TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
TẠP CHÍ KHOA HỌC
KHOA HỌC GIÁO DỤC
ISSN: 1859-3100
Tập 14, Số 4 (2017): 119-130

HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
EDUCATION SCIENCE
Vol. 14, No. 4 (2017): 119-130

Email: tapchikhoahoc@hcmue.edu.vn; Website: http://tckh.hcmue.edu.vn

ÁP DỤNG LẤY MẪU GIBBS VÀO ĐO LƯỜNG VÀ ĐÁNH GIÁ ĐỘ KHÓ CÂU HỎI TRONG MÔ HÌNH RASCH

Lê Anh Vũ[1]*, Phạm Hoàng Uyên[1], Đoàn Hồng Chương[1], Lê Thanh Hoa[1,2]

[1]Trường Đại học Kinh tế - Luật – ĐHQG TPHCM

[2]Trường Đại học Khoa học Tự nhiên – ĐHQG TPHCM

Ngày Tòa soạn nhận được bài: 21-01-2017; ngày phản biện đánh giá: 18-4-2017; ngày chấp nhận đăng: 24-4-2017

TÓM TẮT

Trong nghiên cứu này, chúng tôi áp dụng lấy mẫu Gibbs để ước lượng độ khó của các câu hỏi trong mô hình Rasch. Dữ liệu để phân tích được thu thập ngẫu nhiên từ các bài thi cuối kì môn Toán Cao cấp của sinh viên niên Khóa 2014, Trường Đại học Kinh tế-Luật, ĐHQG TPHCM. Thuật toán trình bày trong nghiên cứu này là đơn giản và có tính ứng dụng cao.

Từ khóa: lấy mẫu Gibbs, phương pháp hợp lí cực đại biên (MML), mô hình Rasch.

ABSTRACT

Using Gibbs Sampler to evaluate item difficulty in Rasch model

In this study, we use Gibbs Sampler to estimate the difficulty of items in Rasch model. Data are based on 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. The investigated algorithm in this study is simple and highly applicable.

Keywords: Rasch model, Marginal Maximum Likelihood, Gibbs Sampler.

1. Mở đầu

Lí thuyết trắc nghiệm cổ điển (Classical Test Theory, viết tắt là CTT) ra đời từ 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 (Bechger et al., 2003). Mặc dù, CTT rất dễ áp dụng để đo lường và đánh giá đề thi trắc nghiệm khách quan vì nó hầu như không đòi hỏi bất kì giả thiết nào khi chạy mô hình, nhưng phương pháp này tồn tại một số hạn chế (Morales, 2009). Các hạn chế đó là sự phụ thuộc của các tham số đặc trưng (độ khó, độ phân biệt…) của các câu hỏi vào mẫu thí sinh tham gia kiểm tra và sự ảnh hưởng của đề thi đến việc đo lường và đánh giá năng lực của thí sinh. Theo Rasch (1960), phân tích trong đo lường và đánh giá đề thi trắc nghiệm khách quan chỉ đáng giá khi dựa vào

* Email: *vula@uel.edu.vn*

từng cá nhân thí sinh, với các thuộc tính của thí sinh và của các câu hỏi được tách riêng. Quan điểm của Rasch đã đánh dấu sự chuyển tiếp từ mô hình CTT sang mô hình lí thuyết ứng đáp câu hỏi (Item Response Theory, viết tắt là IRT), là mô hình xác suất mô tả xác suất trả lời đúng các câu hỏi trong đề thi trắc nghiệm khách quan đối với sự ứng đáp của thí sinh đối với các câu hỏi đó (Camilli và Shepard, 1994). Điều này có nghĩa là trong mô hình IRT, các tham số đặc trưng của các câu hỏi độc lập đối với mẫu được khảo sát (Hambleton và Swaminathan).

Những ý tưởng sơ khởi của mô hình IRT được đề cập đến đầu tiên trong bài báo của Thurstone (1925). Sau đó Lord (1952), đề xuất khái niệm đường cong đặc trưng câu hỏi (Item Characteristic Curve, viết tắt là ICC). ICC mô tả mối liên hệ giữa xác suất trả lời đúng câu hỏi *j* với năng lực của thí sinh *i*, năng lực này thường được kí hiệu là θ_i. Birnbaum (1968), đề xuất dùng mô hình logistic cho IRT. Sau đó Lord và Novick (1968), Bock và Aitkin (1981) đã mở rộng và hoàn thiện các mô hình IRT đồng thời xây dựng các phương pháp ước lượng các tham số của mô hình bằng phương pháp hợp lí cực đại biên (Marginal Maximum Likelihood, viết tắt là MML). Tiếp cận theo một cách khác, năm 1960, Rasch giới thiệu một mô hình mà sau này được gọi là mô hình Rasch. Mô hình của Rasch được dựa trên giả thiết cơ bản sau:

> Nếu một người có năng lực cao hơn người khác thì xác suất 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 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 (Rasch, 1960, p. 117).

Điểm nổi bật của mô hình này, cũng như của các mô hình IRT khác, 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 (Wright và Stone, 1979; Baker, 2001). Theo Rasch (1960), ứng với mỗi mức năng lực θ_i, khả năng trả lời đúng câu hỏi của thí sinh là xác suất $P(\theta_i)$. Xác suất này chỉ phụ thuộc vào năng lực của thí sinh và các tham số đặc trưng của mỗi câu hỏi.

Thông thường, đối với các mô hình IRT, phương pháp ước lượng tham số phổ biến là MML. Tuy nhiên, trong thời gian gần đây, sự phát triển mạnh mẽ của thống kê Bayes đã thu hút nhiều tác giả quan tâm nghiên cứu. Nghiên cứu của Gelfand và Smith (1990) về cách dùng MCMC (Markov chain Monte Carlo) cho các phân phối hậu nghiệm (posterior distribution) trong thống kê Bayes là một bước ngoặt lớn và đưa MCMC trở thành một phương pháp phổ biến trong thống kê hiện đại (Lynch, 2007). Kĩ thuật MCMC cho phép tạo ra các mẫu từ một hàm mật độ xác suất định trước bằng cách rút ra các phần tử của mẫu từ hàm mật độ xác suất đơn giản hơn (Liu, 2004; Liang, Liu và Caroll, 2010). Hai kĩ

thuật MCMC phổ biến nhất là thuật toán Metropolis-Hasting (MH) và lấy mẫu Gibbs (Gibbs Sampling) (Rober và Casella, 2004). Lấy mẫu Gibbs được sử dụng rộng rãi trong các quá trình thu phát tín hiệu (Ruanaidh và Fitzgeral, 1996) hoặc trong máy học (Machine Learning) (Doucet, và Wang, 2005) để tạo ra các mẫu từ hàm mật độ đa chiều bằng cách rút ra các phần tử từ hàm mật độ xác suất có điều kiện đơn chiều tương ứng. Lấy mẫu Gibbs đặc biệt có hiệu quả đối với các trường hợp xác suất có điều kiện có dạng phức tạp (Martino, Read, và Luengo, 2015).

Nguyễn Thị Hồng Minh và Nguyễn Đức Thiện (2004) đã trình bày phương pháp PROX để ước lượng các tham số cho mô hình Rasch. Nguyễn Bảo Hoàng Thanh (2008) và Nguyễn Thị Ngọc Xuân (2014) đã nghiên cứu mô hình IRT 2 tham số bằng phần mềm Quest và ConQuest. Lê Anh Vũ và các tgk (2016) đã nghiên cứu mô hình IRT 3 tham số và đo lường năng lực của các thí sinh theo mô hình này. Tuy nhiên, việc áp dụng thống kê Bayes vào ước lượng các tham số của mô hình Rasch cũng như các mô hình IRT khác chưa được các tác giả ở trên quan tâm và nghiên cứu. Thêm nữa, việc ước lượng các tham số của mô hình trong các nghiên cứu nói trên tương đối khó thực hiện với đa số giáo viên. Vì vậy, việc xây dựng thuật toán đơn giản là nhu cầu thiết yếu của nghiên cứu này.

Nghiên cứu này nhằm giải quyết các mục tiêu sau:

- Làm cách nào để ước lượng được các tham số trong mô hình Rasch bằng phương pháp lấy mẫu Gibbs của thống kê Bayes?
- So sánh việc ước lượng các tham số trong mô hình Rasch bằng phương pháp lấy mẫu Gibbs và phương pháp MML. Phương pháp nào phù hợp với dữ liệu của nghiên cứu hơn?

Để thực hiện các mục tiêu nói trên, chúng tôi tiến hành khảo sát bài thi cuối kì môn Toán Cao cấp của sinh viên Khóa 14, Trường Đại học Kinh tế - Luật, ĐHQG TPHCM. Chúng tôi lấy mẫu ngẫu nhiên gồm 388 bài thi của 800 sinh viên tham gia kì thi (chiếm tỉ lệ 46,74%). Sau đó, 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, các trường hợp khác được gán giá trị 0.

Lấy mẫu Gibbs được áp dụng vào bộ dữ liệu thô nói trên và các tham số của mô hình được ước lượng từ mẫu rút ra từ lấy mẫu Gibbs. Hệ số tương quan Pearson được áp dụng để đo lường mức độ tương quan của các tham số ước lượng từ mô hình bằng phương pháp lấy mẫu Gibbs và phương pháp MML.

Bài viết được trình bày thành 5 mục. Mục 1 là phần mở đầu nhằm giới thiệu vấn đề nghiên cứu, tổng quan các nghiên cứu trước đây về mô hình IRT ở Việt Nam và mục đích, phương pháp nghiên cứu. Mục 2 dành cho việc trình bày tóm lược cơ sở lí thuyết về mô hình Rasch và phương pháp lấy mẫu Gibbs trong thống kê Bayes. Mục 3 trình bày chi tiết phương

pháp của nghiên cứu và thuật toán cụ thể cho việc ước lượng các tham số của mô hình. Mục 4 trình bày kết quả đo lường độ khó của các câu hỏi trong mô hình Rasch bằng cách áp dụng lấy mẫu Gibbs và so sánh mức độ mức độ tương quan của các kết quả khi dùng cách lấy mẫu Gibbs và mô hình dùng MML. Trong mục 5, mục cuối cùng, chúng tôi trình bày một số kết luận về kết quả của nghiên cứu cũng như định hướng phát triển sau này.

2. Cơ sở lí thuyết

2.1. Mô hình Rasch

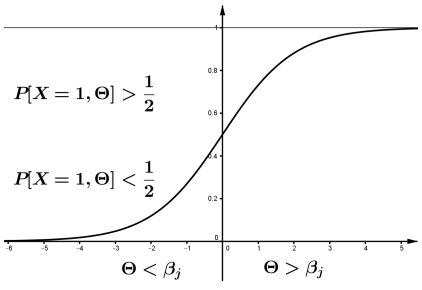
Theo Rasch (1960), phân tích trắc nghiệm chỉ có ý nghĩa khi dựa vào từng cá nhân thí sinh, với các thuộc tính của thí sinh và câu hỏi được tách riêng. Do đó, Rasch cho rằng:

> 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 ngườiđó trả lời đúng câu hỏi kia (Rasch, 1960, p. 117).

Dựa trên quan điểm này, Rasch xây dựng một mô hình toán học cho sự ứng đáp câu hỏi của mỗi thí sinh. Công thức của mô hình có dạng sau:

$$P\left(X_{ij}=1\,|\,\theta_i,b_j\right)=\frac{\exp\left(\theta_i-b_j\right)}{1+\exp\left(\theta_i-b_j\right)}, \tag{1}$$

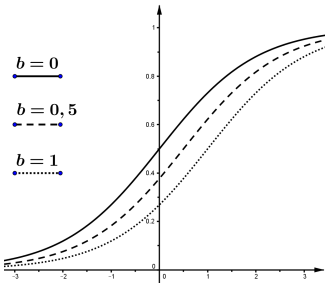
trong đó: θ_i là năng lực của thí sinh i, b_j là độ khó của câu hỏi j, exp(.) là kí hiệu của hàm số mũ cơ số e, và X_{ij} là ứng đáp của thí sinh i đối với câu hỏi j. $X_{ij}=1$ nếu thí sinh trả lời đúng câu hỏi và $X_{ij}=0$ trong các trường hợp còn lại. Nếu chúng ta vẽ đồ thị của hàm số trong công thức (1) theo biến θ_i thì đồ thị này sẽ có dạng hình chữ S như Hình 1.



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

Trong lí thuyết ứng đáp câu hỏi, đường cong hình chữ S này được gọi là đường cong đặc trưng câu hỏi (Item Characteristic Curve, viết tắt là ICC). ICC có độ dốc hướng lên

biểu thị cho xác suất trả lời đúng câu hỏi của thí sinh tỉ lệ thuận với năng lực của thí sinh và xác suất này sẽ tiến dần về 1 khi năng lực của thí sinh tiến đến dương vô cùng. Trong mô hình Rasch, nếu năng lực θ_i của thí sinh bằng với độ khó câu hỏi b_j thì khả năng trả lời đúng câu hỏi của thí sinh là 50%. Mức năng lực này được gọi là ngưỡng của câu hỏi (threshold). Nói một cách khác, độ khó của mỗi câu hỏi chính là ngưỡng mà với năng lực đó, khả năng trả lời đúng câu hỏi của thí sinh là 50%. Hình 2 tiếp theo đây cho thấy câu hỏi nào có ngưỡng càng cao thì càng khó. Cụ thể hơn, câu hỏi có ngưỡng cao thì xác suất trả lời đúng câu hỏi đó của thí sinh sẽ thấp. Một cách trực quan, câu hỏi khó thì ICC của nó sẽ nằm dưới ICC của câu hỏi dễ.



Hình 2. *ICC trong mô hình Rasch ứng với các câu hỏi có độ khó khác nhau*

Theo Baker (2001), độ khó của các câu hỏi được chia thành 5 mức: rất khó (very hard), khó (hard), trung bình (medium), dễ (easy) và rất dễ (very easy). Cụ thể việc phân loại như sau: một câu hỏi được xếp vào loại rất khó nếu ngưỡng của nó (hay độ khó) có giá trị lớn hơn hay bằng 2; loại khó nếu ngưỡng của nó thuộc khoảng 0,5 đến 2; loại trung bình nếu ngưỡng của nó thuộc khoảng -0,5 đến 0,5; loại dễ nếu ngưỡng của nó thuộc khoảng -2 đến -0,5 và một câu hỏi được xếp vào loại rất dễ nếu ngưỡng của nó nhỏ hơn -2. Như vậy, mô hình Rasch chỉ quan tâm đến độ khó của câu hỏi bởi vì Rasch cho rằng đối với dữ liệu dạng nhị phân thì chỉ có độ khó của câu hỏi là có thể ước lượng được một cách ổn định và đầy đủ. Vì vậy, mặc dù mô hình Rasch là mô hình đơn giản nhất trong các mô hình IRT nhưng mô hình Rasch vẫn được sử dụng nhiều nhất trong các nghiên cứu tâm lí và giáo dục.

2.2. Phương pháp lấy mẫu Gibbs

Gibbs Sampler là kĩ thuật cho phép chúng ta tạo ra một mẫu số liệu từ hàm phân phối xác suất đồng thời mà không đòi hỏi phải biết đầy đủ thông tin về phân phối này. Khi áp dụng Gibbs Sampler, chúng ta chỉ cần biết thông tin về các phân phối xác suất có điều kiện. Ví dụ đơn giản sau minh họa cho trường hợp hàm phân phối xác suất đồng thời có 2 biến $f(x, y)$. Trước tiên, chúng ta thấy rằng:

$$f(x,y) = f(y|x).f(x) = \frac{f(y|x)}{\left[\frac{1}{f(x)}\right]},$$

và

$$\frac{1}{f(x)} = \int \frac{f(y)}{f(x)} dy = \int \frac{\frac{f(x,y)}{f(x)}}{\frac{f(x,y)}{f(y)}} dy = \int \frac{f(y|x)}{f(x|y)} dy$$

Do đó, hàm phân phối xác suất đồng thời $f(x,y)$ có thểđược viết lại như sau:

$$f(x,y) = \frac{f(y|x)}{\int \frac{f(y|x)}{f(x|y)} dy}. \quad (2)$$

Công thức (2) cho thấy hàm phân phối xác suất $f(x,y)$ có thể xác định được nếu chúng ta biết đầy đủ các hàm phân phối xác suất có điều kiện.

Tổng quát hơn, giả sử chúng ta muốn tính toán một số đại lượng liên quan đến hàm phân phối xác suất đồng thời $f(x_1, x_2, ..., x_k)$ và nó không dễ để tính trực tiếp. Khi đó chúng ta có thể dùng Gibbs Sampler để tạo ra một mẫu xấp xỉ với đại lượng cần tính từ $f(x_1, x_2, ..., x_k)$ bằng cách dùng các phân phối xác suất có điều kiện của các biến $x_i, i = 1, ..., k$. Quá trình tính toán trên được mô tả trong thuật toán Gibbs Sampler sau:

Thuật toán Gibbs Sampler

Bước 1. chọn giá trị xuất phát $x^{(0)} = \left(x_1^{(0)}, x_2^{(0)}, ..., x_k^{(0)}\right)$.

Bước 2. thực hiện vòng lặp

for t = 1 to M do

for i = 1 to k do

chọn mẫu $x_i^{(t)}$ từ phân phối $f\left(x_i \mid x_1^{(t)}, ..., x_{i-1}^{(t)}, x_{i+1}^{(t-1)}, ..., x_k^{(t-1)}\right)$

end for

end for

Bước 3. xuất ra mẫu gồm các giá trị $x^{(1)}, x^{(2)}, ..., x^{(M)}$.

Mẫu thu được từ lấy mẫu Gibbs là một xích Markov mà phân phối ổn định của nó bằng với hàm phân phối xác suất mục tiêu dưới một số điều kiện tổng quát. Do đó mẫu này được xem như xấp xỉ từ hàm phân phối xác suất mục tiêu khi xích Markov hội tụ. Trong lấy mẫu Gibbs, giá trị xuất phát không ảnh hưởng đến sự hội tụ do đó trong thực hành người ta thường bỏ đi một vài giá trị đầu của dãy $x^{(1)}, x^{(2)}, ..., x^{(M)}$ để đảm bảo là mẫu được chọn xấp xỉ tốt nhất từ hàm phân phối xác suất mục tiêu. Các giá trị ban đầu bị bỏ đi được gọi là burn-in.

3. Phương pháp

Để thực hiện các mục tiêu nghiên cứu, chúng tôi tiến hành khảo sát đề thi cuối kì môn Toán Cao Cấp của sinh viên khóa 14, Trường Đại học Kinh Tế-Luật, ĐHQG TPHCM. Đề thi gồm 20 câu hỏi trắc nghiệm khách quan 4 lựa chọn và được hoán vị thành 4 mã đề khác nhau. Mẫu ngẫu nhiên được chọn gồm 388 bài thi của hơn 800 sinh viên tham gia kì thi (chiếm tỉ lệ 46,74%). 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, các trường hợp còn lại (bao gồm việc thí sinh trả lời sai hoặc không chọn bất kì phương án nào hoặc chọn nhiều hơn 1 phương án trả lời) được gán giá trị 0. Kết quả mã hóa dữ liệu được lưu thành dạng ma trận gồm 20 cột ứng với 20 câu hỏi và 388 dòng ứng với 388 thí sinh có bài thi được chọn.

Trước tiên, chúng tôi áp dụng Gibbs Sampler cho mẫu dữ liệu vừa thu thập được. Thuật toán được trình bày như sau:

Bước 1. giá trị xuất phát $x^{(0)} = \left(x_1^{(0)}, x_2^{(0)}, ..., x_{20}^{(0)}\right)$, trong đó $x_1^{(0)}, x_2^{(0)}, ..., x_{20}^{(0)}$ lấy các giá trị tương ứng của thí sinh thứ nhất khi trả lời câu hỏi từ 1 đến 20.

Bước 2. thực hiện vòng lặp

for t = 2 to 388 do

for i = 1 to 20 do

chọn mẫu $x_i^{(t)} = \dfrac{x_1^{(t)} + ... + x_{i-1}^{(t)} + x_{i+1}^{(t-1)} + ... + x_{20}^{(t-1)}}{20}$

end for

end for

Bước 3. xuất ra các mẫu gồm các giá trị $x^{(2)}, x^{(3)}, ..., x^{(388)}$. Nếu kích thước mẫu lớn thì chúng ta có thể bỏ đi các giá trị ban đầu của dãy các giá trị ở trên để mẫu xấp xỉ tốt nhất đối với hàm phân phối mục tiêu.

Để ước lượng các tham số của mô hình Rasch, chúng tôi xét điều kiện được mô tả theo công thức sau:

$$OR\left(AND\left(x_i^t > 0,5, x_i^{*t} = 1\right), AND\left(x_i^t = 0,5, x_i^{*t} = 1\right), AND\left(x_i^t < 0,5, x_i^{*t} = 0\right)\right) \quad (3)$$

Trong công thức (3), toán tử AND có giá trị là TRUE nếu cả 2 điều kiện cùng thỏa mãn và toán tử OR có giá trị là TRUE nếu chỉ cần một trong các điều kiện thỏa mãn. Các trường hợp khác sẽ nhận giá trị là FALSE. Các giá trị x_i^t, x_i^{*t} là giá trị của ứng đáp của thí sinh *t* ứng với câu hỏi *i* sau khi lấy mẫu Gibbs và trước khi lấy mẫu Gibbs. Độ khó của mỗi câu hỏi trong mô hình Rasch khi đó được tính theo công thức

$$b_j = \frac{\sum TRUE}{\sum TRUE + \sum FALSE}, \quad (4)$$

trong đó: $\sum TRUE$ là tổng số các kết quả có giá trị TRUE trong công thức (3) và $\sum FALSE$ là tổng số các kết quả có giá trị FALSE trong công thức (3) của cùng câu hỏi thứ *j*.

Tiếp theo, độ khó của các câu hỏi trong mô hình Rasch được ước lượng bằng phương pháp MML. Cuối cùng, hệ số tương quan Pearson được dùng để đo lường mức độ tương quan của các kết quả đo lường được tính bằng phương pháp MML và bằng Gibbs Sampler.

4. Kết quả

4.1. Ước lượng độ khó của các câu hỏi bằng Gibbs Sampler

Áp dụng thuật toán trình bày trong mục 3 chúng tôi thu được kết quả dưới đây.

Bảng 1. *Ước lượng độ khó của các câu hỏi bằng Gibbs Sampler*

	Độ khó câu hỏi
Item1	0.581395
Item2	0.788114
Item3	0.770026
Item4	0.775194
Item5	0.568475
Item6	0.392765
…	

4.2. Ước lượng độ khó của các câu hỏi bằng MML

Để ước lượng độ khó của các câu hỏi trong mô hình Rasch bằng phương pháp MML, chúng tôi sử dụng câu lệnh `rasch()` của phần mềm R (là một phần mềm mã nguồn mở). Chi tiết về các câu lệnh có thể tham khảo trong (Rizopoulos, 2006). Kết quả ước lượng được thể hiện trong Bảng 2.

Bảng 2. *Ước lượng độ khó của các câu hỏi bằng phương pháp MML*

	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 các giá trị của cột `z.vals`, cột cuối cùng chỉ độ khó của các câu hỏi được quy đổi sang dạng chuẩn. Sử dụng câu lệnh `coeff`, chúng tôi thu được độ khó của các câu hỏi ở dạng tỉ lệ phần trăm như trong Bảng 3.

Bảng 3. *Độ khó của các câu hỏi dùng MML*

	Dffclt	P(x=1|z=0)
Item2	-2.21399306	0.9014991
Item3	-2.21372903	0.9014756
Item4	-1.88483056	0.8681650
Item14	-1.60904625	0.8332789
Item12	-1.53722284	0.8230606
Item17	-1.45083709	0.8101272
…		

4.3 So sánh mức độ tương quan

So sánh mức độ tương quan của kết quả tính toán bằng lấy mẫu Gibbs và bằng phương pháp MML, chúng tôi có kết quả ghi trong Bảng 4.

Bảng 4. *Hệ số tương quan Pearson*

	Column 1	***Column 2***
Column 1	1	
Column 2	0.975592	1

Trong bảng 4, cột thứ nhất Column 1 tương ứng với kết quả đo lường độ khó bằng lấy mẫu Gibbs và cột thứ hai Column 2 tương ứng với kết quả đo lường độ khó bằng phương pháp MML. Hệ số tương quan $r = 0.975592$ cho thấy mức độ tương quan tuyến tính cao của 2 kết quả ước lượng. Điều này cho thấy rằng việc đo lường, ước lượng độ khó của các câu hỏi trong mô hình Rasch bằng phương pháp lấy mẫu Gibbs là nhất quán cao với phương pháp MML trước đây.

5. Kết luận

Nghiên cứu đã trình bày một thuật toán ước lượng độ khó của các câu hỏi trong mô hình Rasch bằng cách dùng lấy mẫu Gibbs. Cách tiếp cận của nghiên cứu là mới vì cho đến nay, việc áp dụng thống kê Bayes vào trong đo lường và đánh giá ở Việt Nam chưa được phổ biến.

Thêm nữa, việc thực thi thuật toán được trình bày trong nghiên cứu là khá đơn giản vì chỉ cần dùng phần mềm bảng tính Excel, thay vì phải dùng các phần mềm thống kê chuyên dùng. Do đó, chúng tôi cho rằng thuật toán này có tính ứng dụng cao và phù hợp với đa số giáo viên.

Mức độ tương quan cao của 2 kết quả trình bày trong bài viết này cho thấy độ tin cậy của phương pháp mà chúng tôi trình bày. Do đó thuật toán này đảm bảo được tính chính xác trong thực hành và đo lường, đánh giá trong giáo dục.

Nghiên cứu chỉ dừng lại ở việc đo lường và ước lượng độ khó trong mô hình Rasch do đó việc mở rộng phương pháp ước lượng các tham số cho các mô hình IRT là vấn đề trong những nghiên cứu tiếp theo.

TÀI LIỆU THAM KHẢO

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 nhiều lựa chọn. *Tạp chí Khoa học - Trường Đại học Sư phạm TPHCM,* 7(85), 174-184.

Nguyễn Thị Hồng Minh & Nguyễn Đức Thiện (2004). Đo lường vàđánh giátrong thi trắc nghiệm khách quan: Độ khó câu hỏi và khả năng của thí sinh. *Tạp chí Khoa học - Đại học Quốc gia Hà Nội*, 197-214.

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.

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.

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

Bechger, T. M., Maris, G., Verstralen, H. H. F. M., & Beguin, A. A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement*, 27(5), 319–334.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4), 443-459.

Camilli, G. & Shepard, L. A. (1994). *Methods of identifying biased test items*. Thousand Oaks, CA: Sage.

Doucet, A., & Wang, X. (2005). Monte Carlo methods for signal processing. *IEEE Signal Process. Mag.*, 22(6), 152-170.

Gelfand, A.E., & Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. USA: Kluwer-Nijhoff Publishing.

Liu, J. S. (2004). *Monter Carlo strategies in scientific computing*. Berlin, Germany: Springer-Verlage.

Liang, F., Liu, C., & Caroll, R. (2010). *Advanced Markov Chain Monter Carlo methods: learning from past sample*. London, U.K.: Wiley Series in Comput., Statist.

Lord, F. M. (1952). A theory of test scores. Psychometric Monographs, 7. Richmond, VA: Psychometric Coporation. Retrieved from http://www.psychometrika.org/journal/online/MN07.pdf

Lord, F. M. & Novick, M. R. (1968). *Statistical theory of mental test scores*. Reading, MA: Addition-Wesley.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York: Springer.

Martino, L., Read, J., & Luengo, D. (2015). Independent doubly adaptive rejection Metropolis sampling within Gibbs sampling. *IEEE Transactions on Signal Processing*, 63(12), 3123-3138.

Morales, R. A. (2009). Evaluation of mathematics achievement test: A comparison between CTT and IRT. *The International Journal of Educational and Psychological Assessment*, 1(1), 19-26.

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

Rizopoulos, D. (2006), ltm: An R package for latent variable modeling and item response theory analysis, *Journal of Statistical software*, 17, 1-25.

Rober, C.P., & Casella, G. (2004). *Monte Carlo statistical methods*. Berlin, Germany: Springer-Verlag.

Ruanaidh, K. O., & Fitzgeral, W. J.. (1996). *Numerical Bayesian methods applied to signal processing*. Berlin, Germany: Springer-Verlag.

Thurstone, L. L. (1925). A method of scaling psychological and education test. *Journal of Education Psychology*, 16, 433-451.

Wright, B. D. & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

PHỤ LỤC

Độ khó của các câu hỏi ước lượng bằng lấy mẫu Gibbs và MML

	Gibbs Sampler	MML
Item1	0.581395349	0.687496
Item2	0.788113695	0.901499
Item3	0.77002584	0.901476
Item4	0.775193798	0.868165
Item5	0.568475452	0.589576
Item6	0.392764858	0.296839
Item7	0.470284238	0.378976
Item8	0.493540052	0.52212
Item9	0.5374677	0.52801
Item10	0.617571059	0.589572
Item11	0.50129199	0.495647
Item12	0.772609819	0.823061
Item13	0.472868217	0.390505
Item14	0.788113695	0.833279
Item15	0.434108527	0.384727
Item16	0.620155039	0.630265
Item17	0.772609819	0.810127
Item18	0.645994832	0.667595
Item19	0.645994832	0.641806
Item20	0.552971576	0.519191

Nguồn: Kết quả nghiên cứu

Số lượng chi tiết được lưu tại địa chỉ:

https://drive.google.com/drive/folders/0B0lrGQ4YEF3PYkxoaUNfMlRKcmM?usp=sharing