SỬ DỤNG PHẦN MỀM QUEST, CONQUEST ĐỂ PHÂN TÍCH ĐỀ THI TỰ LUẬN

• ThS. Võ Bình Nguyên(*)

Tóm tắt

Chương trình Quest/ConQuest là một phần mềm phân tích và đánh giá đề thi theo lý thuyết khảo thí hiện đại, được xây dựng dựa trên Lý thuyết đáp ứng câu hỏi (IRT). Trong bài viết này, tác giả trình bày các kết quả sử dụng để phân tích, đánh giá một đề thi tự luận nhằm hướng đến việc xây dựng ngân hàng đề thi tự luận đảm bảo tính khoa học, khách quan trong quá trình đánh giá.

Từ khóa: Lý thuyết hồi đáp, Quest, ConQuest, ngân hàng đề thi.

1. Đặt vấn đề

Công tác kiểm tra, đánh giá trong giáo dục xuất phát từ nhiều khía cạnh, trong đó đánh giá kết quả học tập của học sinh (HS) từ trước tới nay luôn được coi trọng. Chính kết quả học tập của HS là tiêu chí để thấy được sự trưởng thành, mức độ thành đạt và cũng là thước đo quan trọng để đánh giá trình độ tổ chức giáo dục trong các trường học hiện nay. Việc kiểm tra, kết quả học tập đề cập tới rất nhiều yếu tố mà mối quan hệ giữa những yếu tố ấy rất phức tạp, vì vậy trong đánh giá cần phải coi trọng đến việc nghiên cứu những kinh nghiệm của giảng viên đã tích lũy được trong quá trình kiểm tra, đánh giá kết quả học tập cho HS, mặt khác còn phải xuất phát từ những lý luận về đánh giá trong giáo dục, lý luận giáo dục học nói chung và các chính sách giáo dục để tiến hành nghiên cứu tổng hợp. Nếu công tác kiểm tra, đánh giá kết quả học tập của HS được tổ chức và tiến hành một cách hợp lý và đúng đắn, đảm bảo khách quan, công bằng sẽ là động lực thúc đẩy người học chủ động, tích cực sáng tạo và không ngừng nâng cao chất lượng đào tạo.

Trong kiểm tra, đánh giá kết quả học tập của HS hiện nay, "đối với mục tiêu nắm vững kiến thức, nếu phương pháp trắc nghiệm khách quan phù hợp với việc đánh giá các hiểu biết đơn lẻ tách biệt thì phương pháp tự luận thuận lợi hơn khi đánh giá một gói bao gồm nhiều kiến thức gắn kết với nhau, các ý tưởng có quan hệ với nhau" [5].

Trong phạm vi của bài viết này, tác giả sẽ sử dụng lý thuyết khảo thí hiện đại dựa trên Thuyết đáp ứng câu hỏi hay còn gọi là Lý thuyết hồi đáp (Item Response Theory - IRT) của Georg Rasch để phân tích đề thi môn Toán giải tích tại Trường Phổ thông Năng khiếu - Đại học Quốc gia Thành phố Hồ Chí Minh.

2. Giới thiệu về Lý thuyết hồi đáp IRT

Người ta thường phân chia lý thuyết khảo thí ra làm hai loại, lý thuyết khảo thí cổ điển và lý thuyết khảo thí hiện đại với việc sử dụng Lý thuyết hồi đáp IRT. Trong bài viết này, ta quan tâm đến các vấn đề về lý thuyết khảo thí hiện đại. Lý thuyết hồi đáp được xây dựng trên khoa học về xác suất và thống kê. Các công trình quan trọng của lý thuyết này ra đời vào 3 thập niên cuối của thế kỷ vừa qua và đạt được nhiều thành tựu quan trọng, được công nhận và áp dụng phổ biến trong thực tiễn. IRT đã đạt những thành tựu quan trọng góp phần nâng cao độ chính xác của đề thi.

"Lý thuyết hồi đáp được xây dựng dựa trên việc nghiên cứu mọi cặp tương tác nguyên tố "HS - câu hỏi". Mỗi HS đứng trước một câu hỏi sẽ ứng đáp như thế nào, điều đó phụ thuộc vào năng lực tiềm ẩn của HS và các đặc trưng của câu hỏi. Hành vi ứng đáp này được mô tả bằng một hàm đặc trưng câu hỏi cho biết xác suất trả lời đúng câu hỏi P(θ) tùy theo tương quan giữa năng lực HS(θ) và các tham số đặc trưng cho câu hỏi" [4]. Hiện nay, có 3 mô hình toán học trong IRT được sử dụng rộng rãi nhất: mô hình một tham số (mô hình Rasch) chỉ xét đến độ khó b của câu hỏi; mô hình 2 tham số có xét đến độ khó b và độ phân biệt a của câu hỏi; và mô hình 3 tham số, so với mô hình 2 tham số, còn xét thêm mức độ đoán mò c của HS khi trả lời câu hỏi. Dạng toán cụ thể của mô hình 3 tham số [2]:

$$P(\theta) = c+(1-c)\frac{\exp a(\theta - b)}{1+\exp a(\theta - b)}$$

(*) Trường Đại học Khoa học Xã hội và Nhân văn, Đại học Quốc gia Thành phố Hồ Chí Minh.

Trong lý thuyết khảo thí cổ điển, độ khó của câu hỏi trắc nghiệm được tính bằng cách lấy số HS trả lời đúng câu hỏi đó chia cho tổng số HS tham gia trả lời câu hỏi đó. Tuy nhiên, đối với một bài thi tự luận thì mỗi một câu hỏi được chia ra làm nhiều ý nhỏ, và ứng với mỗi một ý nhỏ sẽ có số điểm tương ứng của ý đó. Với việc ra đề thi theo hướng hiện đại, mỗi câu hỏi trong cùng một đề thi tự luận phải tương đối độc lập với nhau, nghĩa là kết quả của câu hỏi này sẽ không phụ thuộc vào kết quả trả lời câu hỏi khác. Chính vì vậy, việc cho điểm trong đề thi tự luận cũng sẽ khác rất nhiều so với đề thi trắc nghiệm. Điều này dẫn đến việc tính độ khó của câu hỏi trong đề thi tự luận trở nên hết sức phức tạp và mất nhiều thời gian.

3. Xử lý dữ liệu bằng phần mềm Quest/ ConQuest [1],[7]

Sau khi thi xong, các bài làm của HS sẽ được chấm vào phiếu chấm điểm chi tiết theo từng ý nhỏ theo thang điểm đã cho sẵn. Dữ liệu gồm điểm các câu hỏi mà HS đạt được sẽ được xử lý bằng phần mềm Quest hoặc ConQuest. Kết quả thu được gồm có: Độ tin cậy của bài thi, năng lực của HS, sự phù hợp của các câu hỏi, độ khó, độ phân biệt, độ tin cậy thống kê (P-Value) và độ tin cậy của từng câu hỏi thi, sai số,…

Dưới đây là bảng mô tả kết quả phân tích đề thi môn Toán giải tích - lớp 12 gồm ba câu hỏi do 300 HS thực hiện:

3.1. Mức độ phù hợp với mô hình Rasch

Trong lý thuyết hồi đáp IRT có nhiều mô hình đo. Trong bài này, mô hình Rasch sẽ được sử dụng để đo nếu kết quả phân tích cho thấy dữ liệu làm bài của HS phù hợp với mô hình Rasch. Nếu không phù hợp sẽ phải sử dụng mô hình khác. Dữ liệu nghiên cứu được cho là phù hợp với mô hình Rasch nếu trị số kỳ vọng Mean và độ lệch chuẩn SD của độ khó trung bình của các câu hỏi (hay độ khó chung của toàn bài) xấp xỉ bằng 0,0 và xấp xỉ bằng 1,0 một cách tương ứng [6].

Bảng 1. Mức độ phù hợp với mô hình Rasch

```
Summary of item Estimates
=========================
Mean                          .00
SD                           1.10
SD (adjusted)                1.06
Reliability of estimate       .94
Fit Statistics
==============
Infit Mean Square     Outfit Mean Square
Mean      .99             Mean    1.02
SD        .13             SD       .34

Summary of case Estimates
=========================
Mean                          .91
SD                            .76
SD (adjusted)                 .48
Reliability of estimate       .40
```

Kết quả tính toán ở bảng 1 cho thấy dữ liệu phù hợp với mô hình Rasch. Độ tin cậy của tính toán các thông số liên quan đến câu hỏi rất đáng tin cậy vì có giá trị bằng 0,94.

Bảng 1 cũng cho thấy: năng lực trung bình của mẫu thí sinh (case estimate) tham gia bài kiểm tra (0,91) lớn hơn nhiều so với độ khó chung của bài kiểm tra (0,0), nói cách khác là bài kiểm tra khá dễ so với năng lực của HS.

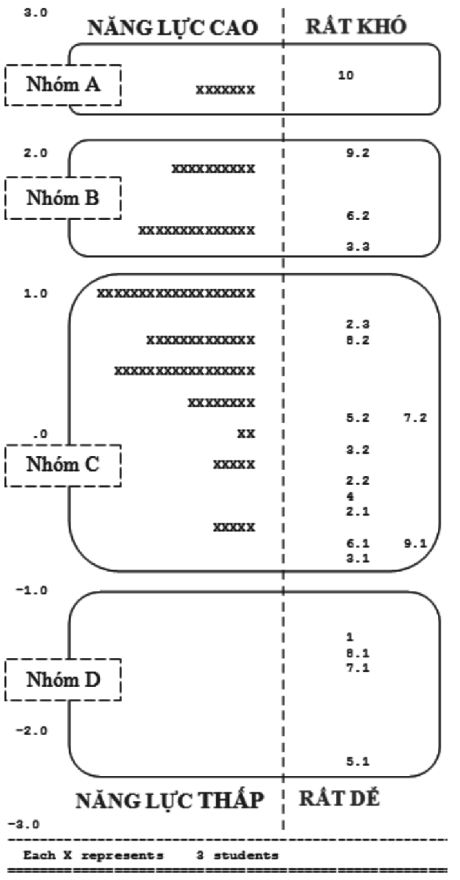
3.2. Mức độ phù hợp của các câu hỏi

Trong biểu đồ Item Fit các câu hỏi được biểu thị bằng dấu "*", các câu hỏi có chỉ số Infit MNSQ nằm trong khoảng cho phép từ 0,77 đến 1,30; các giá trị của từng chỉ số Infit MNSQ cũng được thể hiện cụ thể trong bảng 2, phân bố trong khoảng 0,81 - 1,19 nên tất cả câu hỏi trên đều phù hợp với mô hình Rasch, không có câu hỏi nào phải loại bỏ [1].

Bảng 2. Chỉ số Infit MNSQ

Câu	Infit MNSQ	Câu	Infit MNSQ
1	0,85	6	0,99
2	0,81	7	1,10
3	0,94	8	0,90
4	0,94	9	1,12
5	1,11	10	1,19

3.3. Sơ đồ phân bổ năng lực HS và độ khó câu hỏi thi



Hình 1. Sơ đồ phân bổ năng lực HS và độ khó câu hỏi thi

Sơ đồ phân bố độ khó câu hỏi thi và năng lực HS cho thấy mức độ phù hợp của đề thi đối với HS dự thi. Kết quả xử lý bằng phần mềm QUEST cho một bản đồ phân bố năng lực HS và độ khó đề thi.

Sử dụng lý thuyết khảo thí hiện đại, năng lực của HS và độ khó của câu hỏi được đánh giá bằng thang logistic, kết quả phân tích độ khó từ hình 1 cho thấy đề thi có các câu hỏi khó chiếm 5%, mức tương đối khó chiếm 15%, mức độ khó trung bình chiếm 55%, mức độ dễ chiếm 20%. Do vậy, khả năng phân loại HS của đề thi này là rất tốt.

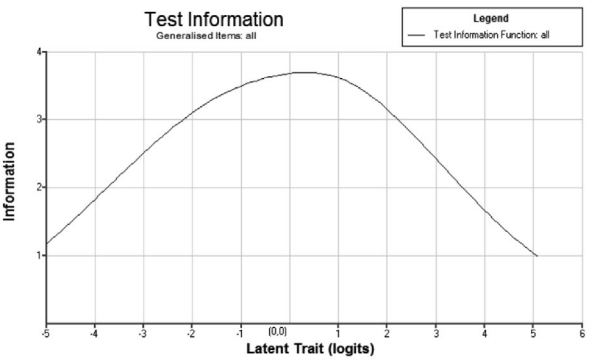
Bên cạnh đó, kết quả phân tích cũng chỉ ra được ngưỡng năng lực cần thiết để trả lời đúng các câu hỏi là từ -2,25 đến 2,54, trong khi năng lực của HS được phân bố từ -0,5 đến 2,40. Kết quả phân tích từ hình 1 cho thấy đây là đề thi phù hợp với năng lực của HS tham gia làm bài kiểm tra, cụ thể:

- Nhóm năng lực cao (A) chiếm 7%, gồm những câu hỏi khó với mức ngưỡng cao hơn 2,03 (cụ thể: câu 10 có ngưỡng thresholds là 2,54). Ở mức này HS có thể làm được hầu hết các bài toán về Giải tích trong đề thi.

- Nhóm năng lực tương đối cao (B) chiếm 24%, gồm những câu hỏi tương đối khó, có ngưỡng thresholds từ 1,39 - 2,03, các câu hỏi mà HS có thể đạt điểm tối đa là câu 3, 6, 9. Ở mức năng lực này, HS có thể khảo sát hàm số, hàm số luỹ thừa, số mũ và hàm số lôgarit, tính nguyên hàm, tích phân, có hiểu biết về hàm phức trong chương trình.

- Nhóm năng lực thấp (D) và năng lực trung bình (C) chiếm 69%, gồm những câu ở mức độ dễ, có ngưỡng thresholds từ -2,25 đến 0,18 gồm các câu hỏi 2, 4, 5, 8 và câu 3 (mức điểm 1-2), câu 6, 9 (mức điểm 1), câu 7 (mức điểm 2).

Kết quả phân tích đề thi bằng phần mềm ConQuest cung cấp đường cong đặc trưng thông tin của cả bài thi (hình 2) và cho thấy đây là một đề thi hay, có khả năng đánh giá cũng như phân loại được năng lực của các HS tham gia bài kiểm tra này.



Hình 2. Đường cong đặc trưng thông tin của cả bài thi

3.4. Độ tin cậy của đề thi

Kết quả tính toán bằng phần mềm Quest cho thấy độ tin cậy của đề thi đạt 0,94, điều này chứng tỏ đề thi có độ tin cậy cao.

3.5. Phân tích đề thi theo các tiêu chí khác

Tiếp tục xem xét các chỉ số khác thu được từ kết quả phân tích bằng phần mềm ConQuest theo các tiêu chí sau:

- *Discrimination:* Độ phân biệt hay mối tương quan giữa điểm của người trả lời với tổng điểm của toàn bài. Tốt nhất là lớn hơn 0.4;

- *Item Threshold(s):* Ngưỡng để thí sinh có thể vượt qua được câu hỏi đó (trả lời đúng);

- *Weighted MNSQ:* Sự phù hợp của mô hình, khi có sự xáo trộn giữa các loại lựa chọn thì giá trị này sẽ có xu hướng lớn hơn 1, điều này có thể dẫn đến những phát hiện về sự không phù hợp với mô hình.

- *Score:* Sự lựa chọn câu trả lời cho câu hỏi.

- *% of tot:* Tỷ lệ phần trăm trả lời cho từng lựa chọn.

- *Pt Bis:* Hệ số tương quan điểm Biserial (Point Biserial Correlation). Cần loại bỏ những câu hỏi có mối tương quan thấp hoặc dưới 0 sẽ làm tăng độ tin cậy của bài kiểm tra, nhưng tốt nhất là giá trị nằm trong khoảng 0,25 - 0,75 [3].

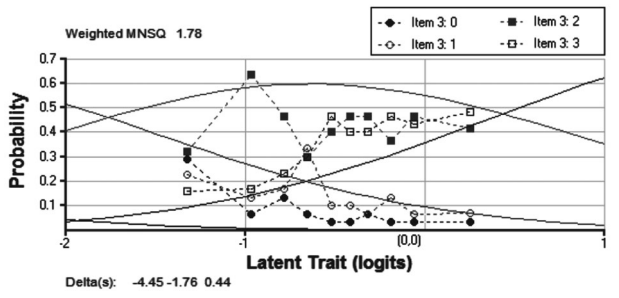
Dưới đây là kết quả phân tích câu 3:

- Độ khó p = 0,43: câu hỏi tương đối khó.

- Độ phân biệt: Discrimination = 0,57. Đây là câu hỏi có thể phân biệt HS, thể hiện được mối tương quan giữa điểm của HS trả lời với tổng điểm toàn bài.

- Weighted MNSQ = 1,78: Không phù hợp với mô hình, tuy nhiên đây là câu hỏi có độ phân biệt tốt và có sự xáo trộn giữa các lựa chọn (mức điểm của HS rải đều từ 0,0 đến 3,0). Chính vì vậy, theo tác giả câu 3 vẫn phù hợp với mô hình.

Hình 3. Đường cong đặc trưng thông tin của câu 3



Hình 4. Đường cong đặc trưng theo mức điểm (ICC)

Bảng dữ liệu trong hình 5 chính là dữ liệu ước tính giá trị hợp lý cho câu 3 khi thể hiện trên đường cong đặc trưng theo mức điểm (ICC).

- Trong khoảng năng lực thấp (> -3,00) và (-3,00 đến -2,25): không có HS nào đạt được điểm trong khoảng này.

- Khoảng năng lực từ (-2,25 đến -0,75): Có 81 HS có năng lực trong khoảng này, tuy nhiên chỉ có 13 HS không có điểm (chiếm 16,05%), 68 HS còn lại có 13 HS đạt 1,0 điểm (chiếm 16,05%), 41 HS đạt 2,0 điểm (chiếm 50,6%) và 14 HS đạt 3,0 điểm (chiếm 17,3%).

- Khoảng năng lực trong khoảng từ (-0,75 đến trên 0,0): Có 188 HS có năng lực trong khoảng này, và chỉ có 10 HS (chiếm 5,32%) không có điểm, 178 HS còn lại có 27 HS đạt 1,0 điểm (chiếm 14,36%), 75 HS đạt 2,0 điểm (chiếm 39,9%) và 76 HS đạt 3,0 điểm (chiếm 40,42%).

- Khoảng năng lực từ (0,0 đến 0,75): Chỉ có 31 HS, trong đó có 01 HS không có điểm (chiếm 3,22%), 02 HS đạt 1,0 điểm (chiếm 6,45%), 13 HS đạt 2,0 điểm (chiếm 41,94%) và 15 HS đạt 3,0 điểm (chiếm 48,39%).

Với kết quả phân tích như trên, ta thấy câu 3 có đến 276/300 HS có điểm câu này và 105 HS đạt điểm tối đa.

Kết quả này chỉ ra rằng, đây là một câu hỏi hay, chất lượng rất tốt, có khả năng đánh giá cũng như phân loại được năng lực của các HS.

4. Kết luận

Đánh giá kết quả đề thi tự luận là một trong những lĩnh vực còn rất mới lạ, chưa có nhiều nghiên cứu chuyên sâu cũng như chưa có những đánh giá đề thi hay thiết kế bộ câu hỏi tự luận cụ thể. Khi mà ngân hàng đề thi trắc nghiệm khách quan chưa được triển khai đồng bộ cho các môn học thì việc đánh giá chất lượng của các đề thi tự luận nhằm hướng đến việc xây dựng ngân hàng đề thi tự luận là một vấn đề cần thiết.

item:3 (3)

Bins			Frequencies					Proportions			
Low	High	Mean	0	1	2	3	Tot.	0	1	2	3
<-	-3.00000	0.00000	0	0	0	0	0	0.000	0.000	0.000	0.000
-3.00000	-2.25000	0.00000	0	0	0	0	0	0.000	0.000	0.000	0.000
-2.25000	-1.50000	-1.62269	3	1	2	1	7	0.429	0.143	0.286	0.143
-1.50000	-0.75000	-1.01164	10	12	39	13	74	0.135	0.162	0.527	0.176
-0.75000	0.00000	-0.38213	10	27	75	76	188	0.053	0.144	0.399	0.404
0.00000	0.75000	0.23874	1	2	13	15	31	0.032	0.065	0.419	0.484
0.75000	1.50000	0.00000	0	0	0	0	0	0.000	0.000	0.000	0.000
1.50000	2.25000	0.00000	0	0	0	0	0	0.000	0.000	0.000	0.000
2.25000	3.00000	0.00000	0	0	0	0	0	0.000	0.000	0.000	0.000
3.00000	->	0.00000	0	0	0	0	0	0.000	0.000	0.000	0.000

Hình 5. Bảng dữ liệu ước tính giá trị hợp lý cho biểu đồ ICC

Với sự hỗ trợ của 2 phần mềm chuyên dụng Quest và ConQuest, chúng ta đã phân tích đề thi tự luận một cách nhanh chóng, tiện lợi và có được một cái nhìn toàn diện về kết quả như sau: Chất lượng đề thi tốt, độ tin cậy của đề thi ở mức rất cao và đề thi có khả năng phân loại được năng lực tất cả HS tham gia kiểm tra đánh giá.

Tài liệu tham khảo

[1]. Adams, R. J., & Khoo, S. T. (1996), *Quest: Interactive test analysis system* [Computer Software], Victoria, Australia: Australian Council for Educational Research.

[2]. Baker, F. (1995), "The basics of item response theory", Wisconsin, EE. UU.: University of Wisconsin. Retrieved Octuber 13, 2005, http://edres.org/irt/baker/final.pdf.

[3]. Griffin, P. (1997), *An Introduction to the Rasch Model*, Melbourne: The University of Melbourne, Assessment Research Centre.

[4]. Phạm Xuân Thanh (2013), *Bài giảng môn Lý thuyết Đo lường đánh giá cho học viên cao học chuyên ngành Đo lường đánh giá*, Viện Đảm bảo chất lượng giáo dục, Đại học Quốc gia Hà Nội.

[5]. Lâm Quang Thiệp (2012), *Đo lường và đánh giá hoạt động học tập trong nhà trường*, NXB Đại học Sư phạm Hà Nội.

[6]. Wright, B. D., & Masters, G. N. (1982), *Rating scale analysis*, Chicago: Mesa Press.

[7]. Wu, M.L., Adams, R.J., and Wilson, M.R. (1997), *ConQuest*: *Multi-Aspect Test Software*, [computer program], Camberwell: Australian Council for Educational Research.

USING QUEST/ CONQUEST SOFTWARE TO ANALYZE ESSAY EXAMS

Summary

Quest/Conquest software is used for analyzing exams based on modern exam theories. It is grounded on Item Response Theory (IRT). This article presents those used results to analyse and evaluate an essay exam; as such to ensure that essay-exam banks are written scientifically and objectively in the evaluation process.

Key words: Item Response Theory, Quest, ConQuest, exam bank.

Ngày nhận bài: 3/9/2015; Ngày nhận lại: 29/9/2015; Ngày duyệt đăng: 5/1/2016.