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

ĐOÀN HỒNG CHƯƠNG[*], LÊ ANH VŨ[**], PHẠM HOÀNG UYÊN[***]

TÓM TẮT

*Trong bài viết này, chúng tôi sử dụng mô hình IRT 3 tham số để đo lường độ khó, độ phân biệt của các câu hỏi trong đề thi trắc nghiệm khách quan nhiều lựa chọn, đồng thời khảo sát sự ảnh hưởng của mức độ dự đoán của thí sinh khi trả lời câu hỏi đối với việc đo lường và đánh giá năng lực của thí sinh. Dữ liệu trong bài viết được thu thập từ một mẫu ngẫu nhiên các bài thi cuối kì môn Toán Cao cấp của sinh viên Khóa 14 Trường Đại học Kinh tế - Luật, ĐHQG TP Hồ Chí Minh. Việc xử lí dữ liệu được thực hiện bằng gói lệnh "*ltm*" của phần mềm R. Kết quả của bài viết giúp giáo viên đánh giá đúng chất lượng của đề thi và năng lực của thí sinh.*

Từ khóa: lí thuyết ứng đáp câu hỏi, mô hình IRT 3 tham số, trắc nghiệm khách quan nhiều lựa chọn, phần mềm R.

ABSTRACT

Applying 3-parameter logistic model in validating the level of difficulty, discrimination and guessing of items in a multiple choice test

In this study, we use 3-parameter logistic model to validate the level of difficulty and discrimination of items in a multiple choice test; as well as examine the effect of test takers' guessing in answering questions for assessing test takers' competence. Data was gathered from a random sample of the 2014 Intake students taking the Advanced Mathematics Final Test of University of Economics and Law, Vietnam National University, Ho Chi Minh City. "Ltm" package of the freeware R was used to analyze the data. The findings of this study, therefore, suggest the way to assess the test's quality and examinees' competence.

Keywords: Item response theory, 3-parameter logistic model, multiple choice test, R software.

1. Mở đầu

1.1. Xuất xứ vấn đề nghiên cứu

Lí thuyết trắc nghiệm cổ điển (Classical Test Theory – CTT) ra đời từ khoảng cuối thế kỉ XIX và hoàn thiện vào những năm 60 của thế kỉ XX, đã có nhiều đóng góp quan trọng cho hoạt động đo lường và đánh giá trong giáo dục. Tuy nhiên, phương pháp này cũng bộc lộ một số hạn chế: Trước tiên là sự phụ thuộc của các tham số (độ khó, độ phân biệt) của các câu hỏi vào mẫu thí sinh tham gia kiểm tra; tiếp theo là ảnh

[*] ThS, Trường Đại học Kinh tế - Luật, ĐHQG TPHCM; *Email: chuongdh@uel.edu.vn*

[**] PGS TS, Trường Đại học Kinh tế - Luật, ĐHQG TPHCM

[***] TS, Trường Đại học Kinh tế - Luật, ĐHQG TPHCM

hưởng của các câu hỏi đến việc đo lường và đánh giá năng lực tiềm tàng (*latent trait*) của thí sinh (từ đây về sau, năng lực tiềm tàng được viết gọn là năng lực). Chẳng hạn, cùng một đề thi, khi được tiến hành với nhóm thí sinh giỏi, thì đề thi này thường được đánh giá là đề thi dễ; trong khi đối với nhóm thí sinh kém, đề thi này có khả năng được đánh giá là đề thi khó. Tương tự như vậy, cùng một thí sinh, khi làm đề thi dễ thì năng lực của thí sinh đó được đánh giá cao hơn so với khi làm đề thi khó.

Để khắc phục những nhược điểm này, mô hình lí thuyết ứng đáp câu hỏi (Item Response Theory – IRT) đã được nghiên cứu và áp dụng vào đo lường và đánh giá các câu hỏi trong đề thi. Mô hình IRT dựa trên giả thiết cơ bản sau: "nếu một người có năng lực cao hơn người khác thì xác suất để người 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 người bất kì trả lời đúng câu hỏi đó phải nhỏ hơn xác suất để người đó trả lời đúng câu hỏi kia" [8]. Điểm nổi bật của mô hình này là mô tả được mối liên hệ giữa năng lực của mỗi thí sinh với các tham số của các câu hỏi thông qua sự ứng đáp của mỗi thí sinh đối với mỗi câu hỏi trong đề thi [6,11]. Một điểm đặc biệt nữa là mô hình IRT tách biệt được các tham số của các câu hỏi với mẫu thí sinh tham gia kiểm tra, cũng như năng lực tiềm tàng của mỗi thí sinh với đề thi [6,11]. Do đó các giáo viên cũng như các nhà quản lí giáo dục có thể áp dụng mô hình IRT để thiết kế các đề thi trắc nghiệm tiêu chuẩn có mức độ tương đương cao và đo chính xác năng lực của thí sinh.

1.2. Tổng quan các nghiên cứu ở Việt Nam trước đây

Ở Việt Nam, mô hình IRT đã và đang được nhiều tác giả quan tâm và nghiên cứu. Ví dụ như Dương Thiệu Tống [4], Lâm Quang Thiệp [3], Nguyễn Bảo Hoàng Thanh [2], Nguyễn Thị Ngọc Xuân [5], Nguyễn Thị Hồng Minh [1]... Tuy nhiên, việc đo lường, phân tích và đánh giá của các tác giả ở trên chỉ dừng lại với mô hình Rasch (là một dạng mô hình IRT một tham số, hoặc mô hình IRT hai tham số). Thực tế trong đề thi trắc nghiệm khách quan nhiều lựa chọn cho thấy, khi gặp một câu hỏi có độ khó cao hơn năng lực bản thân, các thí sinh có khuynh hướng dự đoán câu trả lời (theo cách chọn ngẫu nhiên một phương án hoặc theo cách loại suy dựa trên kinh nghiệm bản thân). Do đó, Birnbaum đề xuất thêm tham số dự đoán vào mô hình để đo lường mức độ dự đoán của thí sinh trong mỗi câu hỏi. [7]

1.3. Mục đích nghiên cứu

Mục đích của bài viết là áp dụng mô hình IRT 3 tham số của Birbaum vào việc đo lường độ khó, độ phân biệt của 20 câu hỏi trong đề thi cuối kì môn Toán Cao cấp năm 2014 của Trường Đại học Kinh tế - Luật, ĐHQG TP Hồ Chí Minh; đồng thời khảo sát ảnh hưởng dự đoán của thí sinh khi trả lời câu hỏi trắc nghiệm đối với việc đo lường và đánh giá năng lực của thí sinh. Bên cạnh đó, chúng tôi cũng tiến hành phân tích mô hình Rasch và mô hình IRT 3 tham số về mức độ phù hợp của mô hình đối với dữ liệu

được khảo sát. Từ đó suy ra mô hình tốt nhất cho việc đo lường và đánh giá chất lượng của đề thi cũng như năng lực của thí sinh.

1.4. Phương pháp nghiên cứu

Chúng tôi sử dụng phương pháp mẫu trong phân tích thống kê các dữ liệu với sự hỗ trợ của các phần mềm chuyên dụng thích hợp. Cụ thể, trên cơ sở hơn 800 bài thi cuối kì môn Toán Cao cấp của sinh viên Khóa 14 Trường Đại học Kinh tế - Luật, chúng tôi đã trích xuất một cách ngẫu nhiên 388 bài thi. Sau đó dùng gói lệnh *ltm* của phần mềm R để phân tích. Đây là gói lệnh có thể tải dễ dàng và miễn phí trên mạng tại địa chỉ *http://CRAN.R-project.org* [9]). Gói lệnh này chuyên được dùng để đo lường độ khó, độ phân biệt và mức độ dự đoán của các câu hỏi trong đề thi. Chúng tôi cũng dùng gói lệnh này để ước lượng năng lực của sinh viên và phân tích ảnh hưởng của dự đoán của thí sinh khi trả lời câu hỏi trắc nghiệm đến việc đánh giá năng lực của thí sinh. Ngoài ra, chúng tôi cũng phân tích phương sai để chọn lựa mô hình thích hợp với dữ liệu được khảo sát.

1.5. Bố cục của bài viết

Bài viết được trình bày thành 5 mục. Mục 1 là phần mở đầu nhằm giới thiệu xuất xứ vấn đề nghiên cứu, tổng quan các nghiên cứu trước đây tại Việt Nam, mục đích và phương pháp nghiên cứu. Mục 2 dành cho việc trình bày tóm lược cơ sở lí thuyết về các mô hình IRT. Mục 3 và mục 4 trình bày phương pháp và kết quả đo lường độ khó, độ phân biệt của các câu hỏi; kết quả phân tích ảnh hưởng của dự đoán của thí sinh khi trả lời câu hỏi trắc nghiệm đến việc đánh giá năng lực của thí sinh; kết quả so sánh mức độ phù hợp của các mô hình với dữ liệu được khảo sát. Mục cuối cùng, chúng tôi trình bày các kết luận và định hướng phát triển của bài viết.

2. Tóm lược về lí thuyết ứng đáp câu hỏi

Trong [8], Rasch cho rằng "*nếu một người có năng lực cao hơn người khác thì xác suất để người 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 người bất kì trả lời đúng câu hỏi đó phải nhỏ hơn xác suất để người đó trả lời đúng câu hỏi kia*". Dựa trên cơ sở này, Rasch đã mô tả mối liên hệ giữa xác suất trả lời đúng câu hỏi của mỗi thí sinh với năng lực của thí sinh đó thông qua hàm đặc trưng câu hỏi (Item Chacracteristics Function – ICF):

$$P\left(X_{jk}=1/\theta_k, b_j\right)=\frac{\exp\left(\theta_k-b_j\right)}{1+\exp\left(\theta_k-b_j\right)}, \quad (1)$$

với θ_k là năng lực của thí sinh thứ *k*, b_j là độ khó của câu hỏi thứ *j* và X_{jk} là ứng đáp của thí sinh thứ *k* đối với câu hỏi thứ *j*. $X_{jk}=1$ nếu thí sinh trả lời đúng câu hỏi và $X_{jk}=0$ nếu thí sinh trả lời sai câu hỏi.

Độ khó của câu hỏi 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 của thí sinh càng thấp. Trong [6], Baker phân loại độ khó của các câu hỏi theo 5 mức sau: *rất khó, khó, trung bình, dễ, rất dễ*. Theo Baker, một câu hỏi thuộc loại rất khó nếu tham số $b_j \geq 2$, thuộc loại khó nếu $0,5 \leq b_j < 2$, thuộc loại trung bình nếu $-0,5 \leq b_j < 0,5$, thuộc loại dễ nếu $-2 \leq b_j < -0,5$ và thuộc loại rất dễ nếu $b_j < -2$.

Trong [10], Thissen và Orlando đề xuất dạng mô hình sau, gọi là mô hình IRT 1 tham số:

$$P\left(X_{jk} = 1/\theta_k, \mathrm{a}, b_j\right) = \frac{\exp\left[a\left(\theta_k - b_j\right)\right]}{1 + \exp\left[a\left(\theta_k - b_j\right)\right]}, \quad (2)$$

trong đó tham số *a* gọi là độ phân biệt của các câu hỏi trong đề thi.

Trong [7], Birnbaum đề xuất mở rộng mô hình IRT 1 tham số bằng cách gán cho mỗi câu hỏi trong đề thi trắc nghiệm ứng với một độ phân biệt a_j khác nhau. Mô hình này được gọi là mô hình IRT 2 tham số. Hàm đặc trưng câu hỏi của mô hình có dạng:

$$P\left(X_{jk} = 1/\theta_k, \mathrm{a}_j, b_j\right) = \frac{\exp\left[a_j\left(\theta_k - b_j\right)\right]}{1 + \exp\left[a_j\left(\theta_k - b_j\right)\right]} \quad (3)$$

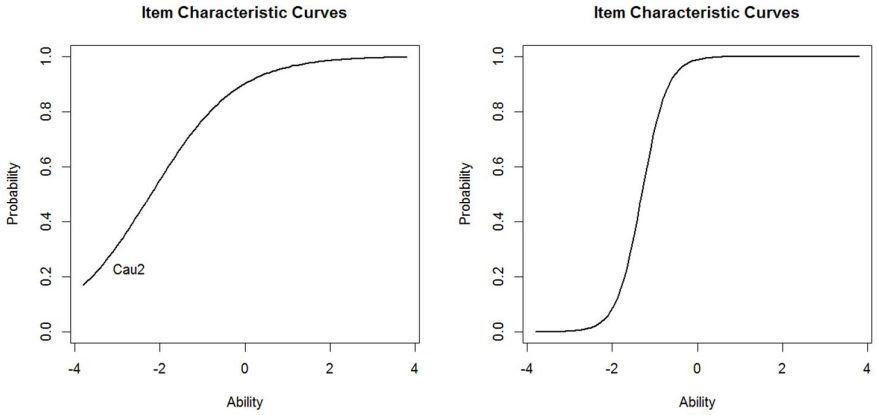
Độ phân biệt của câu hỏi đặc trưng cho khả năng phân loại thí sinh. Thông thường độ phân biệt của câu hỏi có giá trị dương. Trong trường hợp câu hỏi sai hoặc mắc lỗi thiết kế thì độ phân biệt có thể mang giá trị âm [6]. Câu hỏi có độ phân biệt dương càng lớn thì sự chênh lệch về xác suất trả lời đúng của các thì sinh có năng lực cao và năng lực thấp càng lớn. Nói một cách khác, câu hỏi có độ phân biệt cao phân loại thí sinh tốt hơn câu hỏi có độ phân biệt thấp.

Trong [6], Baker chia độ phân biệt của các câu hỏi thành 5 mức: *rất tốt, tốt, bình thường, kém và rất kém*. Cụ thể một câu hỏi được gọi là có độ phân biệt rất tốt nếu tham số $a_j \geq 1,7$, loại tốt nếu $1,35 \leq a_j < 1,7$, loại bình thường nếu $0,65 \leq a_j < 1,35$, loại kém nếu $0,35 \leq a_j < 0,65$ và loại rất kém nếu $a_j < 0,35$.

Thực tế cho thấy, trong quá trình kiểm tra trắc nghiệm khách quan nhiều lực chọn, thí sinh luôn dự đoán câu trả lời (theo cách chọn ngẫu nhiên một phương án hoặc theo cách loại suy dựa trên kinh nghiệm bản thân). Trong lí thuyết trắc nghiệm cổ điển, người ta giảm việc dự đoán của thí sinh khi trả lời câu hỏi bằng cách đưa vào điểm may rủi. Tuy nhiên, cách làm này có nhược điểm là xem các câu hỏi có độ may rủi như nhau. Điều này trái với thực tiễn vì thí sinh thường dự đoán để trả lời đúng câu hỏi khi gặp câu hỏi khó hơn là khi gặp câu hỏi dễ. Vì vậy, Birnbaum đề xuất thêm tham số $c_j \in (0,1)$ vào mô hình IRT 2 tham số để đo lường mức độ dự đoán của thí sinh khi trả lời câu hỏi trắc nghiệm trong mỗi câu hỏi [7]. Mô hình với tham số đo lường mức độ

dự đoán của thí sinh được gọi là mô hình IRT 3 tham số. Hàm đặc trưng câu hỏi của mô hình có dạng sau:

$$P\left(X_{jk}=1/\theta_k,a_j,b_j,c_j\right)=c_j+\left(1-c_j\right).\frac{\exp\left[a_j\left(\theta_k-b_j\right)\right]}{1+\exp\left[a_j\left(\theta_k-b_j\right)\right]} \quad (4)$$



Hình 1. *Mô hình Rasch và mô hình IRT 3 tham số*

Trong 1, đồ thị bên phải là đường cong đặc trưng câu hỏi của mô hình IRT 3 tham số và đồ thị bên trái là đường cong đặc trưng của mô hình Rasch, là mô hình không xét đến yếu tố dự đoán của thí sinh khi trả lời câu hỏi. So với đường cong đặc trưng của mô hình Rasch, đường cong đặc trưng của mô hình IRT 3 tham số có độ dốc lớn hơn và lệch về bên phải. Điều này có nghĩa là độ khó và độ phân biệt của câu hỏi tăng lên khi xét thêm yếu tố dự đoán của thí sinh. Sự gia tăng độ khó, độ phân biệt của câu hỏi này trong mô hình IRT 3 tham số dẫn đến điểm số của thí sinh đạt được khi có câu trả lời đúng tăng lên. Nói một cách khác, yếu tố dự đoán đã tác động đến việc đánh giá năng lực của thí sinh.

3. Mô tả cụ thể phương pháp

Trên cơ sở hơn 800 bài thi cuối kì cuối kì môn Toán Cao cấp của sinh viên Khóa 14 Trường Đại học Kinh tế - Luật, ĐHQG TPHCM, chúng tôi trích xuất ngẫu nhiên 388 bài thi (chiếm tỉ lệ xấp xỉ 46,74%) và lấy kết quả từng câu hỏi để phân tích. Đề thi gồm 20 câu hỏi trắc nghiệm khách quan 4 lựa chọn. Chúng tôi mã hóa dữ liệu thành dạng nhị phân theo quy tắc: Ứng với mỗi câu hỏi, mỗi thí sinh khi trả lời đúng thì được gán giá trị 1, ngược lại được gán giá trị 0.

Trước tiên, chúng tôi áp dụng mô hình Rasch để đo lường độ khó của các câu hỏi trong đề thi nói trên. Tiếp theo, mô hình IRT 3 tham số được áp dụng để ước lượng độ khó, độ phân biệt và mức độ dự đoán của mỗi câu hỏi trong đề thi. Căn cứ vào các kết quả này, chúng tôi tiến hành phân loại và đánh giá các câu hỏi dựa theo các thang đo của [6]. Năng lực của mỗi thí sinh ứng với mỗi mô hình được tính toán từ các công

thức (1) và (4). So sánh các kết quả này, chúng tôi đánh giá được ảnh hưởng của các tham số đến việc đánh giá năng lực của mỗi thí sinh. Cuối cùng, phân tích phương sai được chúng tôi áp dụng để so sánh mức độ phù hợp của mô hình Rasch và mô hình IRT 3 tham số với dữ liệu được khảo sát. Việc ước lượng các tham số của các mô hình nói trên cũng như ước lượng năng lực của mỗi thí sinh và phân tích phương sai được thực hiện bằng gói lệnh *ltm* của phần mềm R. [9]

4. Kết quả cụ thể

4.1. Phân tích độ khó, độ phân biệt và mức dự đoán của các câu hỏi

Để ước lượng độ khó của các câu hỏi trong mô hình Rasch, chúng tôi dùng lệnh `rasch()`. Bảng sau đây mô tả kết quả ước lượng độ khó của các câu hỏi trong đề thi.

Bảng 1. *Độ khó của các câu hỏi trong mô hình Rasch*

	value	std.err	z.vals
Item1	– 0.7884	0.1256	– 6.2775
Item2	– 2.2140	0.1700	– 13.0020
Item3	– 2.2137	0.1700	– 13.0215
Item4	– 1.8848	0.1549	– 12.1664
Item5	– 0.3622	0.1211	– 2.9918
…			

Các giá trị của cột `value` chỉ độ khó của các câu hỏi, các giá trị của cột `std.err` chỉ sai số của độ lệch chuẩn và cột `z.vals`, cột cuối cùng, chỉ độ khó của các câu hỏi được quy đổi sang dạng chuẩn. Theo các mức phân loại trong [6], đề thi này có 1 câu thuộc loại khó, 9 câu thuộc loại trung bình, 8 câu thuộc loại dễ và 2 câu ở mức rất dễ.

Đối với mô hình IRT 3 tham số, chúng tôi dùng câu lệnh `tpm()` và `coeff()` để ước lượng độ khó, độ phân biệt và mức độ dự đoán của thí sinh trong mỗi câu hỏi. Kết quả được trình bày trong bảng sau:

Bảng 2. *Mô hình IRT 3 tham số*

	Gussng	Diffclt	Dscrmn
Item1	0.0000	– 1.0481	0.7403
Item2	0.0000	– 1.3040	3.4131
Item3	0.2352	– 1.3347	1.9398
Item4	0.4526	– 0.6019	3.9070
Item5	0.0001	– 0.6927	0.4882
…			

Các giá trị của cột `Gussng` chỉ mức dự đoán của thí sinh của các câu hỏi, cột `Diffclt` chỉ độ khó của các câu hỏi và cột cuối `Dscrmn` chỉ độ phân biệt của các câu hỏi. Từ các kết quả này, chúng tôi có một số đánh giá như sau: Đề thi có 1 câu hỏi ở mức rất khó, 6 câu hỏi ở mức khó, 1 câu hỏi ở mức trung bình, 10 câu hỏi ở mức dễ và

2 câu hỏi ở mức rất dễ. So với kết quả đánh giá trong mô hình Rasch, kết quả của mô hình IRT 3 tham số sát với dữ liệu thực tế của chúng tôi hơn; đồng thời kết quả này tương đối phù hợp với ma trận câu hỏi và chuẩn đầu ra môn học của chúng tôi. Tiếp tục với độ phân biệt của các câu hỏi, đề thi có 6 câu hỏi ở mức phân biệt rất tốt, 1 câu hỏi ở mức tốt, 6 câu hỏi ở mức bình thường, 4 câu hỏi ở mức kém và 3 câu hỏi ở mức rất kém. Tổng hợp các kết quả phân tích độ khó và độ phân biệt của các câu hỏi, chúng tôi thấy các câu hỏi 6, 9 là câu hỏi có chất lượng rất tốt. Các câu hỏi ở mức tương đối tốt là 1, 2, 3, 4, 11, 12, 13, 14, 16, 17. Câu câu hỏi 19, 8, 5, 7 ở mức khá; tuy nhiên, cần điều chỉnh mồi nhử để đạt được độ phân biệt tốt hơn. Câu hỏi 10 và 18 cần thay thế hoặc cải tiến nhiều hơn vì độ phân biệt rất kém. Đối với câu 15, độ phân biệt có giá trị âm. Điều này có nghĩa là thí sinh có năng lực cao lại có khả năng sai nhiều hơn thí sinh có năng lực thấp. Phân tích câu 15, chúng tôi cho rằng nguyên nhân xảy ra hiện tượng này có thể do cách sử dụng từ đa nghĩa và cấu trúc phủ định của phủ định. Do đó, thí sinh hiểu sai ý câu hỏi hoặc không hiểu câu hỏi. Vì vậy, thí sinh chọn đáp án theo cách ngẫu nhiên hoặc chọn sai đáp án. Thông thường, với câu hỏi dễ, thí sinh thường sẽ chọn ngay câu trả lời đúng mà không cần dự đoán. Tuy nhiên, trong trường hợp câu hỏi 15 (là câu hỏi ở mức rất dễ), mức độ dự đoán là khá cao so với những câu hỏi ở cùng mức độ. Điều này khẳng định suy đoán có lỗi thiết kế trong câu hỏi 15 của chúng tôi là hợp lí. Tiếp theo, chúng tôi tiến hành phân tích mức độ dự đoán của thí sinh trong mỗi câu hỏi để xác định ảnh hưởng của chúng đến việc đánh giá năng lực của thí sinh. Theo Bảng 2, chúng tôi thấy rằng, đối với các câu hỏi dễ, mức dự đoán của thí sinh thường nhỏ, thậm chí gần bằng 0; chẳng hạn như câu hỏi 1, 2, 12, 14, 17. Các câu hỏi càng khó thì tỉ lệ phỏng đoán của thí sinh càng tăng; Ví dụ: câu hỏi 9, là câu hỏi khó, có mức dự đoán gần đến 50%.

4.2. Phân tích ảnh hưởng của dự đoán đến năng lực của thí sinh

Từ các tham số được ước lượng trong phần trên, chúng ta có thể ước lượng được năng lực của mỗi thí sinh thông qua câu lệnh `factor.scores()`. Kết quả ở Bảng 3 và Bảng 4 mô tả tương ứng năng lực của thí sinh khi đánh giá bằng mô hình Rasch và mô hình IRT 3 tham số.

Bảng 3. *Năng lực của thí sinh ứng với mô hình Rasch*

	Abilities	Std.Err	No.
Person1	-1.922	0.489	3
Person2	-1.922	0.489	3
Person3	-1.061	0.446	7
Person4	-1.061	0.446	7
Person5	-1.061	0.446	7
Person6	-0.865	0.442	8
Person7	-1.263	0.453	6
…			

Bảng 4. *Năng lực của thí sinh ứng với mô hình IRT 3 tham số*

	Abilities	Std.Err	No.
Person1	-1.675	0.418	3
Person2	-2.224	0.57	3
Person3	-1.548	0.398	7
Person4	-1.663	0.42	7
Person5	-1.842	0.461	7
Person6	-1.477	0.38	8
Person7	-1.91	0.486	6
…			

Các giá trị trong cột `Abilities` là ước lượng năng lực của thí sinh; `Std.Err` là sai số của ước lượng và `No.` là tổng số câu trả lời đúng của thí sinh. Kết quả ở Bảng 3 cho thấy đối với mô hình Rasch, 2 thí sinh có tổng số câu trả lời đúng bằng nhau thì năng lực của các thí sinh được đánh giá là như nhau. Trong khi đó kết quả ở Bảng 4 cho thấy khi dùng mô hình IRT 3 tham số để đánh giá, năng lực của thí sinh phụ thuộc vào độ khó, độ phân biệt và mức độ dự đoán của mỗi câu hỏi. Ví dụ: hai thí sinh 1 và 2 có tổng số câu trả lời đúng như nhau (thí sinh thứ nhất trả lời đúng câu hỏi 10, 11, 12 còn thí sinh thứ hai trả lời đúng câu hỏi 9, 11, 15). Tuy nhiên, kết quả đánh giá năng lực của thí sinh thứ nhất cao hơn thí sinh thứ hai vì mức độ dự đoán câu trả lời của các câu hỏi 9, 11, 15 cao hơn rất nhiều so với mức độ dự đoán câu trả lời của các câu hỏi 10, 11, 12. Điều này chứng tỏ ảnh hưởng của mức độ dự đoán câu trả lời của các câu hỏi đến việc đánh giá năng lực của thí sinh.

4.3. So sánh mức độ phù hợp của các mô hình

Kết quả trong bảng tiếp theo cho phép chúng ta đánh giá và chọn lựa mô hình tối ưu cho dữ liệu được khảo sát.

Bảng 5. *So sánh mô hình Rasch và mô hình IRT 3 tham số*

	Likelihood ratio table					
	AIC	BIC	log.Lik	LRT	df	p.value
Rasch	9271.18	9350.40	- 4615.59			
3PL	9098.79	9336.45	- 4489.39	252.39	40	<0.001

Theo lí thuyết chọn lựa mô hình, mô hình tốt hơn là mô hình có các chỉ số AIC, BIC và log.Lik nhỏ hơn [9]. Bảng 5 cho thấy mô hình IRT 3 tham số (3PL) là mô hình tốt hơn, theo nghĩa phù hợp với dữ liệu thực tế hơn. Điều này hoàn toàn nhất quán với các phân tích ở phần trên về sự phù hợp của độ khó, độ phân biệt của các câu hỏi và đánh giá năng lực của thí sinh đối với dữ liệu được khảo sát.

5. Kết luận

Bài viết đã nêu được quy trình chi tiết cho việc đo lường, đánh giá độ khó, độ phân biệt và mức độ dự đoán của thí sinh khi trả lời các câu hỏi trắc nghiệm khách quan nhiều lựa chọn. Và cũng đã đánh giá ảnh hưởng của các tham số của mô hình đến việc đánh giá năng lực của thí sinh; đồng thời so sánh và chọn lựa được mô hình thích hợp cho dữ liệu được khảo sát.

Kết quả đo lường độ khó, độ phân biệt và mức dự đoán câu trả lời của các câu hỏi trong đề thi trắc nghiệm môn Toán Cao cấp ở Trường Đại học Kinh tế - Luật là cơ sở để giáo viên và nhà quản lí giáo dục đánh giá chất lượng đề thi, năng lực thí sinh và xây dựng ngân hàng câu hỏi trắc nghiệm.

Quy trình đo lường và đánh giá này có thể áp dụng không chỉ cho môn Toán Cao cấp mà còn cho nhiều môn học khác; và không chỉ cho hình thức trắc nghiệm khách quan nhiều lựa chọn mà còn cho nhiều hình thức kiểm tra khác. Vì vậy theo chúng tôi, bài viết có tính ứng dụng cao.

Kết quả của bài viết khuyến khích việc đánh giá năng lực của thí sinh theo hình thức mới, dựa vào độ khó, độ phân biệt và mức dự đoán câu trả lời. Tuy nhiên, chúng tôi ý thức được rằng, cách đánh giá này sẽ vấp phải một số khó khăn. Một trong số các khó khăn đó là việc thí sinh cũng như các giáo viên đã quen với cách tính điểm theo tổng số câu trả lời đúng. Họ chưa sẵn sàng thay đổi cách đánh giá và chấp nhận sự đánh giá mới.

Mục đích cuối cùng của kiểm tra là đánh giá năng lực của người học. Tuy nhiên kết quả đánh giá năng lực người học của mô hình IRT thường không quen thuộc với người học cũng như giáo viên. Do đó, việc nghiên cứu và áp dụng cách chuyển đổi từ kết quả của mô hình IRT sang các hình thức cho điểm thông thường, chẳng hạn thang điểm 10, là vấn đề tiếp theo bài viết này.

TÀI LIỆU THAM KHẢO

1. Nguyễn Thị Hồng Minh, Nguyễn Đức Thiện (2004), "Đo lường đánh giá trong thi trắc nghiệm khách quan: Độ khó câu hỏi và khả năng của thí sinh", *Tạp chí khoa học*, ĐHQG Hà Nội, 197-214.
2. 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*, Đại học Đà Nẵng, (2), 119-126.
3. Lâm Quang Thiệp (2003), *Giới thiệu về đo lường và đánh giá trong giáo dục*, Nxb Giáo dục.
4. Dương Thiệu Tống (2005), *Trắc nghiệm và đo lường thành quả học tập*, Nxb Khoa học xã hội.

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, (12), 24-27.
6. Baker, F. (2001), *The basic of item response theory*, ERIC Clearinghouse on Assessment and Evaluation.
7. Birnbaum, A. (1968), "Some latent trait models and their use in inferring an examinee's ability", *Statistical theory of Mental test scores*, Reading: Addison Wesley, 395-479.
8. Rasch, G. (1960), *Probabilistic Models for some Intelligence and Attainment Tests*, Copenhagen, Denmark.
9. Rizopoulos, D. (2006), "ltm: An R package for latent variable modeling and item response theory analysis", *Journal of Statistical software*, 17, 1-25.
10. Thissen, D. & Orlando, M. (2001), *Chapter 3 – Item response theory for item scores in two categories. In D. Thissen & H. Wainer (Eds), Test scoring*, Hillsdale, NJ: Erlbaum.
11. Benjamin, D. Wright & Stone, M. H. (1979), *Best test design*, SMESA PRESSA, Chicago.

PHỤ LỤC

PHỤ LỤC 1. Kết quả ước lượng độ khó của các câu hỏi trong mô hình Rasch

Coefficients:

	Value	Std.err	z.vals
Dffclt.Cau1	-0.7884	0.1256	-6.2775
Dffclt.Cau2	-2.2140	0.1700	-13.0220
Dffclt.Cau3	-2.2137	0.1700	-13.0215
Dffclt.Cau4	-1.8848	0.1549	-12.1664
Dffclt.Cau5	-0.3622	0.1211	-2.9918
Dffclt.Cau6	0.8624	0.1262	6.8349
Dffclt.Cau7	0.4939	0.1218	4.0561
Dffclt.Cau8	-0.0885	0.1199	-0.7385
Dffclt.Cau9	-0.1122	0.1199	-0.9351
Dffclt.Cau10	-0.3622	0.1211	-2.9917
Dffclt.Cau11	0.0174	0.1198	0.1454
Dffclt.Cau12	-1.5372	0.1425	-10.7900
Dffclt.Cau13	0.4452	0.1214	3.6678
Dffclt.Cau14	-1.6090	0.1448	-11.1143
Dffclt.Cau15	0.4695	0.1216	3.8623
Dffclt.Cau16	-0.5334	0.1225	-4.3545
Dffclt.Cau17	-1.4508	0.1399	-10.3729
Dffclt.Cau18	-0.6973	0.1243	-5.6080
Dffclt.Cau19	-0.5832	0.1230	-4.7417
Dffclt.Cau20	-0.0768	0.1199	-0.6407

PHỤ LỤC 2. Kết quả ước lượng độ khó, độ phân biệt và mức độ dự đoán của các câu hỏi trong mô hình IRT 3 tham số

	Gussng	Dffclt	Dscrmn
Cau1	1.872309e-05	-1.0480792	0.74033620
Cau2	1.597029e-08	-1.3040327	3.41314886
Cau3	2.352452e-01	-1.3347035	1.93978292
Cau4	4.526242e-01	-0.6019112	3.90700529
Cau5	9.283560e-05	-0.6927461	0.48816302
Cau6	3.030104e-01	2.0426714	8.83408331
Cau7	2.148219e-02	1.3966637	0.35883916
Cau8	2.536327e-01	1.0917708	0.57895799
Cau9	4.798526e-01	1.3967295	7.03038792
Cau10	1.201517e-04	-1.1309911	0.28978012
Cau11	1.460698e-01	0.4256194	1.04176835
Cau12	2.955705e-08	-1.0977862	1.94249834
Cau13	9.672185e-06	0.6502781	0.65602596
Cau14	9.532632e-06	-1.9215262	0.84491280
Cau15	1.682643e-02	-4.5616876	-0.09893687
Cau16	3.835617e-01	0.5642493	1.12563487
Cau17	4.405779e-06	-1.3204629	1.23967710
Cau18	1.758819e-02	-2.4287461	0.24584817
Cau19	1.269043e-04	-0.8906735	0.62764588
Cau20	3.190117e-01	0.8152564	1.54708412

(Ngày Tòa soạn nhận được bài: 04-5-2016; ngày phản biện đánh giá: 25-5-2016; ngày chấp nhận đăng: 22-7-2016)