong Bảng 3 cho thấy độ khó của câu hỏi trong đề thi tập trung nhiều ở 3 mức: mức *dễ* với 15 câu hỏi chiếm tỉ lệ 30%, mức *trung bình* với 11 câu hỏi chiếm tỉ lệ 22%, mức *khó* với 13 câu hỏi chiếm tỉ lệ 26%. Ngoài ra, đề thi cũng có 6 câu hỏi ở mức *rất dễ* chiếm tỉ lệ 12% và 5 câu hỏi ở mức *rất khó* chiếm tỉ lệ 10%. Trong đó 6 câu hỏi có độ khó quá thấp hoặc quá cao cần phải loại bỏ ra khỏi đề thi là: câu 4, 8, 32, 38, 40, 44. Độ khó trung bình của câu hỏi là -0,026, độ lệch chuẩn là 1,669. Bên cạnh đó, năng lực trung bình của thí sinh tham gia làm bài trắc nghiệm là 0,055, độ lệch chuẩn là 1,218. Như vậy, độ khó trung bình của các câu hỏi trong đề thi thấp hơn mức năng lực trung bình của thí sinh dự thi, tuy nhiên mức chệnh lệch không đáng kể.

Sự phân bố độ khó của câu hỏi và năng lực của thí sinh được ước lượng bằng phần mềm ConQuest thể hiện qua Hình 3.



Hình 3. Biểu đồ phân bố độ khó của câu hỏi và năng lực của thí sinh

Kết quả phân bố độ khó của câu hỏi và năng lực của thí sinh cho thấy độ khó của các câu hỏi trong đề thi có sự phân bố tương ứng với năng lực của thí sinh, từ những thí sinh có năng lực thấp đến những thí sinh có năng lực cao. Tuy nhiên, kết quả hiển thị trong Hình 3 đã cho thấy rất cụ thể các câu hỏi 4, 8, 38, 44, 32, 40 là những câu hỏi có độ khó quá thấp hoặc quá cao, không tương ứng với năng lực của thí sinh. Do đó, những câu hỏi trên cần phải được loại bỏ ra khỏi đề thi.

2.4.3. Độ phân biệt của câu hỏi

Baker (2001) đã chia độ phân biệt của các câu hỏi theo IRT thành 5 mức: *rất kém* (giá trị độ phân biệt bé hơn 0,35); *kém* (từ 0,35 đến dưới 0,65); *trung bình* (từ 0,65 đến dưới 1,35); *tốt* (từ 1,35 đến dưới 1,70) và *rất tốt* (từ 1,70 trở lên). Tuy nhiên, độ phân biệt của câu hỏi trong đề thi nên có giá trị từ 0,5 đến dưới 2,0 [13]. Những câu hỏi có giá trị độ phân biệt quá thấp hoặc quá cao thường không có ý nghĩa hoặc có ý nghĩa rất thấp trong việc đo lường và phân loại năng lực của thí sinh.

Độ phân biệt của các câu hỏi trong đề thi theo mô hình IRT hai tham số được ước lượng bằng phần mềm ConQuest là giá trị Score (hoặc Slope). Theo cách phân loại trên, mức độ phân biệt của các câu hỏi trong đề thi này được thể hiện qua Bảng 4.

Bảng 4. Thống kê mức độ phân biệt của câu hỏi trong đề thi

Giá trị độ phân biệt	Mức độ	Số lượng	Tỉ lệ %	Các câu hỏi
Dưới 0.35	Rất kém	9	18,0	4, 5, 7, 8, 29, 32, 38, 41, 46
Từ 0.35 đến dưới 0.65	Kém	10	20,0	2, 3, 23, 25, 30, 31, 35, 40, 44, 45
Từ 0.65 đến dưới 1.35	Trung bình	21	42,0	1, 6, 9, 10, 11, 12, 14, 16, 17, 18, 9, 20, 21, 22, 27, 33, 34, 43, 47, 48, 50
Từ 1.35 đến dưới 1.7	Tốt	5	10,0	26, 36, 37, 42, 49
Từ 1.7 trở lên	Rất tốt	5	10,0	13, 15, 24, 28, 39

Kết quả thống kê trong Bảng 4 cho thấy độ phân biệt của các câu hỏi chủ yếu tập trung ở các mức như: mức trung bình 21 câu (42%), mức kém 10 câu (20%), mức rất kém 9 câu (18%). Còn lại là các câu hỏi có mức phân biệt tốt với 5 câu hỏi (10%), và mức rất tốt với 5 câu hỏi (10%). Điều này cho thấy, đây là một đề thi có mức độ phân biệt chưa cao. Ngoài ra, trong đề thi có một số câu hỏi có độ phân biệt quá thấp (dưới 0,5) như: câu 3, 4, 5, 7, 8, 29, 30, 31, 32, 35, 38, 40, 41, 44, 45, 46. Đây là những câu hỏi cần phải loại bỏ ra khỏi đề thi.

2.4.4. Đánh giá chất lượng của từng câu hỏi trong đề thi

Chất lượng của từng câu hỏi trong đề thi được thể hiện qua các yếu tố độ khó, độ phân biệt và các phương án nhiễu. Các phương án nhiễu của một câu hỏi được gọi là có chất lượng khi xác suất lựa chọn các phương án đó của thí sinh giảm dần khi năng lực của thí sinh tăng dần, đồng thời xác suất này sẽ dần về 0 khi năng lực của thí sinh ở mức rất cao.

Sự khác biệt của phần mềm ConQuest so với các phần mềm khác khi phân tích câu hỏi theo IRT đó là việc hiển thị đường biểu diễn xác suất phản hồi các phương án nhiễu của thí sinh. Thông qua các đường biểu diễn này, người biên soạn đề thi sẽ phát hiện được những phương án nhiễu có vấn đề để chỉnh sửa hoặc thay thế.

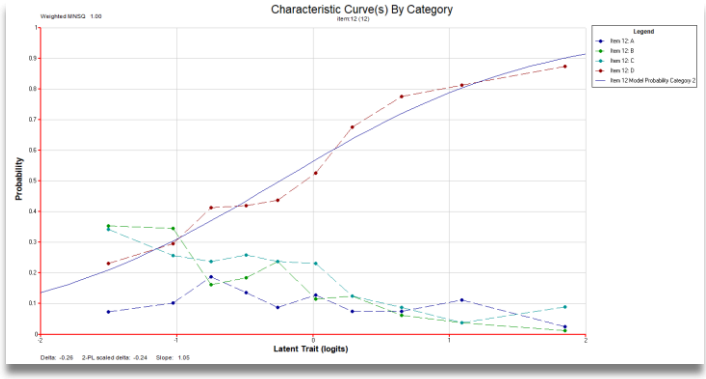
Trong phần tiếp theo của bài viết này, chúng tôi sẽ đánh giá một số câu hỏi được phân tích bằng phần mềm ConQuest theo mô hình IRT hai tham số, cụ thể như trong Bảng 5.

Giá trị 2-PL Scale delta cho biết độ khó, giá trị Score hoặc Slope cho biết độ phân biệt của câu hỏi. Giá trị Pt Bis cho biết mối tương quan giữa số lượng thí sinh lựa chọn từng phương án trả lời với tổng điểm bài thi. Giá trị Pt Bis > 0 cho biết phương án đó có số lượng thí sinh năng lực cao chọn nhiều hơn các thí sinh năng lực thấp. Điều này hợp lí đối với phương án đúng và ngược lại đối với các phương án nhiễu. Giá trị Sig < 0,05 cho biết phép kiểm định t đối với hệ số tương quan Pt Bis có ý nghĩa thống kê với độ tin cậy 95%. Những phương án nhiễu kém chất lượng khi giá trị Sig ≥ 0,05.

Bảng 5. Thông tin phân tích câu hỏi 12

item:12 (12)								
Cases for this item 798			Item-Rest Cor. 0.40				Item-Total Cor. 0.46	
Item Delta: -0.26			2-PL scaled delta: -0.24		Slope: 1.05		Weighted MNSQ 1.00	
Label	**Score**	**Count**	**% of tot**	**Pt Bis**	**t**	**sig**	**PV1Avg:1**	**PV1 SD:1**
A	0	80	10,03	-0,11	-3	0,003	-0,199	0,787
B	0	131	16,42	-0,27	-8,01	0,000	-0,658	0,762
C	0	152	19,05	-0,17	-4,78	0,000	-0,442	0,817
D	1,05	435	54,51	0,4	12,27	0,000	0,358	0,963

Kết quả thống kê trong Bảng 5 cho thấy, câu hỏi 12 có giá trị độ khó là -0,24 (mức trung bình); giá trị độ phân biệt bằng 1,05 (mức trung bình). Ngoài ra, các phương án nhiễu đều có giá trị Pt Bis < 0 và giá trị Sig < 0,05. Như vậy, các phương án nhiễu trong câu hỏi trên đều có chất lượng tốt. Xác suất lựa chọn các phương án của câu hỏi này được thể hiện qua các đường biểu diễn trong Hình 4.



Hình 4. Đường cong đặc trưng của câu hỏi 12

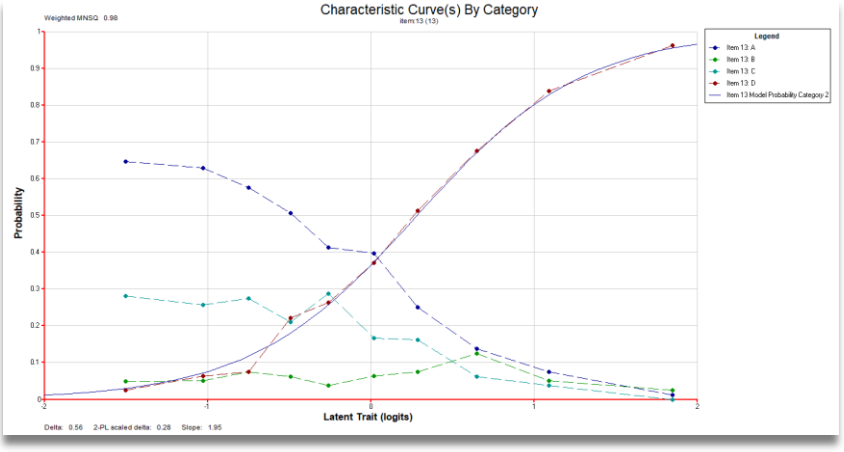
Đường biểu diễn các phương án trả lời cho thấy, khi năng lực của thí sinh càng cao thì xác suất lựa chọn các phương án nhiễu A, B, C càng thấp, đây là điều hợp lí với phương án nhiễu có chất lượng tốt. Như vậy, mặc dù đây là câu hỏi có độ khó và độ phân biệt ở mức trung bình nhưng đây là một câu hỏi có chất lượng tốt.

Bảng 6. Thông tin phân tích câu hỏi 13

item:13 (13)								
Cases for this item		798	Item-Rest Cor. 0.56				Item-Total Cor. 0.61	
Item Delta: 0.56			2-PL scaled delta: 0.28		Slope: 1.95		Weighted MNSQ 0.98	
Label	**Score**	**Count**	**% of tot**	**Pt Bis**	**t**	**sig**	**PV1Avg:1**	**PV1 SD:1**
A	0	291	36,47	-0,42	-12,88	0,000	-0,604	0,697
B	0	49	6,14	0,05	1,54	0,124	-0,023	0,801
C	0	139	17,42	-0,23	-6,62	0,000	-0,528	0,613
D	1,95	319	39,97	0,56	18,98	0,000	0,742	0,847

Kết quả thống kê trong cho thấy câu hỏi 13 có giá trị độ khó đạt 0,28 (mức trung bình); giá trị độ phân biệt đạt 1,95 (mức rất tốt). Ngoài ra, các phương án nhiễu A, C đều có giá trị Pt Bis < 0 và giá trị Sig < 0,05, phương án nhiễu B có giá trị Pt Bis = 0,05 > 0 và giá trị Sig = 0,125 > 0,05. Như vậy, trong các phương án nhiễu của câu hỏi 13, phương án B là phương án kém chất lượng. Người biên soạn đề thi cần phải quan tâm đến câu hỏi này và điều chỉnh phương án B nhằm nâng cao chất lượng câu hỏi.

Xác suất lựa chọn các phương án của câu hỏi này được thể hiện qua các đường biểu diễn trong Hình 5.



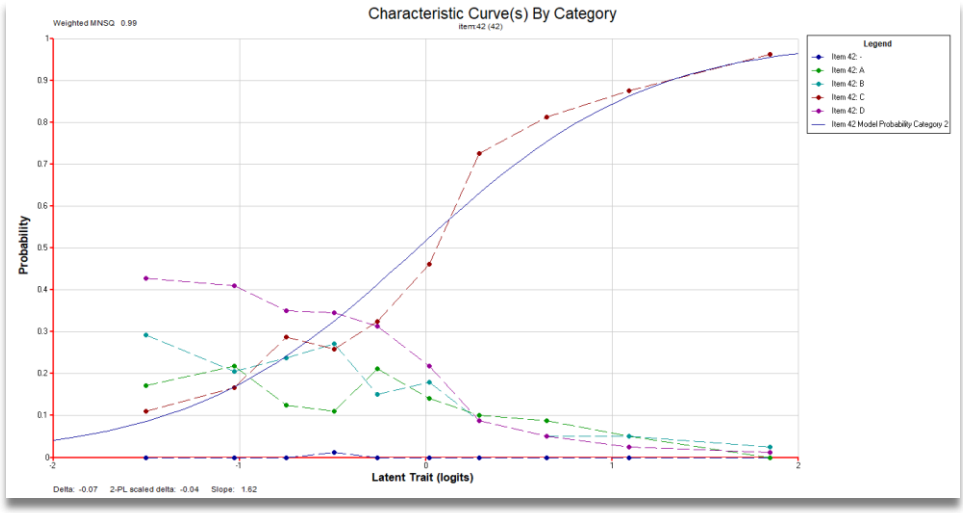
Hình 5. Đường cong đặc trưng của câu hỏi 13

Đường biểu diễn của các phương án trả lời trong câu hỏi trên cho thấy xác suất lựa chọn các phương án nhiễu A, C giảm dần khi năng lực của thí sinh tăng dần. Tuy nhiên, đối với phương án nhiễu B thì ngược lại, có những thí sinh có năng lực cao nhưng xác suất lựa chọn phương án này cũng cao hơn những thí sinh có năng lực thấp. Điều này chứng tỏ phương án B là phương án nhiễu kém chất lượng.

Bảng 7. Thông tin phân tích câu hỏi 42

item:42 (42)								
Cases for this item 798				Item-Rest Cor. 0.49			Item-Total Cor. 0.55	
Item Delta: -0.07		2-PL scaled delta: -0.04		Slope: 1.62			Weighted MNSQ 0.99	
Label	**Score**	**Count**	**% of tot**	**Pt Bis**	**t**	**sig**	**PV1Avg:1**	**PV1 SD:1**
-	0	1	0,13	-0,01	-0,18	0,856	-0,437	0,000
A	0	97	12,16	-0,12	-3,37	0,001	-0,414	0,697
B	0	124	15,54	-0,2	-5,74	0,000	-0,584	0,755
C	1,62	397	49,75	0,49	16,02	0,000	0,546	0,922
D	0	179	22,43	-0,33	-9,7	0,000	-0,656	0,604

Kết quả thống kê trong Bảng 7 cho thấy câu hỏi 42 có giá trị độ khó bằng -0,04 (mức trung bình); giá trị độ phân biệt bằng 1,62 (mức tốt); các phương án nhiễu A, B, D đều có chất lượng tốt (do giá trị Pt Bis < 0 và giá trị Sig < 0,05). Xác suất lựa chọn các phương án của câu hỏi này được thể hiện qua các đường biểu diễn trong Hình 6.



Hình 6. Đường cong đặc trưng câu hỏi 42

Đường biểu diễn các phương án trả lời cho thấy xác suất lựa chọn các phương án nhiễu A, B, D giảm dần khi năng lực của thí sinh tăng dần. Điều này cho thấy các phương án nhiễu trong câu hỏi này đều có chất lượng tốt. Như vậy, đây là một câu hỏi có chất lượng tốt.

Việc phân tích đề thi trắc nghiệm khách quan bằng phần mềm ConQuest đã cho thấy trong đề thi này có một số câu hỏi chất lượng tốt, có thể đưa vào ngân hàng câu hỏi thi dùng để đánh giá kết quả học tập của sinh viên. Những câu hỏi này được thể hiện trong Bảng 8.

Bảng 8. Thống kê các câu hỏi có chất lượng tốt

Câu hỏi	Mức độ khó	Mức độ phân biệt	Phương án nhiễu	Câu hỏi	Mức độ khó	Mức độ phân biệt	Phương án nhiễu
1	Trung bình	Trung bình	Tốt	24	Dễ	Rất tốt	Tốt
6	Khó	Trung bình	Tốt	26	Dễ	Tốt	Tốt
10	Trung bình	Trung bình	Tốt	28	Dễ	Rất tốt	Tốt
12	Trung bình	Trung bình	Tốt	33	Trung bình	Trung bình	Tốt
14	Dễ	Trung bình	Tốt	34	Trung bình	Trung bình	Tốt
15	Dễ	Rất tốt	Tốt	36	Dễ	Tốt	Tốt
16	Trung bình	Trung bình	Tốt	37	Dễ	Tốt	Tốt
17	Khó	Trung bình	Tốt	39	Dễ	Rất tốt	Tốt
18	Dễ	Trung bình	Tốt	42	Trung bình	Tốt	Tốt
19	Dễ	Trung bình	Tốt	43	Dễ	Trung bình	Tốt
20	Trung bình	Trung bình	Tốt	47	Khó	Trung bình	Tốt
21	Dễ	Trung bình	Tốt	49	Rất dễ	Tốt	Tốt
22	Dễ	Trung bình	Tốt	50	Khó	Trung bình	Tốt

Bên cạnh đó, một số câu hỏi có độ khó và độ phân biệt đáp ứng yêu cầu nhưng cần chỉnh sửa lại các phương án nhiễu để nâng cao chất lượng câu hỏi và có thể đưa vào ngân hàng câu hỏi thi. Những câu hỏi này được thể hiện trong Bảng 9.

Bảng 9. Thống kê các câu hỏi cần phải điều chỉnh phương án nhiễu

Câu hỏi	Mức độ khó	Mức độ phân biệt	Phương án nhiễu cần điều chỉnh
2	Dễ	Kém	A, D
9	Khó	Trung bình	B
11	Trung bình	Trung bình	D
13	Trung bình	Rất tốt	B
23	Khó	Kém	A
25	Khó	Kém	B
27	Trung bình	Trung bình	A
48	Rất dễ	Trung bình	B

Ngoài ra, trong đề thi này có những câu hỏi không đạt yêu cầu về độ khó và độ phân biệt. Đó là những câu hỏi có độ khó, độ phân biệt quá thấp hoặc quá cao không có ý nghĩa trong việc đo lường năng lực của người học (giá trị độ khó nằm ngoài đoạn [-3; 3], độ phân biệt nằm ngoài khoảng [0.5; 2)). Những câu hỏi này được thể hiện qua Bảng 10.

Bảng 10. Thống kê các câu hỏi kém chất lượng

Câu hỏi	Mức độ khó	Mức độ phân biệt	Câu hỏi	Mức độ khó	Mức độ phân biệt
3	Khó	Kém	32	Rất khó	Rất kém
4	Rất dễ	Rất kém	35	Rất khó	Kém
5	Khó	Rất kém	38	Rất dễ	Rất kém
7	Rất khó	Rất kém	40	Rất khó	Kém
8	Rất dễ	Rất kém	41	Dễ	Rất kém
29	Khó	Rất kém	44	Rất dễ	Kém
30	Khó	Kém	45	Khó	Kém
31	Khó	Kém	46	Rất khó	Rất kém

Kết quả thống kê trong Bảng 10 cho thấy những câu hỏi kém chất lượng trong đề thi này là những câu hỏi có độ khó ở mức *Rất khó/Khó* và độ phân biệt ở mức *Rất kém/Kém* như: câu 3, 5, 7, 29, 30, 31; hoặc những câu hỏi có độ khó ở mức *Rất dễ/Dễ* và độ phân biệt ở mức *Rất kém/Kém* như: câu 4, 8, 38, 41, 44. Đây là những câu hỏi kém chất lượng cần phải được loại bỏ ra khỏi đề thi.

3. Kết luận

Thông qua việc ứng dụng phần mềm ConQuest với mô hình IRT hai tham số, bài viết đã trình bày cách phân tích, đánh giá chất lượng của một đề thi trắc nghiệm khách quan dựa trên độ khó, độ phân biệt và chất lượng các phương án nhiễu của từng câu hỏi. Kết quả nghiên cứu trong bài viết đã chỉ ra những câu hỏi có chất lượng tốt, có thể để đưa vào ngân hàng câu hỏi thi dùng để đánh giá kết quả học tập của sinh viên. Bên cạnh đó, những câu hỏi có độ khó, độ phân biệt không đáp ứng yêu cầu đã được khuyến nghị loại bỏ ra khỏi đề thi. Đồng thời, những câu hỏi đáp ứng yêu cầu về độ khó và độ phân biệt nhưng có các phương án nhiễu chưa tốt, chưa có khả năng phân loại năng lực của thí sinh cũng đã được khuyến nghị điều chỉnh hoặc thay thế.

TÀI LIỆU THAM KHẢO

[1] Dương Thị Thúy Hà, 2017. Đánh giá kết quả học tập của người học theo định hướng hình thành năng lực và định hướng vận dụng trong thực tiễn giáo dục đại học. *Tạp chí khoa học Trường Đại học Sư phạm Hà Nội*, Tập 62, Số 1A, tr. 171-180.

[2] Nguyễn Thị Hồng Minh, Nguyễn Đức Thiện, 2006. Đo lường đánh giá trong thi trắc nghiệm khách quan: Độ khó câu hỏi và năng lực của thí sinh. *Tạp chí khoa học Trường Đại học Quốc gia Hà Nội*, Số 4, tr. 34-47.

[3] Lâm Quang Thiệp, Lâm Ngọc Minh, Lê Mạnh Tấn, Vũ Đình Bổng, 2007. Phần mềm Vitesta và việc phân tích số liệu trắc nghiệm. *Tạp chí Giáo dục*, Số 176.

[4] Nguyễn Bảo Hoàng Thanh, 2008. Sử dụng phần mềm Quest để phân tích câu hỏi trắc nghiệm khách quan. *Tạp chí Khoa học và Công nghệ, Trường Đại học Đà Nẵng*, Số 2, tr. 119-126.

[5] Nguyễn Thị Ngọc Xuân, 2014. Sử dụng phần mềm Quest/ConQuest để phân tích câu hỏi trắc nghiệm khách quan. *Tạp chí Khoa học, Trường Đại học Trà Vinh*, Số 12, tr. 24-27.

[6] Bùi Ngọc Quang, 2017. Đánh giá chất lượng ngân hàng đề thi trắc nghiệm khách quan môn Nhân học đại cương bằng mô hình Rasch và phần mềm Quest. *Tạp chí Phát triển Khoa học và Công nghệ*, Tập 20, Số X3, tr. 42-54.

[7] Bùi Anh Kiệt và Bùi Nguyên Phương, 2018. Sử dụng phần mềm IATA để phân tích, đánh giá và nâng cao chất lượng câu hỏi trắc nghiệm khách quan trong chương trình hàm số lũy thừa, hàm số mũ, hàm số logarit. *Tạp chí khoa học Trường Đại học Cần Thơ*, Tập 54, Số 9C, tr. 81-93.

[8] Đoàn Hồng Chương, Lê Anh Vũ, Phạm Hoàng Uyên, 2016. Áp dụng mô hình IRT 3 tham số vào đo lường và phân tích độ khó, độ phân biệt và mức độ dự đoán của các câu hỏi trong đề thi trắc nghiệm khách quan. *Tạp chí khoa học Trường Đại học Sư phạm Thành phố Hồ Chí Minh*, Tập 85, Số 7, tr. 174-184.

[9] Lê Anh Vũ, Phạm Hoàng Uyên, Đoàn Hồng Chương, Lê Thanh Hoa, 2017. Áp dụng lấy mẫu GIBBS vào đo lường và đánh giá độ khó câu hỏi trong mô hình Rasch. *Tạp chí khoa học Trường Đại học Sư phạm Thành phố Hồ Chí Minh*, Tập 14, Số 4, tr. 119-130.

[10] Lâm Quang Thiệp, 2010. *Đo lường trong giáo dục, lí thuyết và ứng dụng*. Nhà xuất bản Đại học Quốc gia Hà Nội.

[11] Hambleton, R. K., & Swaminathan, H., 2013. *Item response theory: Principles and applications*. Springer Science & Business Media.

[12] Rasch, G., 1980. *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

[13] Baker, F. B., 2001. *The basic of item response theory*. College Park, MD: University of Maryland, ERIC Clearinghouse on Assessment and Evaluation.

[14] Birnbaum, A. L., 1968. *Some latent trait models and their use in inferring an examinee's ability*. Statistical Theories of Mental Test Scores.

[15] Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A., 2007. *ACER conquest version 2.0*. Camberwell, Victoria, Australia: ACER Press, Australian Council for Educational Research.

[16] Adams, R. J. & Macaskill, G., 2012, *Score Estimation and Generalised Partial Credit Models*. Note 6, ACER ConQuest, Notes and Tutorials.

ABSTRACT

Applying ConQuest software with the two-parameter IRT model to evaluate the quality of multiple-choice test

Nguyen Van Canh and Nguyen Quoc Tuan

Department of Education Quality Assurance, Dong Thap University

The paper presents the results of applying the two-parameter IRT model in analyzing and evaluating the quality of items in a multiple-choice test through the use of ConQuest software. The data used in this paper is the survey is from English 1 module of 798 students in the final exam at Dong Thap University. The research results help exam-preparation-teachers find good items to fulfil the exam-bank and identify low-quality items to adjust or remove.

Keywords: ConQuest Software, test, IRT, two-parameter model.