



Original Article

Developing Computerized Adaptive Testing: An Experimental Research on Assessing the Mathematical Ability of 10th Graders

Le Thai Hung¹, Tang Thi Thuy¹, Tran Lan Anh¹, Nguyen Tien Dung²,
Nguyen Phuong Anh², Nguyen Thi Quynh Giang^{3,*}

¹*Faculty of Quality Management, VNU University of Education,
144 Xuan Thuy, Cau Giay, Hanoi, Vietnam*

²*High school of Education Sciences, VNU University of Education,
144 Xuan Thuy, Cau Giay, Hanoi, Vietnam*

³*Viettel Digital Service Corporation, 01 Giang Van Minh, Kim Ma, Ba Dinh, Hanoi, Vietnam*

Received 23 September 2019

Revised 12 October 2019; Accepted 28 October 2019

Abstract: Computerized Adaptive Testing (CAT) is a form of assessment test which requires fewer test questions to arrive at precise measurements of examinees' ability. One of the core technical components in building a CAT is mathematical algorithms which estimate examinees' ability and select the most appropriate test questions for the estimation. Mathematical algorithms serve as a locomotive in operating the system of adaptive multiple-choice questions on computers. This research aims to develop essential mathematical algorithms for a computerized system of adaptive multiple-choice tests. A question bank of 500 multiple-choice questions standardized by IRT theory with the difficulty level following the normal distribution satisfying Kolmogorov-Smirnov test, to measure the mathematical ability of 10th graders is also built. The experimenting of the question bank shows that it satisfies the requirements of a psychometric model and the constructed mathematical algorithms meet the criteria for applying in computerized adaptive testing.

Keywords: Computerized Adaptive Testing, ability measurement, mathematical ability, IRT.

* Corresponding author.

E-mail address: qgiang.nguyen@gmail.com

<https://doi.org/10.25073/2588-1159/vnuer.4301>



Phát triển hệ thống trắc nghiệm thích ứng trên máy tính: Nghiên cứu thử nghiệm đánh giá năng lực toán học của học sinh lớp 10

Lê Thái Hưng¹, Tăng Thị Thuỳ¹, Trần Lan Anh¹, Nguyễn Tiến Dũng²,
Nguyễn Phương Anh², Nguyễn Thị Quỳnh Giang^{3,*}

¹*Khoa Quản trị Chất lượng, Trường Đại học Giáo dục, Đại học Quốc gia Hà Nội,
144 Xuân Thủy, Cầu Giấy, Hà Nội, Việt Nam*

²*Trường Trung học phổ thông Khoa học Giáo dục, Trường Đại học Giáo dục,
Đại học Quốc gia Hà Nội, 144 Xuân Thủy, Cầu Giấy, Hà Nội, Việt Nam*

³*Tổng công ty Dịch vụ số Viettel, Số 1 Giang Văn Minh, Kim Mã, Ba Đình, Hà Nội, Việt Nam*

Nhận ngày 23 tháng 9 năm 2019

Chỉnh sửa ngày 12 tháng 10 năm 2019; Chấp nhận đăng ngày 28 tháng 10 năm 2019

Tóm tắt: Trắc nghiệm thích ứng trên máy tính (Computerized Adaptive Testing - CAT) là một hình thức kiểm tra đánh giá cho phép rút ngắn số lượng câu hỏi nhưng vẫn đảm bảo độ chính xác về đánh giá năng lực của thí sinh. Một trong những phần cốt lõi của hệ thống trắc nghiệm thích nghi là các thuật toán ước lượng năng lực thí sinh và lựa chọn câu hỏi. Các thuật toán này đóng vai trò quan trọng trong quá trình vận hành hệ thống trắc nghiệm thích nghi trên máy tính. Nghiên cứu này sẽ phát triển các thuật toán cốt lõi trong hệ thống trắc nghiệm thích nghi từ đó lập trình hệ thống trắc nghiệm thích ứng. Nghiên cứu cũng tiến hành xây dựng ngân hàng gồm 500 câu hỏi trắc nghiệm thích ứng được chuẩn hoá theo lý thuyết IRT với điều kiện độ khó tuân theo phân phối chuẩn thoả mãn kiểm định Kolmogorov-Smirnov, để đánh giá năng lực toán học của học sinh lớp 10. Kết quả vận hành thử nghiệm với hệ thống ngân hàng câu hỏi bước đầu cho thấy: bộ câu hỏi xây dựng đã đáp ứng yêu cầu mô hình ước lượng năng lực và thuật toán cốt lõi đáp ứng được yêu cầu của trắc nghiệm thích ứng.

Từ khóa: Trắc nghiệm thích ứng trên máy tính, đánh giá năng lực, năng lực toán học, IRT.

1. Mở đầu

Trắc nghiệm thích ứng trên máy tính (CAT) được phát triển vào năm 1960 sau khi có sự phát triển mô hình Rasch và lý thuyết ứng đáp

câu hỏi [1], và được thử nghiệm đầu tiên bởi ASVAB (Armed Services Vocational Aptitude Battery) với bài kiểm tra thích ứng về năng lực cá nhân. Từ năm 1979 - 1996, Trung tâm Nghiên cứu và Phát triển Nhân lực Hải quân (NPRDC) đã phát triển, triển khai mô hình đánh giá CAT-ASVAB trong việc xây dựng ngân hàng trắc nghiệm chuẩn hóa quy mô lớn với người được tuyển dụng vào quân đội. Một số

* Tác giả liên hệ.

Địa chỉ email: qgiang.nguyen@gmail.com

<https://doi.org/10.25073/2588-1159/vnuer.4301>

nhà khoa học như Anh, Reckase [2]; Bejar và Weiss [3] ... đã nghiên cứu rất nhiều các báo cáo về mô hình đánh giá CAT và mô hình đánh giá truyền thống trên giấy là tương đương nhau về kết quả phân tích. Trên thế giới, đã có nhiều nghiên cứu cho thấy hiệu quả trong việc sử dụng CAT trong đánh giá người học. Một số kì thi sử dụng CAT: trắc nghiệm thích ứng toán học (CAT-Math) và đọc (CAT-Reading); CAT đánh giá độ thành thạo ngôn ngữ: tiếng Tây Ban Nha, tiếng Ả Rập và Trung Quốc thực hiện bởi Kenyon và Malabonga (2001) [4]; CAT English thử nghiệm tại Thái Lan với đa số sinh viên của họ đều hài lòng với bài kiểm tra và tỏ ra thích thú với hình thức CAT.

Một trong những ưu thế của CAT là chúng ta có thể tổ chức đánh giá đồng thời trên diện rộng với số lượng lớn học sinh tham gia làm bài trên hệ thống máy tính được kết nối mạng. Hơn nữa, CAT cho phép phân tích các chỉ số về năng lực của thí sinh ngay sau khi thí sinh trả lời câu hỏi và thông tin về năng lực của học sinh được cập nhật thường xuyên trong quá trình làm bài cho đến khi đo được năng lực thực sự của họ. CAT không những giúp đánh giá chính xác năng lực mà còn đưa ra thông tin đầy đủ và toàn diện về năng lực học sinh tại từng thời điểm đánh giá. Kết quả này là cơ sở quan trọng triển khai các mô hình học tập thích ứng, đây cũng là một trong những xu thế được quan tâm của giáo dục hiện nay. Kết quả đánh giá thích ứng cùng từng học sinh sẽ được cung cấp thông tin cho người dạy trong suốt quá trình học tập để người dạy có thể đưa ra những quyết định phù hợp. Hệ thống trắc nghiệm thích ứng sẽ giúp người học chủ động tham gia đánh giá và nhận được kết quả tại từng thời điểm để có chiến lược học tập phù hợp. Tại Việt Nam, nghiên cứu về CAT chưa phổ biến, các công trình chủ yếu nghiên cứu về cơ sở lý luận và đưa ra khung lý thuyết xây dựng trắc nghiệm thích ứng trên máy tính mà chưa xây dựng được cơ sở dữ liệu để dùng CAT đánh giá người học. Vì vậy, CAT cũng chưa được đưa ra để sử dụng rộng rãi trong đánh giá năng lực người học. Bài báo này tập trung trình bày kết quả nghiên cứu phát triển hệ thống đánh giá thích ứng từ việc lựa chọn và phát triển thuật toán, xây dựng hệ

thống website, xây dựng ngân hàng 500 câu hỏi trắc nghiệm thích ứng môn Toán lớp 10 và thực nghiệm để kiểm nghiệm lại mô hình thuật toán.

2. Nguyên lý của đánh giá thích ứng

Trắc nghiệm thích ứng, tiếng Anh gọi là “Adaptive Test” là thuật ngữ để chỉ một phương pháp đánh giá thí sinh (học sinh, sinh viên, bệnh nhân, ...) bằng hình thức kiểm tra trắc nghiệm với mục đích đánh giá theo hướng năng lực thông qua bộ câu hỏi tương ứng với mức năng lực của thí sinh. Hệ thống Trắc nghiệm thích ứng là một hệ thống phần mềm được phát triển trên cơ sở mô hình Trắc nghiệm thích ứng để đánh giá thí sinh. Về hoạt động, ta có thể hình dung hệ thống Trắc nghiệm thích ứng cố gắng bắt chước phương pháp đánh giá của một người giáo viên đối với học sinh. Cụ thể, lần đầu tiên hệ thống mặc định năng lực học sinh ở chuẩn trung bình và cung cấp cho thí sinh một câu hỏi khó trung bình. Nếu thí sinh trả lời câu trả lời một cách chính xác, thì sau đó một câu hỏi khó hơn sẽ được đề nghị và nếu không một câu hỏi có độ khó thấp hơn được đề nghị. Quá trình này nên được lặp đi lặp lại cho đến khi có đủ bằng chứng để xác định trình độ kiến thức của thí sinh. Trong Trắc nghiệm thích ứng, quá trình này được thực hiện một cách tự động. Ban đầu có thể tạm thời ước lượng một mức năng lực của thí sinh, sau khi đặt ra một câu hỏi và thí sinh trả lời. Một ước lượng mới về năng lực của thí sinh sẽ được tính toán lại. Với ước tính này, câu hỏi kế tiếp sẽ được chọn một cách chính xác hơn. Ta có thể xem Trắc nghiệm thích ứng như là một thuật toán lặp với thông số đầu vào là ước tính ban đầu về mức độ năng lực của thí sinh. Quy trình để triển khai trắc nghiệm thích ứng được Nathan A. Thompson đưa ra như sau [5]:

Giai đoạn 1. Phát triển ngân hàng câu hỏi và hệ thống trắc nghiệm thích ứng

Giai đoạn 2. Thực hiện đánh giá theo các bước:

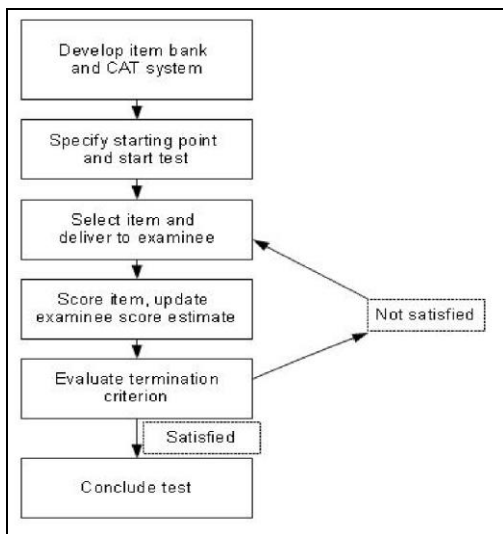
Bước 1: Lựa chọn câu hỏi đầu tiên và bắt đầu quá trình đánh giá;

Bước 2. Câu hỏi tiếp theo phù hợp với năng lực hiện tại của thí sinh được đưa ra và thí sinh trả lời câu hỏi đó;

Bước 3. Theo kết quả câu trả lời của thí sinh, một ước lượng mới của mức độ năng lực được tính toán;

Bước 4. Quay lại Bước 1 nếu các điều kiện dừng của Trắc nghiệm thích ứng chưa thỏa mãn.

Bước 5. Kết thúc quá trình đánh giá nếu điều kiện dừng của Trắc nghiệm thích ứng thỏa mãn.



Hình 1. Quy trình triển khai CAT.

Do vậy, trong mô hình Trắc nghiệm thích ứng: thuật toán lựa chọn câu hỏi tiếp theo phù

Ở đó P_i là xác suất trả lời đúng câu hỏi thứ i và được cho bởi công thức sau theo lý thuyết IRT [6]:

$$P_i(\theta) = P(X_i = 1 / \theta, a_i, b_i, c_i) = c_i + (1 - c_i) \frac{\exp(a_i(\theta - b_i))}{1 + \exp(a_i(\theta - b_i))} \quad (2)$$

Các ước lượng Bayes xem các tham số cần ước lượng là một biến ngẫu nhiên chứ không phải một hằng số. Tham số cần ước lượng sẽ được gắn với một phân bố ban đầu. Ước lượng kiểu Bayes có thể sử dụng khi mà dữ liệu ít và sau đó ước lượng sẽ được cải thiện tốt hơn nếu có thêm dữ liệu. Trong suy diễn Bayes, ban đầu ta giả sử rằng θ tuân theo một phân bố gọi là phân bố tiên nghiệm (prior distribution) $f(\theta)$.

hợp với khả năng hiện tại của thí sinh là khó khăn vì phải được tính toán một cách tối ưu nhất. Bên cạnh đó một ngân hàng câu hỏi được chuẩn hoá theo lý thuyết ứng đáp câu hỏi cần được xây dựng. Số câu hỏi trong ngân hàng cần đủ lớn để đạt được phân bố chuẩn với tham số độ khó.

3. Xây dựng thuật toán cốt lõi và hệ thống trắc nghiệm thích ứng

Các phương pháp phổ biến ước lượng năng lực θ bao gồm: ước lượng hợp lý cực đại (Maximum-Likelihood), ước lượng hậu nghiệm cực đại (Maximum a posteriori) hoặc ước lượng hậu nghiệm trung bình (Expected a posteriori estimator). Tất cả các phương pháp này được mô tả rõ trong Lord (1986), Mislevy (1986). Loại ước lượng mà được sử dụng trong nghiên cứu này sẽ là ước lượng hậu nghiệm cực đại. Dưới đây sẽ mô tả kỹ ước lượng này và thuật toán để tìm ước lượng này.

3.1. Hàm biến cố hợp lý cực đại và ước lượng hậu nghiệm cực đại

Hàm biến cố hợp lý cực đại ứng với $k-1$ câu hỏi được cho bởi:

$$L\left(\frac{\theta}{X_1}, \dots, X_{k-1}\right) = \prod_{i=1}^{k-1} P_i(\theta)^{x_i} (1 - Q_i(\theta))^{1-x_i} \quad (1)$$

Sau đó dựa vào phân bố tiên nghiệm và hàm hợp lý cực đại ta suy ra được phân bố hậu nghiệm của tham số θ :

$$g(\theta / X_1, \dots, X_{k-1}) = \frac{f(\theta)L(\theta / X_1, \dots, X_{k-1})}{\int f(\theta)L(\theta / X_1, \dots, X_{k-1})d\theta} \quad (3)$$

Trong suy diễn Bayes, ước lượng phân phối hậu nghiệm cực đại (Maximum a

Posteriori Estimator- MAP) được giới thiệu trong IRT trong cuốn sách của Lord (1986) [7] và được cho bởi công thức sau:

$$\hat{\theta}_{MAP} = \operatorname{argmax}_{\theta} g(\theta/X_1, \dots, X_{k-1}) \quad (4)$$

Muốn tìm $\hat{\theta}_{MAP}$ ta chuyển bài toán từ tìm maximum của hàm $\ln(f(\theta)L(\theta/X_1, \dots, X_{k-1}))$. Nếu ta chọn phân bố tiên nghiệm của θ là phân bố chuẩn với trung bình 0, độ lệch chuẩn 1 thì $f(\theta) = \frac{1}{\sqrt{2\pi}} e^{-\theta^2/2}$. Khi đó ta có:

$$l(\theta) := \ln(f(\theta)L(\theta/X_1, \dots, X_{k-1})) = C - \theta^2/2 + \sum_{i=1}^{k-1} X_i \ln(P_i(\theta)) + (1 - X_i) \ln(1 - P_i(\theta)) \quad (5)$$

Ở đó C là hằng số. Để tìm maximum của $l(\theta)$ ta dùng thuật toán “Gradient Descent” được mô tả dưới đây.

3.2. Thuật toán Gradient Descent

Thuật toán Gradient Descent là thuật toán hiệu quả được dùng để tìm các điểm cực trị của hàm số khi mà giải phương trình đạo hàm bằng 0 khá phức tạp (Vũ Hữu Tiệp, 2018) [8]. Giả sử ta muốn tìm cực tiểu của hàm một biến $f(x)$. Ta có thể mô tả sơ lược thuật toán như sau: từ một điểm bất kỳ trên đồ thị x , ta cố gắng di chuyển điểm x về điểm mà tại đó $f(x)$ đạt giá trị cực tiểu, ký hiệu là x^* . Điểm x sẽ di chuyển theo hướng ngược với dấu của đạo hàm trong trường hợp ta muốn tìm cực tiểu. Vì giả sử $f'(x) > 0$, thì x nằm về phía bên phải so với x^* , do đó x phải giảm để tiến tới x^* . Ngược lại nếu $f'(x) < 0$, thì x nằm về phía bên trái so với x^* , do đó x phải tăng để tiến tới x^* . Thì công thức cập nhật điểm x như sau:

$$x_k = x_{k-1} - \gamma * f'(x_{k-1}) \quad (6)$$

Ở đó tại bước đầu tiên $x_0 = x$, γ là tốc độ học (learning rate). Sau một số hữu hạn bước, điểm x sẽ di chuyển về gần điểm x^* . Theo thuật toán trên ta phải tính $l'(\theta)$ tại mỗi bước của thuật toán.

$$\begin{aligned} l'(\theta) &= \sum_{i=1}^{k-1} \frac{P'_i(\theta)}{P_i(\theta)} - (1 - X_i) \frac{P'_i(\theta)}{1 - P_i(\theta)} \\ &= \sum_{i=1}^{k-1} X_i \frac{P'_i(\theta)}{P_i(\theta)} \left(\frac{1}{P_i} + \frac{1}{Q_i} \right) - \frac{P'_i(\theta)}{Q_i(\theta)} = \\ &= \sum_{i=1}^{k-1} X_i - P_i(\theta) \end{aligned} \quad (7)$$

Ta mô tả thuật toán như sau:

Bước 1: chọn một điểm khởi tạo $\theta=0$ và tốc độ học γ .

Bước 2: cập nhật

$$\theta := \theta + \gamma \sum_{i=1}^{k-1} \left(X_i - \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)} \right) \quad (8)$$

3.3. Thuật toán tìm kiếm nhị phân (Binary search) để tìm câu hỏi tiếp theo

Tiêu chí chọn câu hỏi. Sau khi thí sinh trả lời câu hỏi $k-1$ thì năng lực tạm thời của thí sinh được ước lượng và ký hiệu là $\hat{\theta}_{k-1}$. Tiếp theo ta phải tìm câu hỏi thứ k phù hợp với mức năng lực này bằng phương pháp lựa chọn câu hỏi theo tiêu chuẩn thông tin tối đa (Maximum-Information Criterion) (Van der Linden and Glas (2010)):

$$i_k = \operatorname{argmax}_j \{ I_{x_j}(\hat{\theta}_{k-1}), j \in R_k \} \quad (9)$$

ở đó $I(\theta)$ là hàm thông tin Fisher:

$$I(\theta) = \frac{(P(\theta))^2}{P(\theta)(1 - P(\theta))} \quad (10)$$

Khi θ cố định, hàm thông tin Fisher đạt giá trị cực đại tại điểm $b = \theta$. Vì vậy câu hỏi i_k được chọn là câu hỏi có độ khó gần với $\hat{\theta}_{k-1}$. Hàm thông tin đạt giá trị maximum khi: Giá trị độ khó b gần bằng giá trị năng lực θ và độ phân biệt a càng lớn. Câu hỏi thứ k được chọn là câu hỏi có độ khó gần bằng năng lực ước lượng $\hat{\theta}_{k-1}$ và có độ phân biệt lớn nhất.

Thuật toán tìm kiếm nhị phân. Cho trước một giá trị a , tìm trong tập hợp b_1, b_2, \dots, b_n giá trị gần nhất với giá trị a . Cách làm đơn giản nhất là ta tính sai số giữa a với tất cả các giá trị trong tập hợp, sau đó tìm sai số nhỏ nhất. Cách làm này rất mất thời gian vì độ phức tạp của thuật toán là $O(n)$. Thuật toán tìm kiếm nhị phân cho phép ta tìm kiếm một cách nhanh hơn. Ta chỉ cần so sánh giá trị a với giá trị trung tâm (là giá trị nằm ở vị trí giữa) trong dãy. Nếu a nhỏ hơn giá trị trung tâm thì ta tìm kiếm trong nửa trái của dãy, nếu a lớn hơn giá trị trung tâm

thì ta tiếp tục tìm kiếm trong nửa phải của dãy, nếu a xấp xỉ giá trị trung tâm thì ta lấy luôn giá trị trung tâm. Độ phức tạp của thuật toán này là $O(\log n)$. Ta có giả thuật như sau:

```

Input:  $a, b_1, b_2, \dots, b_n$ 
Output:  $b$  (giá trị xấp xỉ giá trị  $a$ )
 $L := 0$ 
 $R := n$ 
Do while ( $L < R$ )
   $m := \lfloor (L+R)/2 \rfloor$ 
  if ( $a = b_m$ ) then
     $b = b_m$ 
  else
    if ( $a < b_m$ ) then
       $R := m - 1$ 
    else
       $L := m + 1$ 
  endif
endif
enddo

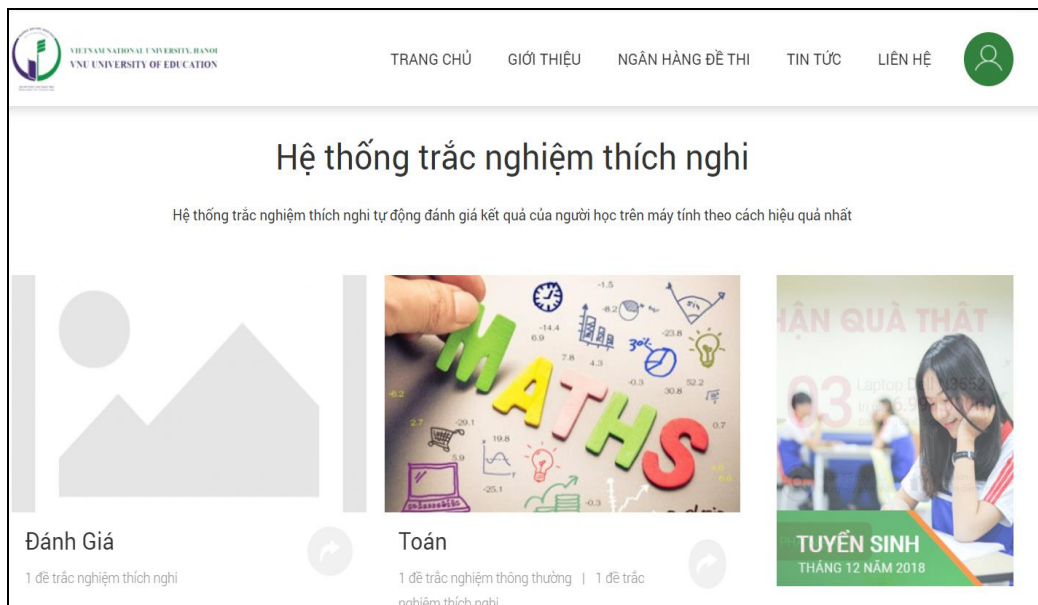
```

Sau khi các thuật toán được xây dựng, các thuật toán được viết bằng ngôn ngữ lập trình web php kết hợp với phần code giao diện web để tạo nên website “trắc nghiệm thích ứng...” Các chức năng chính của website (<http://cat.education.vnu.edu.vn/>) bao gồm:

- Quản lý đăng nhập: Thí sinh chưa có tài khoản thì có thể đăng ký mới để đăng nhập vào hệ thống đánh giá trắc nghiệm thích ứng, hệ thống sẽ khởi tạo thí sinh mới có mức năng lực trung bình. Đối với thí sinh cũ có thể đăng nhập tài khoản đã có.

- Quản trị nội dung website: Phần này cho phép giáo viên quản lý học sinh, quản lý môn học như thêm mới hoặc xóa nội dung môn học, quản lý câu hỏi, quản lý liên hệ... Chi tiết có trong phụ lục hướng dẫn sử dụng web đính kèm.

3.4. Xây dựng website trắc nghiệm thích ứng



Hình 2. Giao diện hệ thống CAT phát triển trên web.

- Giao diện thực hiện quá trình kiểm tra: Đây là nơi thí sinh sẽ thực hiện thao tác trả lời câu hỏi của hệ thống trắc nghiệm thích ứng. Mỗi lượt sẽ xuất hiện câu hỏi và các đáp án lựa chọn. Thí sinh chọn đáp án và gửi về hệ thống.

Hệ thống đánh giá và hiển thị câu hỏi tiếp theo cho thí sinh. Hệ thống sẽ tự động dừng khi đủ cơ sở đánh giá thí sinh. Toàn bộ bài làm của thí sinh được hiển thị: các câu hỏi, độ khó của từng câu, câu trả lời của thí sinh là đúng hay sai, lĩnh

vực của câu hỏi và điểm số của thí sinh quy đổi sang thang điểm 100.

4. Phát triển ngân hàng câu hỏi đánh giá năng lực toán học

Năng lực Toán học là các đặc điểm tâm lý cá nhân (trước hết là các đặc điểm hoạt động trí tuệ) đáp ứng được các yêu cầu của hoạt động toán và tạo điều kiện linh hoạt các kiến thức, kỹ năng, kỹ xảo trong lĩnh vực toán học tương đối nhanh, dễ dàng và sâu sắc trong những điều kiện như nhau. Theo OECD: “**Năng lực Toán học** là khả năng của cá nhân biết lập công thức (formulate), vận dụng (employ) và giải thích (explain) Toán học trong nhiều ngữ cảnh. Nó bao gồm suy luận Toán học và sử dụng các khái niệm, phương pháp, sự việc và công cụ để mô tả, giải thích và dự đoán các hiện tượng. Nó giúp cho con người nhận ra vai trò của Toán học trên thế giới và đưa ra phán đoán và quyết định của công dân biết góp ý, tham gia và suy ngẫm” [9].

Theo Korutecxki thì **cấu trúc của năng lực Toán học** bao gồm:

Về mặt thu nhận thông tin: Năng lực tri giác hình thức hóa tài liệu toán học, năng lực nắm cấu trúc hình thức của bài toán.

Về mặt chế biến thông tin, đó là:

- Năng lực tư duy logic trong phạm vi các quan hệ số lượng và các quan hệ không gian, các ký hiệu, năng lực suy nghĩ với các ký hiệu toán học.

- Năng lực khái quát hóa nhanh chóng và rộng rãi các đối tượng, quan hệ, các phép toán của toán học. Năng lực rút ngắn quá trình suy luận toán học và hệ thống các phép toán tương ứng, năng lực suy nghĩ với cấu trúc được rút gọn.

- Tính mềm dẻo của quá trình tư duy trong hoạt động toán học.

- Khuynh hướng đạt tới sự rõ ràng, sự đơn giản, tính tiết kiệm và tính hợp lý của lời giải.

- Năng lực thay đổi nhanh chóng và dễ dàng hướng suy nghĩ, dạng tư duy thuận chuyển qua tư duy nghịch.

Về mặt lưu trữ các thông tin, đó là trí nhớ toán học tức là trí nhớ khái quát về các quan hệ toán học, về các đặc điểm điển hình, các sơ đồ suy luận và chứng minh, về các phương pháp giải toán và các nguyên tắc xem xét các bài toán ấy.

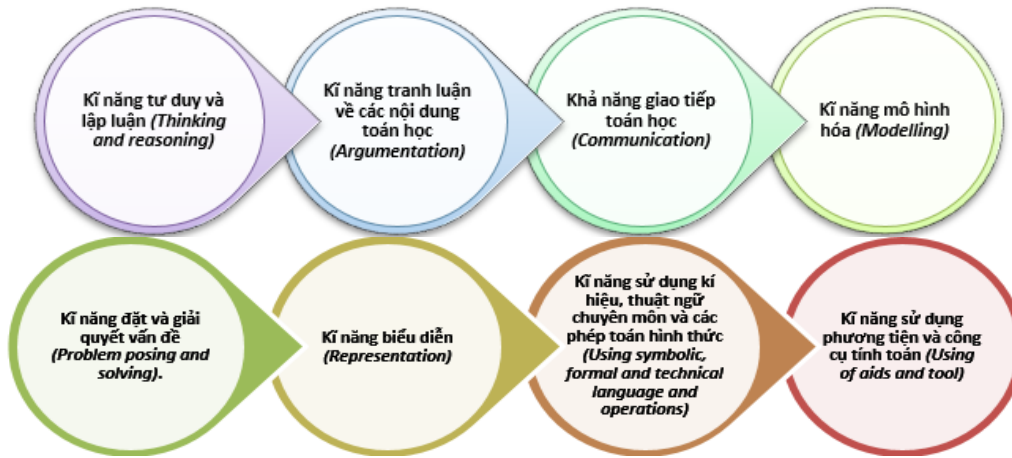
Về thành phần tổng hợp chung, đó là khuynh hướng toán học của trí tuệ. Tuy nhiên, cần chú ý rằng tốc độ tư duy, năng lực tính toán, trí nhớ về các công thức, ... không nhất thiết phải có mặt trong các thành phần của năng lực toán học.

Các cấp độ năng lực Toán học. PISA đề cập đến 3 cấp độ năng lực Toán học phổ thông khác với đánh giá truyền thống, đòi hỏi không chỉ chú ý đến nội dung kiến thức người học đã tiếp thu được, mà còn chú trọng đánh giá những năng lực, quá trình hình thành các kỹ năng (processes skills).

Bảng 1. Các cấp độ năng lực Toán học của PISA

Cấp độ	Đặc điểm
Cấp độ 1 Ghi nhớ, tái hiện	- Nhớ lại các đối tượng, khái niệm, định nghĩa và tính chất toán học. - Thực hiện được một cách làm quen thuộc. - Áp dụng một thuật toán tiêu chuẩn.
Cấp độ 2 Kết nối, tích hợp	- Kết nối, tích hợp thông tin để giải quyết các vấn đề đơn giản. - Tạo những kết nối trong các cách biểu đạt khác nhau. - Đọc và giải thích được các ký hiệu và ngôn ngữ hình thức (toán học) và hiểu mối quan hệ của chúng với ngôn ngữ tự nhiên.
Cấp độ 3 Khái quát hóa, Toán học hóa	- Nhận biết nội dung toán học trong tình huống có vấn đề phải giải quyết. - Vận dụng kiến thức toán học để giải quyết các vấn đề thực tiễn. - Biết phân tích, tổng hợp, suy luận, lập luận, khái quát hóa trong chứng minh toán học.

Khung đánh giá của PISA đối với lĩnh vực Toán học tập trung vào 8 kỹ năng quan trọng dưới đây:



Hình 3. Tám kỹ năng toán học theo PISA.

Phát triển ngân hàng câu hỏi trắc nghiệm thích ứng môn Toán 10. Trên cơ sở về thang đánh giá năng lực toán học, nhóm nghiên cứu tiến hành phát triển ngân hàng câu hỏi theo quy trình khoa học (Hình 4) với một nhóm giáo viên đang trực tiếp dạy môn Toán tại trường THPT. Nghiên cứu này đặc biệt quan tâm đến việc tập huấn kĩ thuật cho giáo viên tham gia nghiên cứu về thiết kế bảng đặc tả đánh giá năng lực toán học, viết và hiệu chỉnh câu hỏi sau khi thẩm định kĩ thuật. Bảng đặc tả đề thi gồm 50 câu hỏi dạng thức trắc nghiệm khách quan, với các chủ đề bao phủ chương trình Đại số lớp 10 chương trình giáo dục phổ thông hiện hành, được thiết kế bám sát theo ba cấp độ đánh giá năng lực toán học theo PISA tương ứng với nhận biết (**K**nowledge) - Áp dụng (**A**rgument) - lập luận (**R**easoning) với 3 mức độ khó của câu hỏi (dễ, trung bình và khó). Tuy nhiên với đặc thù của Toán học phổ thông hiện hành, các bài toán thường bắt đầu ở mức Áp dụng đơn giản, rất ít khi sử dụng mức độ nhận biết lý thuyết thuần túy. Bảng 2 dưới đây sẽ tóm lược bảng đặc tả chi tiết đề thi, các câu hỏi được sắp xếp dựa trên kết quả thảo luận nhóm của tập thể giáo viên tham gia nghiên cứu này (ý kiến chuyên gia) vì vậy sẽ có sự thay đổi sau thử nghiệm.

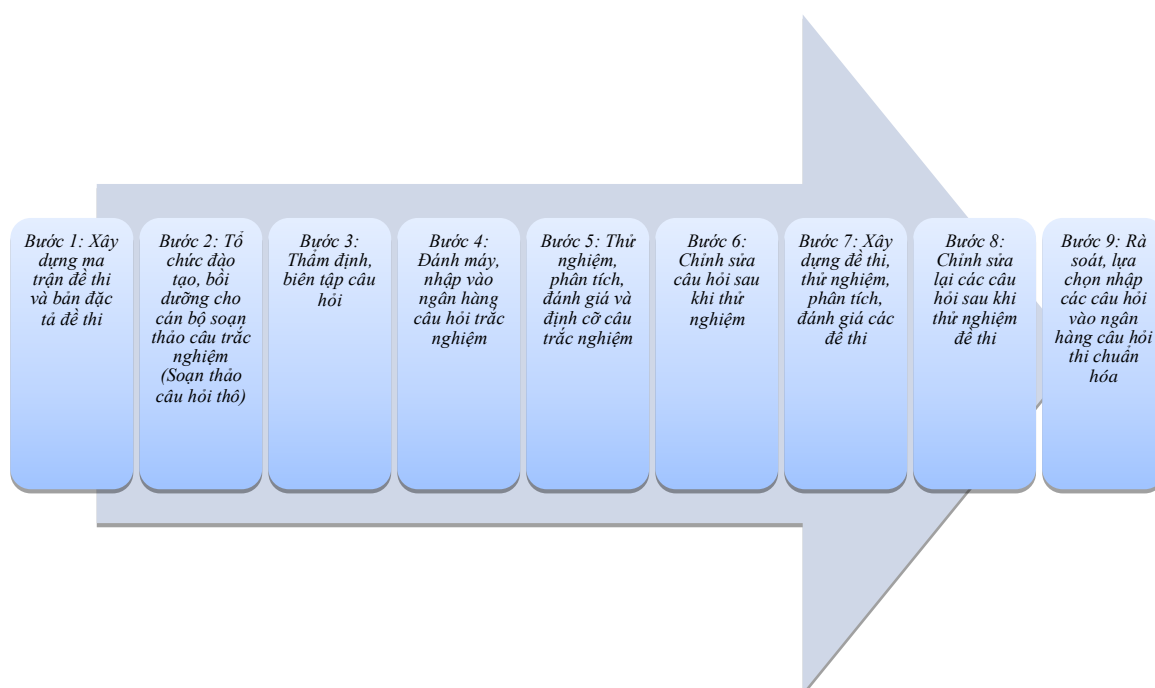
Tiếp theo đó các giáo viên tham gia nghiên cứu này đã được tập huấn kĩ thuật viết câu hỏi theo bảng đặc tả tiến hành viết câu hỏi và đọc kĩ

thuật trước khi thử nghiệm. Mỗi đề thử nghiệm chúng tôi tiến hành tối thiểu trên 150 học sinh để đảm bảo cỡ chuẩn hoá câu hỏi trắc nghiệm thích ứng. Những câu hỏi đáp ứng theo lý thuyết ứng đáp câu hỏi (IRT) sẽ được giữ lại và lưu vào ngân hàng đề. Dựa theo các nghiên cứu trước đây [10] về cơ của ngân hàng trắc nghiệm CAT, chúng tôi sẽ phát triển 500 câu hỏi trắc nghiệm thích ứng đánh giá năng lực toán học của học sinh lớp 10. Dưới đây chúng tôi trình bày kết quả phân tích 01 đề thử nghiệm theo lý thuyết IRT sử dụng phần mềm IATA.

Dưới đây là kết quả chạy sự phù hợp với mô hình IRT, các tham số độ khó (b), độ phân biệt (a), độ phân biệt theo lý thuyết khảo thí cổ điển (Discr) thông qua tham số MNSQ, theo lý thuyết IRT giá trị này phải nằm trong khoảng

$$\left[1 - 2\sqrt{\frac{2}{N}}, 1 + 2\sqrt{\frac{2}{N}}\right]$$

tương ứng với giá trị CI, N là số thí sinh tham gia trả lời câu hỏi. Dữ liệu gồm 247 học sinh tham gia sau khi chúng tôi loại bỏ các trường hợp thí sinh bỏ trống quá nhiều trong 254 thí sinh ban đầu. Trong kết quả chạy IATA, các câu hỏi có kết quả chạy: \blacktriangle là không phù hợp, \blacklozenge phù hợp và \bullet phù hợp tốt với lý thuyết IRT. Kết quả cho thấy có 14 câu (\blacktriangle) có giá trị không thoả mãn điều kiện cần được điều chỉnh hoặc loại bỏ.



Hình 4. Quy trình phát triển ngân hàng đề thi CAT.

Bảng 2. Bảng đặc tả rút gọn bài thi

Nội dung (%)	Dễ		Trung bình			Khó	
	A	R	K	A	R	A	R
1. Mệnh đề, tập hợp 5%	C1			C2			
2. Hàm số 22%	C3	C4		C5			
	C6	C7		C8			C9
		C10	C11	C12		C13	
3. Phương trình 23%	C14,C15			C16			
	C17	C18		C19	C20		
	C21	C22		C23	C24		C25
4. Bất đẳng thức bất phương trình 25%		C26		C27			
	C28	C29		C30			
	C31			C32			
	C33				C34		
5. Thống kê 5%	C35			C36	C37		C38
	C39						
	C40						
6. Lượng giác 20%				C41			
	C42	C43		C44			
	C45	C46		C47, C48	C49	C50	
Tổng	15	9	1	15	5	2	3

Kết quả cho thấy đã số các câu hỏi này không thỏa mãn cả điều kiện về độ phân biệt theo lý thuyết cổ điển ($Discr > 0.2$) và thuộc nhóm câu hỏi quá dễ (Câu 1, 15, 17, 18, 26: $b < -3$) hoặc quá khó (Câu 10, 11, 20, 21, 31, 33, 44, 46, 47), $b > +3$). Biểu diễn phân tích nhân tố của các câu hỏi này cũng không đáp ứng yêu cầu, có sự chênh lệch lớn giữa đường lý thuyết và thực nghiệm (xem trường hợp với câu hỏi 1). Những câu này cần được điều chỉnh, tuy nhiên trong nghiên cứu này chúng tôi loại bỏ để viết

bảng câu hỏi mới và tiếp tục thử nghiệm cho đến khi ngân hàng câu hỏi đủ lớn và đáp ứng yêu cầu phân phối chuẩn về độ khó. Những điểm cần lưu ý khi phân tích cụ thể các câu hỏi không thỏa mãn được tổng hợp cho giai đoạn tiếp theo nhằm giảm thiểu số lượng câu hỏi không đạt yêu cầu. Các câu hỏi còn lại phù hợp với mô hình, là những câu hỏi thường có độ khó nằm trong khoảng $[-3, +3]$; thỏa mãn điều kiện độ phân biệt theo lý thuyết cổ điển ($Discr > 0,2$) và chất lượng đáp án nhiều tốt.

Bảng 3. Kết quả phân tích sự phù hợp với lý thuyết IRT (items fit) và các tham số (tóm lược)

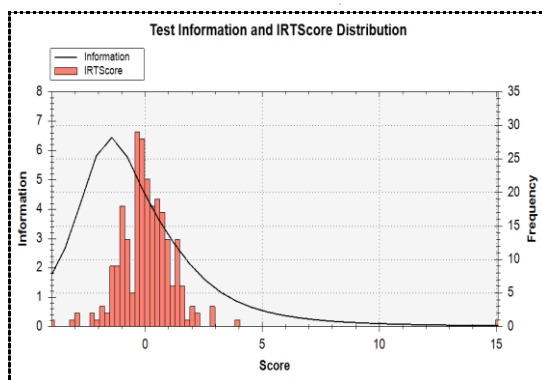
Use	O	Name	Discr	PVal	PBis	a	b	c
<input checked="" type="checkbox"/>	▲	C26	0.01	0.05	0.06	-0.04	-41.19	0.00
<input checked="" type="checkbox"/>	●	C27	0.20	0.92	0.33	1.00	-1.99	0.00
<input checked="" type="checkbox"/>	●	C28	0.42	0.39	0.34	0.44	0.69	0.00
<input checked="" type="checkbox"/>	◆	C29	0.24	0.19	0.22	0.32	2.79	0.00
<input checked="" type="checkbox"/>	●	C30	0.49	0.42	0.34	0.49	0.45	0.00
<input checked="" type="checkbox"/>	▲	C31	0.02	0.30	0.17	0.15	3.37	0.00
<input checked="" type="checkbox"/>	◆	C32	0.22	0.31	0.21	0.19	2.49	0.00
<input checked="" type="checkbox"/>	▲	C33	0.14	0.21	0.16	0.11	7.27	0.00
<input checked="" type="checkbox"/>	●	C34	0.35	0.85	0.46	1.08	-1.41	0.00
<input checked="" type="checkbox"/>	●	C35	0.49	0.74	0.44	0.80	-1.01	0.00
<input checked="" type="checkbox"/>	●	C36	0.40	0.33	0.35	0.52	0.90	0.00
<input checked="" type="checkbox"/>	●	C37	0.43	0.55	0.33	0.44	-0.32	0.00
<input checked="" type="checkbox"/>	◆	C38	0.32	0.59	0.31	0.44	-0.54	0.00
<input checked="" type="checkbox"/>	●	C39	0.44	0.75	0.44	0.87	-1.04	0.00
<input checked="" type="checkbox"/>	●	C40	0.42	0.38	0.33	0.44	0.71	0.00
<input checked="" type="checkbox"/>	◆	C41	0.17	0.71	0.23	0.25	-2.18	0.00
<input checked="" type="checkbox"/>	◆	C42	0.29	0.77	0.32	0.42	-1.87	0.00
<input checked="" type="checkbox"/>	●	C43	0.63	0.42	0.46	0.75	0.32	0.00
<input checked="" type="checkbox"/>	▲	C44	0.17	0.37	0.17	0.09	3.58	0.00
<input checked="" type="checkbox"/>	●	C45	0.37	0.35	0.30	0.44	0.90	0.00
<input checked="" type="checkbox"/>	▲	C46	0.10	0.20	0.17	0.19	4.52	0.00
<input checked="" type="checkbox"/>	▲	C47	-0.01	0.09	0.10	0.07	18.55	0.00
<input checked="" type="checkbox"/>	●	C52	0.39	0.32	0.35	0.53	0.98	0.00
<input checked="" type="checkbox"/>	●	C55	0.38	0.75	0.36	0.57	-1.37	0.00
<input checked="" type="checkbox"/>	●	C57	0.40	0.56	0.26	0.25	-0.59	0.00

Use	O	Name	Discr	PVal	PBis	a	b	c
<input checked="" type="checkbox"/>	▲	C18	-0.08	0.06	-0.06	-0.29	-6.02	0.00
<input checked="" type="checkbox"/>	▲	C15	-0.04	0.12	0.03	-0.08	-13.96	0.00
<input checked="" type="checkbox"/>	▲	C17	-0.03	0.02	-0.02	-0.30	-7.86	0.00
<input checked="" type="checkbox"/>	▲	C47	-0.01	0.09	0.10	0.07	18.55	0.00
<input checked="" type="checkbox"/>	▲	C26	0.01	0.05	0.06	-0.04	-41.19	0.00
<input checked="" type="checkbox"/>	▲	C31	0.02	0.30	0.17	0.15	3.37	0.00
<input checked="" type="checkbox"/>	▲	C10	0.03	0.10	0.12	0.16	8.08	0.00
<input checked="" type="checkbox"/>	▲	C20	0.09	0.19	0.12	0.05	16.04	0.00
<input checked="" type="checkbox"/>	▲	C46	0.10	0.20	0.17	0.19	4.52	0.00
<input checked="" type="checkbox"/>	▲	C11	0.14	0.30	0.10	0.01	40.87	0.00
<input checked="" type="checkbox"/>	▲	C33	0.14	0.21	0.16	0.11	7.27	0.00
<input checked="" type="checkbox"/>	◆	C2	0.15	0.91	0.22	0.53	-2.95	0.00
<input checked="" type="checkbox"/>	▲	C21	0.15	0.16	0.19	0.29	3.55	0.00
<input checked="" type="checkbox"/>	◆	C24	0.17	0.18	0.19	0.32	2.94	0.00
<input checked="" type="checkbox"/>	◆	C41	0.17	0.71	0.23	0.25	-2.18	0.00
<input checked="" type="checkbox"/>	▲	C44	0.17	0.37	0.17	0.09	3.58	0.00
<input checked="" type="checkbox"/>	◆	C5	0.18	0.84	0.19	0.35	-2.97	0.00
<input checked="" type="checkbox"/>	◆	C7	0.19	0.92	0.31	0.79	-2.37	0.00
<input checked="" type="checkbox"/>	▲	C1	0.20	0.88	0.17	0.27	-4.51	0.00

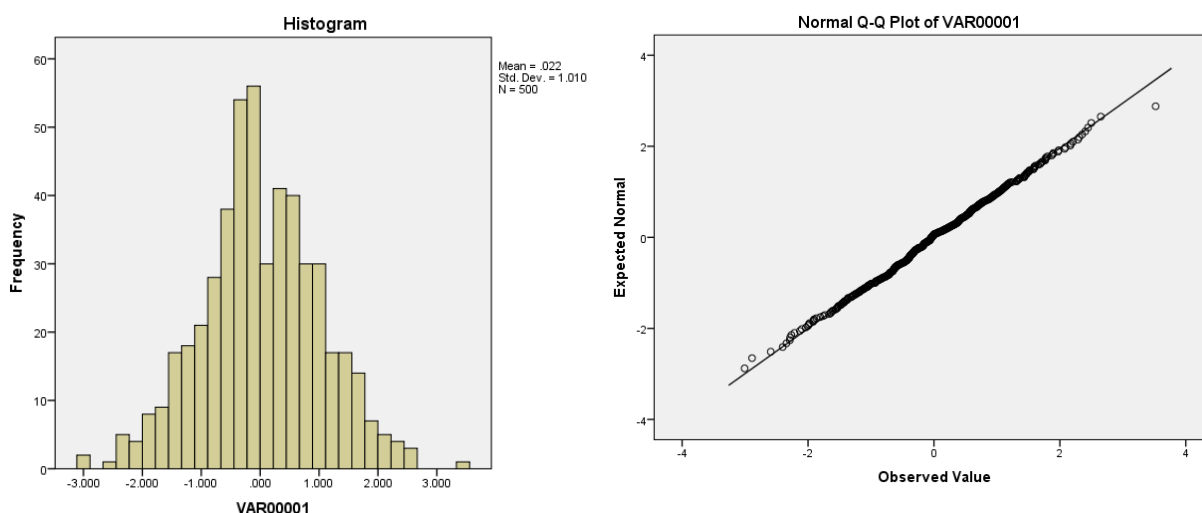
Đề có nhận định chung về đề thi, chúng tôi tiến hành phân tích phổ điểm theo thang năng lực và hàm thông tin của đề thi. Kết quả cho thấy năng lực trung bình của thí sinh là 0,08 gần với mức năng lực trung bình lý thuyết, phân bố năng lực của thí sinh có dáng điệu chuẩn và số câu hỏi đáp ứng mô hình là 36. Hàm thông tin có đỉnh đạt xấp xỉ 6,6, như vậy

độ tin cậy của bài trắc nghiệm sẽ lớn hơn 0,8 đạt yêu cầu (Biểu đồ 1).

Tiếp tục thực hiện quá trình trên cho 11 đề thi, tác giả thu được 500 câu hỏi với tham số độ khó và độ phân biệt theo lý thuyết khảo thí IRT. Tiến hành kiểm nghiệm phân phối chuẩn với kiểm định Kolmogorov-Smirnov thu nhận được kết quả như sau (Biểu đồ 2):



Biểu đồ 1. Phân bố điểm năng lực và hàm thông tin của đề thi theo lý thuyết IRT.



Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	Df	Sig.	Statistic	df	Sig.
VAR00001	.037	500	.119	.998	500	.917

a. Lilliefors Significance Correction

Biểu đồ 2. Kết quả chạy kiểm định Kolmogorov-Smirnov giá trị độ khó của 500 câu hỏi.

Nhìn vào biểu đồ 2, ta thấy dáng điệu của phân bố độ khó và biểu đồ xác suất chuẩn Q-Q (Normal QQ plot) đều thể hiện đặc trưng của phân phối chuẩn (phân bố hình chuông, các trị số quan sát và mong đợi đều nằm trên đường chéo). Kết quả kiểm định phân phối chuẩn (Test of Normality) trường hợp dữ liệu lớn hơn 30 qua hệ số kiểm định Kolmogorov-Smirnov cho thấy hệ số ý nghĩa (Sig) = 0,119 > 0.05 nên

giả thuyết H_0 thoả mãn, nên độ khó của 500 câu hỏi tuân theo phân phối chuẩn.

5. Kết quả vận hành hệ thống trắc nghiệm thích ứng

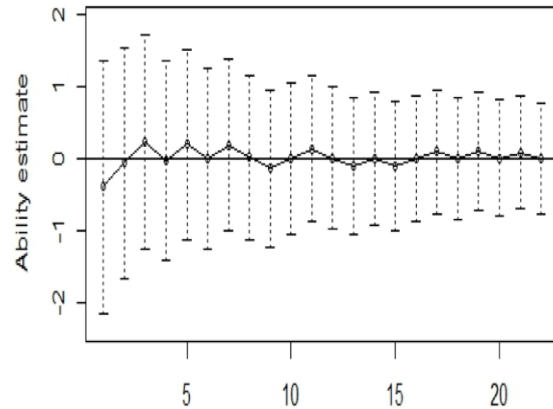
Tác giả tiến hành nhập câu hỏi và cho thí sinh làm bài trên hệ thống để kiểm nghiệm lại tính đúng đắn của thuật toán và ngân hàng câu

hỏi trắc nghiệm thích ứng đánh năng lực Toán học, dưới đây là một số kết quả điển hình. Kết quả thi của thí sinh có năng lực $\theta=1$ cho ta ước lượng năng lực thí sinh sau 22 câu hỏi =1.068. Điểm năng lực này được đổi sang điểm thô bằng cách sử dụng hàm phân phối của phân bố chuẩn. Ta tính:

$$\frac{1}{2\pi} \int_{-\infty}^{1.068} e^{-\frac{x^2}{2}} dx = 0.8572. \quad (11)$$

Điều đó có nghĩa là THÍ SINH làm được 85.72% bài thi tiêu chuẩn. Đổi sang thang điểm 100 ta nhân giá trị này với 100 thì ra điểm của THÍ SINH là 85.72. Nhìn vào hình 1 ta thấy kết quả thi trên hệ thống trắc nghiệm thích ứng hoàn toàn trùng khớp với kết quả mô phỏng bởi phần mềm R được công bố trước đây [11]. Kết quả cho thấy, thí sinh làm đến câu hỏi thứ 22 thoả mãn điều kiện dừng theo mô hình lý CAT

đã được phát triển. Từ đó cho phép ta khẳng định tính chính xác của hệ thống trắc nghiệm thích ứng.



Biểu đồ 3. Mô phỏng CAT với thí sinh có năng lực $\theta=1$ (Giang và cs, 2018).

Thí sinh đạt số điểm: **85.72**

STT	CÂU HỎI	ĐỘ KHÓ	CÂU TRẢ LỜI
1	Câu: 78	0.001105352	1
2	Câu: 192	0.402011779	1
3	Câu: 341	0.739589226	1
4	Câu: 242	1.034107735	0
5	Câu: 22	0.782136301	1
6	Câu: 290	1.005159218	0
7	Câu: 349	0.803283216	1
8	Câu: 215	0.987838267	1
9	Câu: 93	1.160402616	0
10	Câu: 157	1.000028804	0
11	Câu: 444	0.852814994	1
12	Câu: 182	0.98389557	1
13	Câu: 413	1.124772447	0
14	Câu: 417	1.017754227	0
15	Câu: 164	0.886422651	1
16	Câu: 250	1.020463909	0
17	Câu: 301	0.893673702	1
18	Câu: 451	0.951596633	1
19	Câu: 470	1.08342991	0
20	Câu: 179	1.027392439	0
21	Câu: 108	0.910174229	1
22	Câu: 298	0.955136677	1

Hình 1. Kết quả thi của thí sinh có mức năng lực $\theta = 1$.

Để khẳng định mức độ chính xác của hệ thống, ta kiểm tra thêm hai trường hợp: trường hợp 1 thí sinh trả lời đúng tất cả các câu hỏi và trường hợp 2 thí sinh trả lời sai tất cả các câu hỏi. Kết quả bài thi của 2 thí sinh này tương ứng được cho bởi hình 2 và hình 3.

Trường hợp thí sinh trả lời sai tất cả các câu hỏi thì điểm thi cuối cùng của thí sinh là 0.01

trên thang điểm 100. Ngược lại trường hợp thí sinh trả lời đúng tất cả các câu hỏi thì điểm thi cuối cùng của thí sinh là 99.99 trên thang điểm 100. Kết quả cho thấy ứng với mỗi thí sinh có năng lực khác nhau, số lượng câu hỏi cần trả lời trên hệ thống là khác nhau.

Thí sinh đạt số điểm: **0.01**

STT	CÂU HỎI	ĐỘ KHÓ	CÂU TRẢ LỜI
1	Câu: 78	0.001105352	0
2	Câu: 213	-0.400246744	0
3	Câu: 65	-0.743273209	0
4	Câu: 488	-1.043667439	0
5	Câu: 319	-1.294140004	0
6	Câu: 136	-1.536449824	0
7	Câu: 424	-1.759468735	0
8	Câu: 489	-1.972934934	0
9	Câu: 14	-2.214699887	0
10	Câu: 450	-2.34272312	0
11	Câu: 361	-2.59232767	0
12	Câu: 232	-2.888920672	0
13	Câu: 446	-3.008048599	0
14	Câu: 257	-2.403096215	0
15	Câu: 277	-2.28912398	0
16	Câu: 205	-2.285235535	0
17	Câu: 345	-2.264889356	0
18	Câu: 387	-2.129360648	0
19	Câu: 458	-2.09788296	0
20	Câu: 431	-2.033285596	0
21	Câu: 246	-2.000164945	0
22	Câu: 24	-1.989351696	0
23	Câu: 359	-1.91890982	0
24	Câu: 141	-1.914359426	0
25	Câu: 389	-1.904955446	0

Hình 2. Kết quả thi của thí sinh có mức năng lực rất kém.

Thí sinh đạt số điểm: **99.99**

STT	CÂU HỎI	ĐỘ KHÓ	CÂU TRẢ LỜI
1	Câu: 78	0.001105352	1
2	Câu: 192	0.402011779	1
3	Câu: 341	0.739589226	1
4	Câu: 242	1.034107735	1
5	Câu: 321	1.30790152	1
6	Câu: 216	1.519745025	1
7	Câu: 380	1.763552003	1
8	Câu: 303	1.971337386	1
9	Câu: 70	2.17261167	1
10	Câu: 485	2.350554326	1
11	Câu: 206	2.49766159	1
12	Câu: 274	2.649166881	1
13	Câu: 486	2.446531376	1
15	Câu: 495	3.810276681	1
16	Câu: 171	2.307978399	1
17	Câu: 492	2.284659326	1
18	Câu: 166	2.206102465	1
19	Câu: 462	2.165368502	1
20	Câu: 147	2.087166546	1
21	Câu: 178	2.075245009	1
22	Câu: 56	1.980399899	1
23	Câu: 402	1.895654774	1
24	Câu: 350	1.887474463	1
25	Câu: 160	1.869290622	1
26	Câu: 226	1.803141908	1
27	Câu: 295	1.778429287	1

Hình 3. Kết quả thi của thí sinh có mức năng lực rất tốt.

6. Kết luận

Hệ thống trắc nghiệm thích ứng đã được xây dựng và thử nghiệm trên ngân hàng 500 câu hỏi trắc nghiệm đánh giá năng lực Toán học với độ khó đáp ứng phân phối chuẩn trong miền

từ $[-3, +3,5]$. So sánh với kết quả mô phỏng cho thấy kết quả chạy trên hệ thống trắc nghiệm thích ứng hoàn toàn trùng khớp với kết quả mô phỏng với phần mềm R. Từ đó khẳng định tính chuẩn xác của các thuật toán trong hệ thống.

Với một ngân hàng câu hỏi được xây dựng đạt chuẩn chất lượng, hệ thống trắc nghiệm thích ứng cho phép rút ngắn thời gian làm bài của thí sinh thí sinh mà vẫn đạt được độ chuẩn xác. Hệ thống trắc nghiệm thích ứng còn có các tính năng giúp giáo viên có thể dễ dàng quản lý ngân hàng câu hỏi, quản lý thí sinh và thu thập kết quả thi của thí sinh một cách tự động. Từ đó giúp giáo viên dễ dàng theo dõi sự tiến bộ của học sinh và có những trợ giúp kịp thời. Nghiên cứu này tạo tiền đề để phát triển hệ thống đánh giá thích ứng kết hợp dạy học phân hoá cho người học không chỉ ở môn Toán mà các môn học khác nếu các nhóm năng lực hoặc các kỹ năng liên quan được xác định rõ ràng, có thể đánh giá được. Mặt khác để có được hệ thống đánh giá thích ứng tốt, ngoài việc phát triển thuật toán và phần mềm thì phần trọng tâm nhất chính là xây dựng ngân hàng câu hỏi chuẩn hoá theo lý thuyết IRT, đây là công việc đòi hỏi nhiều công sức nhất bởi những giáo viên có kinh nghiệm giảng dạy và am hiểu về kiểm tra đánh giá.

Lời cảm ơn

Kết quả nghiên cứu trình bày ở đây được tài trợ bởi đề tài nghiên cứu khoa học mã số **QS.17.14** và **QS.17.15** của trường Đại học Giáo dục được thực hiện bởi nhóm nghiên cứu của Khoa Quản trị Chất lượng, Trường Đại học Giáo dục.

Tài liệu tham khảo

- [1] Rod Powers, Jennifer Lawler, ASVAB For Dummies, John Wiley & Sons Published house, 2007.
- [2] M.D. Reckase, Item pool design for computerized adaptive tests, Paper presented at annual meeting of the National Council on Measurement in Education, Chicago, IL, 2003.
- [3] D.J. Weiss, G.G. Kingsbury, Application of computerized adaptive testing to educational problems *Journal of Educational Measurement* 21 (1984) 361-375.
- [4] A. Carol, Chapelle, Shannon Sauro, *The Handbook of Technology and Second Language Teaching and Learning*, John Wiley & Sons, 2017.
- [5] Thompson, A. Nathan, Weiss, A. David, *A Framework for the Development of Computerized Adaptive Tests. Practical Assessment, Research & Evaluation*, 16 (1). Available online: <http://pareonline.net/getvn.asp?v=16&n=1/>, 2011.
- [6] Lam Quang Thiep, *Measurement and Evaluation in Education: Theory and Application*, VNU Publishing house, 2011. (in Vietnamese).
- [7] F.M. Lord, Maximum likelihood and Bayesian parameter estimation in item response theory, *Journal of Educational Measurement* 23 (1986) 157-162.
- [8] Vu Huu Tiep, *Basic Machine Learning*, Scientific and Technical Publishing, 2018. (Vietnamese).
- [9] ECD, *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*, OECD Publishing, 2013.
- [10] Alper Şahin, David J. Weiss, Effects of Calibration Sample Size and Item Bank Size on Ability Estimation in Computerized Adaptive Testing, *Educational Sciences: Theory & Practice*, 2015.
- [11] Nguyen Thuy Giang, Le Thai Hung, Simulate an Computerized Adaptive Testing with R, *Vietnam Education Journal* 11 (2018) 6-11. (in Vietnamese).