VJE Tạp chí Giáo dục (2023), 23(19), 8-14 ISSN: 2354-0753

PHÁT TRIỂN NGÂN HÀNG CÂU HỎI TRẮC NGHIỆM THÍCH ỨNG ĐÁNH GIÁ TỪ VỰNG TIẾNG ANH THÔNG DỤNG: ÁP DỤNG IRT VÀ PHƯƠNG PHÁP CÂN BẰNG ĐỀ

Nguyễn Thái Hà[1],
Bùi Thị Kim Phượng[1,2],
Lê Thái Hưng[1,+]

[1]Trường Đại học Giáo dục - Đại học Quốc gia Hà Nội;
[2]Đại học Bách khoa Hà Nội
+Tác giả liên hệ ● Email: lthung@vnu.edu.vn

Article history
Received: 01/6/2023
Accepted: 14/7/2023
Published: 05/10/2023

Keywords
Item bank, adaptive testing, vocabulary assessment, high-frequency English words, IRT, equating

ABSTRACT
In the context of language education in Vietnam, while computerized adaptive testing can be seen as an inevitable development trend in the 4.0 era, only few studies have been published on this topic, including those on UEd-CAT 1.0 - an adaptive testing system of the University of Education - Vietnam National University, Hanoi. This study was conducted to build a standardized item bank to evaluate the Vietnamese EFL learners' knowledge of English high-frequency vocabulary. After the steps of writing 552 items and getting the items evaluated by item writing experts, the authors carried out a try-out with the participation of 1619 university students, analyzed the quality, calibrated the items with the Rasch model, equated the test forms to set item parameters onto a single scale, edited and finalized the item bank of 522 standardized questions. This question bank is believed to meet the requirements of the UEd-CAT 1.0 system, contributing to the content development of the system and to the EFL teaching and learning practices in Vietnam.

1. Mở đầu

Kỉ nguyên 4.0 của chuyển đổi số đã tác động đến mọi mặt của giáo dục và thúc đẩy đổi mới cách thức kiểm tra, đánh giá trong giáo dục. Trong lĩnh vực đánh giá ngôn ngữ, việc ứng dụng công nghệ máy tính đã trở nên phổ biến hơn ở mọi gia đình và trường học, từ đó tạo điều kiện thuận lợi cho một sáng kiến kiểm tra hiệu quả hơn - hệ thống trắc nghiệm thích ứng trên máy vi tính. Những năm gần đây, ngày càng có nhiều bài kiểm tra ngôn ngữ thích ứng trên máy vi tính đã và đang được tiếp tục phát triển; việc áp dụng CAT trong đánh giá ngôn ngữ cũng trở thành đề tài thảo luận nhận được sự quan tâm của nhiều nhà nghiên cứu trong những thập kỉ qua (Canale, 1986; Pathan, 2012; Khoshsima & Toroujeni, 2017; Okhotnikova et al., 2019).

Trong bối cảnh giáo dục Việt Nam, CAT vẫn còn là một lĩnh vực nghiên cứu mới mẻ cùng với mối quan tâm ngày càng tăng về ứng dụng công nghệ trong giáo dục trong những năm gần đây. Một số ít các nghiên cứu đã được thực hiện về việc phát triển và xác trị hệ thống trắc nghiệm thích ứng UEd-CAT 1.0 của Đại học Giáo dục - Đại học Quốc gia Hà Nội (Nguyễn Quỳnh Giang & Lê Thái Hưng, 2018; Le et al., 2019). Le và Nguyen (2021) đã báo cáo những kết quả tích cực thu nhận được từ việc chạy thử nghiệm hệ thống. Theo đó, hệ thống hiện đã cung cấp cho giáo viên và học sinh quyền truy cập miễn phí để đo năng lực làm toán và năng lực đọc hiểu tiếng Việt của HS lớp 10. Bài báo này miêu tả lại quá trình xây dựng ngân hàng câu hỏi trắc nghiệm thích ứng với mục đích kiểm tra từ vựng tiếng Anh, cụ thể là vốn từ vựng tiếng Anh thông dụng của người học tiếng Anh ở Việt Nam, tập trung vào việc biên soạn, thử nghiệm, định cỡ, chỉnh sửa và hoàn thiện ngân hàng câu hỏi, đóng góp vào việc mở rộng nội dung của hệ thống trắc nghiệm thích ứng UEd-CAT 1.0.

2. Kết quả nghiên cứu

2.1. Một số khái niệm liên quan

- Kiểm tra từ vựng tiếng Anh thông dụng của người học ngoại ngữ tiếng Anh: Trong đào tạo ngôn ngữ, cụ thể là tiếng Anh, kiểm tra từ vựng đóng một vai trò quan trọng, mang lại giá trị cả về nghiên cứu và thực tiễn với người dạy và người học cũng như các nhà nghiên cứu. Dù các bài kiểm tra từ vựng được phát triển rất đa dạng, có các cách tiếp cận khác nhau, không thể phủ nhận rằng các bài kiểm tra từ vựng tiếp nhận dạng viết là phổ biến nhất, thường được tiến hành với khía cạnh cơ bản nhất và quan trọng nhất của kiến thức từ vựng, đó là mối quan hệ giữa dạng từ

(form) và nghĩa (meaning), khía cạnh này làm nền móng để tiến hành việc học tập và lĩnh hội các khía cạnh khác của từ vựng (Webb et al., 2012). Hiện nay, nhiều nghiên cứu về kiểm tra từ vựng của người học tiếng Anh tại Việt Nam đều sử dụng các bài kiểm tra từ vựng tiếp nhận dạng viết để kiểm tra từ vựng của người học (Le & Nation, 2011; Nguyen & Webb, 2017; Dang, 2020).

- *Hệ thống trắc nghiệm thích ứng*: Trắc nghiệm thích ứng là một hệ thống kiểm tra trong đó máy tính được sử dụng để tạo ra một bài kiểm tra điều chỉnh theo trình độ của thí sinh. Việc xây dựng một hệ thống trắc nghiệm thích ứng đòi hỏi một ngân hàng câu hỏi chuẩn hóa với các tham số câu hỏi được xác định dựa trên lí thuyết hồi đáp câu hỏi - Item response theory (IRT) - làm nội dung kiểm tra, cụ thể trong nghiên cứu này là ngân hàng câu hỏi trắc nghiệm từ vựng thông dụng tiếng Anh, và các thuật toán để quyết định câu hỏi đầu tiên, thuật toán chọn các câu hỏi tiếp theo, thuật toán ước tính năng lực của thí sinh và thuật toán kết thúc bài kiểm tra (Thompson & Weiss, 2011). Trong một quy trình kiểm tra hoàn chỉnh, quy trình tiến hành kiểm tra bắt đầu với một câu hỏi được chọn từ ngân hàng câu hỏi đã hiệu chuẩn. Câu hỏi bắt đầu này có thể được chọn ngẫu nhiên hoặc từ một nhóm các mục có độ khó trung bình trong ngân hàng câu hỏi (Oppl et al., 2017; Choi & McClenen, 2020). Sau đó, tùy thuộc câu trả lời của thí sinh, hệ thống cung cấp các câu hỏi tiếp theo. Trong trường hợp thí sinh đưa ra câu trả lời chính xác cho câu hỏi được đưa ra, hệ thống sẽ cấp tiếp một câu hỏi có độ khó cao hơn; trong trường hợp thí sinh đưa ra câu trả lời không chính xác, hệ thống sẽ cấp tiếp một câu hỏi có độ khó thấp hơn. Quá trình này được lặp đi lặp lại cho đến khi hệ thống có đủ bằng chứng để xác định trình độ của thí sinh.

- *Ngân hàng câu hỏi*: Ngân hàng câu hỏi là thành phần đầu tiên, quyết định nội dung kiểm tra của hệ thống trắc nghiệm thích ứng. Với một hệ thống trắc nghiệm đã phát triển và đưa vào sử dụng thì các thuật toán đã được xác định từ trước, do đó, chất lượng của ngân hàng câu hỏi chất lượng đóng vai trò quyết định hiệu quả đánh giá năng lực của các thí sinh. Tất cả các câu hỏi trong ngân hàng đều đầu tiên được hiệu chỉnh bởi lí thuyết hồi đáp câu hỏi. Ba mô hình của lí thuyết hồi đáp có thể được áp dụng bao gồm mô hình một tham số kiểm tra các câu hỏi thử nghiệm theo chỉ một tham số, độ khó của câu hỏi; mô hình hai tham số phân tích cả độ khó của câu hỏi và độ phân biệt câu hỏi, và mô hình ba tham số bao gồm độ khó của câu hỏi, độ phân biệt câu hỏi và độ dự đoán hay đoán mò câu trả lời. Nghiên cứu đang thực hiện áp dụng mô hình một tham số, hay mô hình Rasch để hiệu chuẩn các câu hỏi thô. Phương trình của mô hình Rasch như sau:

$$P(u_i = 1 | \theta) = \frac{e^{\theta - b_i}}{1 + e^{\theta - b_i}} \quad (1)$$

Trong đó, θ là ước lượng năng lực thí sinh, P là khả năng trả lời đúng câu hỏi i của thí sinh với mức năng lực θ, u_i là câu trả lời của thí sinh cho câu hỏi, b_i là độ khó của câu hỏi i. Các tham số về độ khó của câu hỏi và ước tính năng lực thí sinh có cùng thang đo, thường nằm trong khoảng từ -3 đến +3.

Ở Việt Nam, mô hình Rasch được sử dụng phổ biến để phát triển đề kiểm tra trong nhiều nghiên cứu gần đây như Le và cộng sự (2019), Nguyễn Thái Hà và cộng sự (2021). Thompson và Weiss (2011) nhấn mạnh sự cần thiết của việc xây dựng ngân hàng câu hỏi không chỉ cần lưu ý đến số lượng câu hỏi trong ngân hàng, mà còn đến sự phân bố của các thông số câu hỏi và những cân nhắc thực tế như phân phối nội dung và các dự đoán về mức độ phân phối từng câu hỏi. Các tác giả cũng cho rằng việc xây dựng ngân hàng câu hỏi cần dựa trên những nghiên cứu thực nghiệm, cụ thể là tiến hành thử nghiệm bộ câu hỏi. Nhờ đó, các tham số của câu hỏi được ước tính thông qua phân tích thống kê về phản hồi thực tế của thí sinh đối với câu hỏi.

- *Cân bằng đề thi*: Để ước lượng năng lực của thí sinh bằng trắc nghiệm thích ứng cần một ngân hàng câu hỏi đã chuẩn hoá, nghĩa là các thông số của câu hỏi phải được xác định một cách chính xác và duy nhất. Để đảm bảo tính khoa học của việc chuẩn hoá ngân hàng câu hỏi, một phương pháp tiến hành cân bằng độ khó của các câu hỏi được thực hiện trong các nghiên cứu trước đây là phương pháp "câu hỏi neo" (David & Ida, 2019). Phương pháp cân bằng tham số của câu hỏi thi bằng câu hỏi neo là phương pháp sử dụng một số câu hỏi cùng xuất hiện trong các đề thử nghiệm khác nhau. Các câu hỏi neo này có thể là câu hỏi chung giữa tổ hợp 2 đề hoặc tất cả các đề.

Trong phương pháp cân bằng độ khó của câu hỏi thi bằng câu hỏi neo, các tham số độ khó (*b*), độ phân biệt (*a*), mức năng lực (θ) của câu hỏi trong đề thi $s+1$ được tính thông qua đề thi s bằng công thức sau:

$$\theta_{s+1} = A_{s,s+1}\theta_s + B_{s,s+1} \quad (2)$$

$$a_{s+1} = \frac{a_s}{A_{s,s+1}} \quad (3)$$

$$b_{s+1} = A_{s,s+1}b_s + B_{s,s+1} \quad (4)$$

Trong đó $n_{s,s+1}$ là các câu hỏi neo giữa đề thi S và đề thi $s+1$; $A_{s,s+1}$, $B_{s,s+1}$ là các hệ số cân bằng độ khó giữa hai đề thi. Trong nghiên cứu này, để tính các hệ số cân bằng chúng tôi sử dụng phương pháp "mean - mean" (Loyd & Hoover, 1980), hệ số cân bằng được tính như sau:

$$A_{s,s+1} = \left(\prod_{j=1}^{n_{s,s+1}} \frac{a_{sj}}{a_{(s+1)j}} \right)^{\frac{1}{n_{s,s+1}}} \quad (5)$$

$$B_{s,s+1} = \frac{1}{n_{s,s+1}} \sum_{j=1}^{n_{s,s+1}} b_{(s+1)j} - A_{s,s+1} \frac{1}{n_{s,s+1}} \sum_{j=1}^{n_{s,s+1}} b_{sj} \quad (6)$$

2.2. Phương pháp nghiên cứu

2.2.1. Công cụ nghiên cứu

Công cụ được thiết kế để thực hiện nghiên cứu là bộ câu hỏi thô kiểm tra từ vựng tiếng Anh thông dụng. Quy trình thiết kế công cụ nghiên cứu được chia làm hai giai đoạn chính: Giai đoạn một là biên soạn bộ câu hỏi thô từ một đặc tả được xác định từ trước, cụ thể là đặc tả của đề kiểm tra từ vựng tiếng Anh thông dụng NGSLT phiên bản song ngữ trong nghiên cứu của Bui và cộng sự (2022); Giai đoạn hai là xin ý kiến chuyên gia để chỉnh sửa bộ câu hỏi thô. Sau quá trình thẩm định của chuyên gia, bộ câu hỏi thô được chia thành 7 đề để tiến hành thử nghiệm, đề tuân thủ đặc tả của đề kiểm tra từ vựng tiếng Anh thông dụng NGSLT phiên bản song ngữ như trình bày ở trên, có một số các câu hỏi neo giữa các đề để tiến hành cân bằng đề sau giai đoạn thử nghiệm, đảm bảo các tham số của câu hỏi trong ngân hàng câu hỏi được ước lượng trên cùng một thang đo.

2.2.2. Đối tượng tham gia nghiên cứu

Thử nghiệm được thực hiện tại một trường đại học ở Hà Nội với 1.619 SV tham gia, là các SV khối kĩ thuật, đang tham gia các khóa học tiếng Anh tại trường. Việc chọn mẫu được thực hiện theo phương pháp lấy mẫu thuận tiện. Tất cả đối tượng đều tình nguyện tham gia thử nghiệm bằng cách làm bài kiểm tra từ vựng 100 câu và có thể xem kết quả số câu trả lời đúng ngay sau khi hoàn thành. Sau khi tiến hành làm sạch dữ liệu, nhóm nghiên cứu lấy kết quả và tiến hành phân tích để đánh giá chất lượng bộ câu hỏi thô cũng như định cỡ các tham số của câu hỏi trên cùng thang đo để đưa vào ngân hàng câu hỏi hiệu chuẩn.

2.3. Kết quả thực nghiệm và thảo luận

2.3.1. Đánh giá độ tin cậy của các đề

Bộ câu hỏi thô được chia thành 7 đề để tiến hành thử nghiệm. Conquest đã được sử dụng để thực hiện phân tích các đề kiểm tra. Hệ số Alpha và Separation Reliability của cả 7 đề đều trên 0,9. Các hệ số này cao cho thấy các đề thiết kế có độ tin cậy cao và các tham số câu hỏi có tính độc lập tốt.

Bảng 1. Độ tin cậy theo đề

Đề	Separation Reliability	Coefficient Alpha
1	0,978	0,92
2	0,968	0,93
3	0,958	0,95
4	0,964	0,95
5	0,972	0,95
6	0,960	0,96
7	0,963	0,97

2.3.2. Loại các câu hỏi không phù hợp với mô hình

Về mức độ phù hợp của câu hỏi với mô hình, chỉ số WEIGHTED FIT được sử dụng để phát hiện các câu hỏi không phù hợp. Bảng 2 minh họa kết quả phân tích của 5 câu bị loại của đề 1 do có chỉ số MNSQ không nằm trong phạm vi CI và thống kê T tương ứng của chúng có giá trị tuyệt đối vượt quá 2,0. Sau khi phân tích 7 đề bằng Conquest, có tổng số 30 câu hỏi bị loại do không phù hợp với mô hình phân tích.

Bảng 2. Các câu hỏi không phù hợp với mô hình của đề 1

VARIABLES			UNWEIGHTED FIT			WEIGHTED FIT		
item		ESTIMATE	ERROR^	MNSQ	CI	T	MNSQ	CI	T
39	39	2,020	0,243	2,86	(0,78, 1,22)	11,0	1,36	(0,70, 1,30)	2,1

54	54	-0,365	0,177	1,23 (0,78, 1,22) 1,9	1,18 (0,88, 1,12) 2,7
61	61	-0,751	0,182	0,68 (0,78, 1,22) -3,1	0,76 (0,86, 1,14) -3,7
73	73	0,707	0,184	0,74 (0,78, 1,22) -2,5	0,79 (0,85, 1,15) -3,0
95	95	0,380	0,178	1,29 (0,78, 1,22) 2,4	1,19 (0,87, 1,13) 2,7

2.3.3. Phân loại và chỉnh sửa câu hỏi

Hình 1 phân bố năng lực và độ khó của đề 7, một trong số 7 đề được chọn để minh họa. Từ bản đồ này, mức độ năng lực có thể đo trong khoảng -4 đến 4, điều này cho thấy rằng bài kiểm tra có thể đo được nhiều loại khả năng của con người. Độ khó của câu hỏi cần được chú ý nhiều hơn khi lượng câu hỏi có mức độ khó trên 2 còn hạn chế. Cũng đáng lưu ý là một số câu hỏi đo lường mức năng lực dưới -3; tuy nhiên, với mục đích đo lượng vốn từ vựng của thí sinh, các mục này vẫn có giá trị biểu thị lượng từ vựng mà thí sinh thu được và đánh giá năng lực thí sinh ở trình độ thấp nhất - thí sinh mới bắt đầu học hoặc ở trình độ sơ cấp.

Hình 1. Bản đồ phân bố năng lực và độ khó của đề 7

Để tăng chất lượng câu hỏi, nhóm tác giả sử dụng kết quả phân tích của Conquest với từng câu hỏi. 522 câu hỏi phù hợp với mô hình được phân chia thành hai nhóm: nhóm câu hỏi tốt và nhóm câu hỏi cần chỉnh sửa. Nhóm câu hỏi tốt phù hợp với mô hình Rasch, có độ khó và độ phân biệt chấp nhận được theo lí thuyết khảo thí cổ điển cùng với các phương án gây nhiễu đạt yêu cầu. Hình 2 là kết quả phân tích của một câu hỏi tốt của đề 4.

```
Item 20
-------
item:20 (20)
Cases for this item    150   Discrimination  0.34
Item Threshold(s):     0.05   Weighted MNSQ   1.11
Item Delta(s):      0.05
--------------------------------------------------------------------------------
 Label    Score     Count   % of tot  Pt Bis     t  (p)   WLEAvg:1 WLE SD:1
--------------------------------------------------------------------------------
   a       0.00        26      17.33   -0.15    -1.84(.068)  0.05     0.95
   b       1.00        85      56.67    0.34     4.36(.000)  0.90     1.52
   c       0.00        28      18.67   -0.26    -3.25(.001) -0.17     0.83
   d       0.00        11       7.33   -0.04    -0.46(.647)  0.19     0.61
================================================================================
```

Hình 2. Kết quả phân tích của câu hỏi 20 đề số 4

Với lí thuyết khảo thí cổ điển, độ khó của câu hỏi được tính là 0,57, nằm trong khoảng 0,25-0,75; 56,67% thí sinh trả lời đúng câu hỏi này. Độ phân biệt rất tốt, d = 0,34, cho thấy câu hỏi có thể phân biệt giữa nhóm thí sinh có trình độ cao và nhóm thí sinh có năng lực thấp. Hệ số tương quan (PT BIS) cho chúng ta thấy những phương án gây nhiễu có chỉ số âm và phương án đúng có chỉ số dương. Các phương án đưa ra ở câu 20 có giá trị đánh giá năng lực của thí sinh. Nhóm câu hỏi cần chỉnh sửa gồm các câu hỏi phù hợp với mô hình Rasch, có độ khó và độ phân biệt chấp nhận được theo lí thuyết khảo thí cổ điển, tuy nhiên nhóm nghiên cứu cần đầu tư thời gian và công sức để chỉnh sửa các phương án nhiễu. Tổng hợp kết quả của 7 đề, số lượng câu hỏi cần chỉnh sửa là 84 câu hỏi.

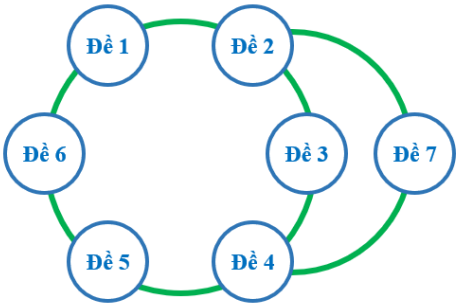
2.3.4. Cân bằng đề và chuẩn hóa ngân hàng câu hỏi

Trong nghiên cứu này, chúng tôi thiết kế lược đồ câu hỏi neo như sau:

Đề	*1*	*2*	*3*	*4*	*5*	*6*	*7*
1	90	**10**				**13**	
2		91	**13**				**4**
3			88	**14**			
4				92	**12**		**22**
5					86	**11**	
6						91	
7							83

Số câu hỏi neo giữa 2 đề

Số câu hỏi phù hợp mô hình sau khi phân tích Conquest



Hình 3. Lược đồ thiết kế câu hỏi neo giữa 7 đề

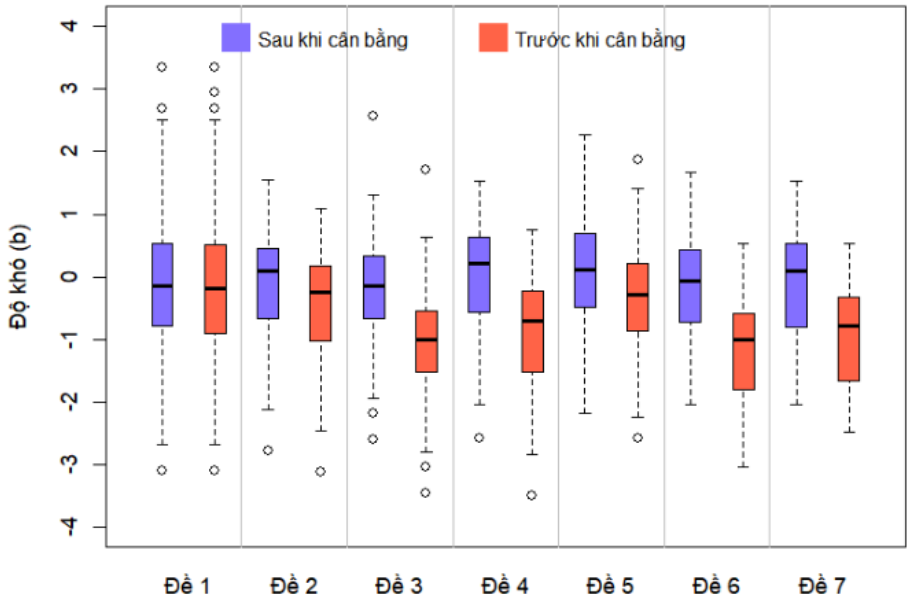
Sử dụng phần mềm thống kê R với phương pháp tính của Loyd và Hoover (1980), nhóm nghiên cứu tính toán được các hệ số cân bằng của 7 đề thi (theo lược đồ neo) như sau:

Bảng 3. Hệ số cân bằng 7 đề thi

Đề thi	Hệ số	Giá trị	Sai số	Hệ số	Giá trị	Sai số
1	A	1,000	0,000	B	0,000	0,000
2	A	1,000	0,000	B	0,343	0,093
3	A	1,000	0,000	B	0,853	0,094
4	A	1,000	0,000	B	0,906	0,100
5	A	1,000	0,000	B	0,398	0,092
6	A	1,000	0,000	B	0,994	0,084
7	A	1,000	0,000	B	0,860	0,105

Tham số độ khó của câu hỏi được ước lượng bằng mô hình Rasch 1 tham số (1PL - model), do vậy các hệ số cân bằng $A_{s,s+1}$ đều bằng 1. Các hệ số cân bằng $B_{s,s+1}$ tương ứng của 7 đề được trình bày trong bảng trên. Tiếp tục sử dụng các công thức của Loyd và Hoover (1980), nhóm nghiên cứu quy độ khó của câu hỏi trong 7 đề về cùng một thang đo. Độ khó của các câu hỏi sử dụng làm câu hỏi neo sau khi cân bằng vẫn đảm bảo khoảng cách tương đối về độ khó. Hơn nữa, sau khi cân bằng độ khó của các đề thi, độ khó của câu hỏi neo thuộc hai đề thi khác nhau đã được quy về một độ khó duy nhất. Các câu hỏi khác trong đề thi cũng được cân bằng theo công thức (4). Hình 4 so sánh độ khó của các câu hỏi thuộc 7 đề trước và sau khi cân bằng.

Hình 4 cho thấy, các đề thi thử nghiệm được xây dựng dựa trên cùng một bản đặc tả đề thi (cấu trúc theo nội dung và các cấp độ năng lực) nhưng chưa đảm bảo điều kiện để các đề thi có sự tương đương về độ khó. Phổ độ khó của đề thi sau khi cân bằng có tính ổn định hơn phổ độ khó của các đề thi khi chưa cân bằng (Hình 4) nên các câu hỏi này đảm bảo đủ tính khoa học và tính tương thích để đưa vào xây dựng một ngân hàng câu hỏi có số câu hỏi lớn.



Hình 4. Độ khó câu hỏi thi trước và sau khi cân bằng

3. Kết luận

Bài báo đã xây dựng ngân hàng câu hỏi chuẩn hóa để kiểm tra từ vựng tiếng Anh thông dụng cho đối tượng người học tiếng Anh ở Việt Nam. Sau khi xem xét kết quả phân tích của 7 đề, 30 câu hỏi thô không phù hợp với mô hình đang sử dụng nên sẽ bị loại bỏ, không đưa vào ngân hàng câu hỏi. 84 câu hỏi yêu cầu chỉnh sửa các phương án nhiễu để nâng cao chất lượng câu hỏi. Ngân hàng sau khi cân bằng có tổng 522 câu hỏi, đã được định cỡ trên cùng một thang đo để có thể sử dụng trong hệ thống trắc nghiệm thích ứng Ued-CAT1.0. Nghiên cứu này, cùng với các nghiên cứu trước đây về phát triển và xác trị đề kiểm (Le et al., 2019; Nguyễn Thái Hà và cộng sự, 2021) mang lại những đóng góp tích cực trong lĩnh vực kiểm tra, đánh giá tại Việt Nam, nhấn mạnh tầm quan trọng của một quy trình xây dựng đề thi nghiêm túc để đảm bảo độ giá trị và độ tin cậy của các công cụ kiểm tra, đánh giá cũng như khuyến khích sử dụng các công cụ phân tích để hỗ trợ các nhà giáo dục cũng như các nhà nghiên cứu thực hiện các nghiên cứu xác trị một cách hiệu quả. Những kết quả tích cực này có thể tạo tiền đề cho việc thiết kế các ngân hàng câu hỏi có giá trị và đáng tin cậy để phục vụ các mục đích và nhu cầu khác nhau của người học, giáo viên và nhà nghiên cứu trong bối cảnh giáo dục tại Việt Nam.

Mặt khác, nghiên cứu có một số hạn chế. Thứ nhất, nghiên cứu chưa thực hiện việc phân loại trình độ của đối tượng tham gia. Các nghiên cứu trong tương lai có thể cân nhắc vấn đề này để đảm bảo việc xác định năng lực thí sinh chính xác hơn, bổ sung thêm các minh chứng xác trị cho ngân hàng câu hỏi; Thứ hai, nghiên cứu này để lại "khoảng trống" cho các nghiên cứu xác trị trong tương lai bằng cách sử dụng các mô hình phân tích hai hoặc ba tham số và các gói phân tích khác nhau trong R để mang lại kết quả đa chiều và hữu ích hơn; Thứ ba, các nghiên cứu trong tương lai có thể hướng tới xây dựng ngân hàng câu hỏi với các nội dung khác hoặc áp dụng trắc nghiệm thích ứng để phục vụ kiểm tra và đánh giá thường xuyên, hỗ trợ thực hành dạy, học và nghiên cứu tại Việt Nam.

Tài liệu tham khảo

Bui, T. K. P., Nguyen, Q. T., & Le, T. H. (2022). The development of A Vietnamese-English Bilingual Version of the New General Service List Test. *2nd Hanoi Forum on Pedagogical and Educational Sciences*, Hanoi, Vietnam.

Canale, M. (1986). The promise and threat of computerized adaptive assessment of reading comprehension. In C. Stansfield (ed.), *Technology and language testing* (pp. 30-45). Washington, DC: TESOL Publications.

Choi, Y., & McClenen, C. (2020). Development of adaptive formative assessment system using computerized adaptive testing and dynamic bayesian networks. *Applied Sciences*, *10*(22), 8196. https://doi.org/10.1016/j.caeai.2022.100104

Dang, T. N. Y. (2020). Vietnamese non-English major EFL university students' receptive knowledge of the most frequent English words. *VNU Journal of Foreign Studies*, *36*(3).

David, A., & Ida, M. (2019). *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences.* Springer.

Khoshsima, H., & Toroujeni, S. M. H. (2017). Computer Adaptive Testing (Cat) Design; Testing Algorithm and Administration Mode Investigation. *European Journal of Education Studies*, *3*(5). https://doi.org/10.5281/zenodo.576047

Le, T. C. N, & Nation, P. (2011). A Bilingual Vocabulary Size Test of English for Vietnamese Learners. *RELC journal*, *42*(1), 86-99. https://doi.org/10.1177/0033688210390264

Le, T. H., & Nguyen, T. H. (2021). *Experimental Research and Application of Computerized Adaptive Tests to assess Learners' Competencies*. 2021 3rd International Conference on Computer Science and Technologies in Education (CSTE), Beijing, China. https://doi.org/10.1109/CSTE53634.2021.00021

Le, T. H., Tang, T. T, Tran, L. A., Nguyen, T. D., Nguyen, P. A., & Nguyen, T. Q. G. (2019). Developing Computerized Adaptive Testing: An Experimental Research on Assessing the Mathematical Ability of 10th Graders. *VNU Journal Of Science: Education Research*, *35*(4). https://doi.org/10.25073/2588-1159/vnuer.4301

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*(3), 179-193. https://doi.org/10.1111/j.1745-3984.1980.tb00825.x

Nguyễn Quỳnh Giang, Lê Thái Hưng (2018). Mô phỏng một bài kiểm tra thích nghi trên máy tính thông qua phần mềm R. *Tạp chí Khoa học Giáo dục Việt Nam, 11*, 6-11.

Nguyễn Thái Hà, Vũ Trọng Lưỡng, Lê Thái Hưng, Phạm Văn Hoằng (2021). Thiết kế câu hỏi trắc nghiệm thích ứng nhằm đánh giá năng lực toán học của học sinh lớp 12. *Tạp Chí Giáo dục*, *508*(2), 33-40.

Nguyen, T. M. H., & Webb, S. (2017). Examining second language receptive knowledge of collocation and factors that affect learning. *Language Teaching Research*, *21*(3), 298-320.

Okhotnikova, A., Daminova, J., Muzafarova, A., & Rasskazova, T. (2019). Challenges of designing and administering computer-adaptive tests. In *13th International Technology, Education and Development Conference (INTED)* (pp. 5633-5636). International Academy of Technology, Education and Development.

Oppl, S., Reisinger, F., Eckmaier, A., & Helm, C. (2017). A flexible online platform for computerized adaptive testing. *International Journal of Educational Technology in Higher Education*, *14*(1), 1-21. https://doi.org/10.1186/s41239-017-0039-0

Pathan, M. M. (2012). Computer Assisted Language Testing [CALT]: advantages, implications and limitations. *Research Vistas, 1*(4), 30-45.

Şahin, A., & Weiss, D. J. (2015). Effects of calibration sample size and item bank size on ability estimation in computerized adaptive testing. *Educational Sciences: Theory & Practice, 15*(6).

Thompson, N. A., & Weiss, D. A. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation*, *16*(1). https://doi.org/10.7275/wqzt-9427

Webb, S. A., & Chang, A. C. S. (2012). Second language vocabulary growth. *RELC Journal*, *43*(1), 113-126.