ĐẢM BẢO TÍNH CÔNG BẰNG THI TUYỂN SINH TRONG GIÁO DỤC NGÀNH KHOA HỌC SỨC KHOẺ BẰNG PHÂN TÍCH THIÊN LỆCH CHỨC NĂNG CÂU HỎI TRẮC NGHIỆM

Phạm Dương Uyển Bình*, Trần Thị Diệu, Vĩnh Sơn**, Nguyễn Anh Vũ*****
***Phòng Đảm bảo chất lượng và Khảo thí, Đại học Y Dược TP.HCM**
****Bộ môn Tin học, Khoa Khoa học cơ bản, Đại học Y Dược Tp.HCM**
*****Bộ môn Toán, Khoa Khoa học cơ bản, Đại học Y Dược Tp.HCM**

Tóm tắt: *Thiên lệch chức năng câu hỏi (DIF) là một khái niệm trong đo lường và đánh giá giáo dục, xảy ra khi một câu hỏi trắc nghiệm (TN) có xác suất trả lời đúng khác nhau giữa các nhóm thí sinh về yếu tố như chủng tộc, giới tính, hoặc nền văn hóa, mặc dù họ có cùng mức độ năng lực. DIF có thể gây ra sự thiên vị và không công bằng trong các bài thi TN. Do đó, nghiên cứu này nhằm phát hiện thiên lệch ở các câu hỏi trong đề thi tại kỳ thi tuyển sinh sau đại học Y khoa năm 2023. Phương pháp nghiên cứu cắt ngang với dữ liệu từ một bài thi 120 câu trắc nghiệm, kỳ thi tuyển sinh chuyên khoa cấp 1 (CKI) môn Sinh lý tại Đại học Y dược TP.HCM với 1058 thí sinh thuộc các nhóm theo giới tính, loại tốt nghiệp và số năm tốt nghiệp. Phần mềm Basicstat và SPSS phiên bản 22.0 được dùng cho thống kê. Kết quả phân tích cho thấy, có thiên lệch chức năng trong số ít câu hỏi giữa các nhóm thí sinh (p<.05). Cụ thể, kết quả xác định được thí sinh nữ lại có ưu thế ở 8/120 câu trắc nghiệm. Tỷ lệ nhỏ số câu được phát hiện có DIF: 6/120 câu MCQ có thiên lệch chức năng đối với năm tốt nghiệp và 10/120 câu có DIF đối với phân loại tốt nghiệp. Đề thi môn Sinh lý cho thấy không có sự thiên lệch đáng kể giữa các nhóm học viên theo giới tính, năm tốt nghiệp và xếp loại tốt nghiệp.*

Từ khoá: *tính công bằng, thi tuyển sinh, ngành khoa học sức khoẻ, phân tích thiên lệch, chức năng, câu hỏi trắc nghiệm.*

ENSURING FAIRNESS ENTRANCE EXAMINATIONS IN HEALTH SCIENCE EDUCATION THROUGH FUNCTIONAL BIAS ANALYSIS OF MULTIPLE-CHOICE QUESTIONS

Abstract: *Differential Item Functioning (DIF) is a core concept in educational measurement and assessment, occurring when a multiple-choice question (MCQ) shows different probabilities of correct responses between groups of examinees based on factors such as race, gender, or culture, despite having the same level of ability. DIF can cause bias and unfairness in objective tests; therefore, this study aims to detect item bias in the 2023 medical postgraduate entrance examination. Using a cross-sectional design, the study analyzed data from a 120-item MCQ Physiology exam for the Specialist Level 1 admission at the University of Medicine and Pharmacy at Ho Chi Minh City, involving 1,058 candidates categorized by gender, graduation classification, and years since graduation. Statistical analyses were performed using Basicstat and SPSS version 22.0. The results indicated that a small number of questions exhibited functional bias between groups (p<.05). Specifically, female candidates held an advantage in 8/120 items, while a low prevalence of DIF was detected elsewhere: 6/120 items for years since graduation and 10/120 items for graduation classification. In conclusion, the Physiology exam showed no significant systemic bias among candidate groups based on gender, graduation year, or academic ranking.*

Keywords: *fairness, entrance examination, health science, bias analysis, function, multiple-choice questions.*

Nhận bài: 22/04/2026 *Phản biện: 22/05/2026* *Duyệt đăng: 25/05/2026*

I. ĐẶT VẤN ĐỀ

Thiên lệch chức năng câu hỏi trắc nghiệm (TN) là gì?

Thiên lệch chức năng câu TN hay diferential item function (DIF) là hiện tượng khi các câu hỏi trong bài kiểm tra có mức độ khó khác nhau đối với các nhóm khác nhau, không phải do sự khác biệt về năng lực cơ bản được đo lường mà do một số đặc điểm khác, chẳng hạn như chủng tộc, giới tính hoặc nền tảng văn hóa, giáo dục. Khi những cá nhân có cùng năng lực nhưng thuộc các nhóm khác nhau thể hiện hiệu suất khác nhau trong một câu hỏi kiểm tra, thì câu hỏi đó được cho là có DIF. Sự hiện diện của DIF có thể ảnh hưởng đáng kể đến độ hợp lệ của điểm số bài kiểm tra, đặc biệt là trong nhóm thí sinh đa dạng, bằng cách thể hiện sự thiên lệch làm suy yếu tính công bằng và độ chính xác của các đánh giá. Điều này có ý nghĩa quan trọng đối với việc giải thích và sử dụng điểm số bài kiểm tra trên các nhóm nhân khẩu học khác nhau.

Tác động đến tính giá trị của bài kiểm tra: Thiên vị trong so sánh điểm số: DIF có thể dẫn đến so sánh điểm số thiên vị giữa các nhóm, vì các câu hỏi có thể ưu tiên một nhóm hơn nhóm khác, không phản ánh sự khác biệt thực sự về năng lực hoặc kiến thức. Điều này có thể dẫn đến lợi thế hoặc bất lợi cho một số nhóm, dẫn đến đánh giá không chính xác về khả năng hoặc đặc điểm của họ. Ví dụ, nếu một câu hỏi dễ dàng hơn đối với một giới tính do định kiến văn hóa, thì điểm thi sẽ không phản ánh chính xác khả năng thực sự của cá nhân mà do khác nhau giới tính. Trong nghiên cứu của các tác giả Hop và cộng sự (2018) trong

kỳ thi tuyển sinh sau đại học, kết quả xác định được DIF theo giới tính và sắc tộc. Nghiên cứu của Phạm Dương Uyển Bình và cộng sự (2023) chỉ ra rằng Thí sinh kỳ thi tuyển sinh SĐH đa dạng giới tính, trường, loại và năm TN và các yếu tố trên có ảnh hưởng đến kết quả của thí sinh. Ảnh hưởng đến ước tính cấp độ nhóm: Nghiên cứu đã chỉ ra rằng sự hiện diện của nhiều câu hỏi DIF có thể thay đổi đáng kể điểm trung bình của nhóm và tăng lỗi chuẩn của các điểm này, điều này có thể làm méo mó so sánh giữa các nhóm. DIF có thể làm sai lệch các thống kê cấp độ nhóm, chẳng hạn như điểm trung bình và độ lệch chuẩn, điều này rất quan trọng đối với các đánh giá quy mô lớn, do đó, điều chỉnh DIF có thể làm giảm đáng kể sự thiên vị.

Phương pháp để hạn chế DIF: Để hạn chế DIF, việc phân tích kết quả kiểm tra giúp phát hiện sự hiện diện của thiên lệch. Một số phương pháp từng được ghi nhận trong các nghiên cứu nhằm phát hiện và giải quyết DIF. Mỗi phương pháp có những ưu điểm và hạn chế riêng, vì vậy thường được khuyến khích sử dụng nhiều phương pháp để thu thập bằng chứng cho DIF. Phương pháp Mantel-Haenszel (MH): Đây là một trong những phương pháp phổ biến nhất để phát hiện DIF. Phương pháp này so sánh tỷ lệ trả lời đúng giữa hai nhóm (thường là nhóm tham chiếu và nhóm mục tiêu) sau khi kiểm soát mức độ năng lực. Phân tích Logistic Regression: Phương pháp này mô hình hóa xác suất trả lời đúng của một câu hỏi dựa trên năng lực của người tham gia và nhóm mà họ thuộc về. Phân tích logistic regression có thể kiểm soát được nhiều yếu tố và xác định DIF một cách chính xác hơn.

II. NỘI DUNG NGHIÊN CỨU

2.1. Tổng quan về các nghiên cứu trước đây

Nghiên cứu của các tác giả Hope và cộng sự (2018) trong kỳ thi tuyển sinh sau đại học (SĐH), kết quả xác định được DIF theo giới tính và sắc tộc. Nghiên cứu của Phạm Dương Uyển Bình và cộng sự (2023) chỉ ra rằng Thí sinh kỳ thi tuyển sinh SĐH, các yếu tố giới tính, trường, loại và năm tốt nghiệp trên có ảnh hưởng đến kết quả của thí sinh. Do đó, điều chỉnh DIF có thể làm giảm đáng kể sự thiên vị.

Để hạn chế DIF, một số phương pháp hay được sử dụng nhằm phát hiện và giải quyết DIF. Phương pháp Mantel-Haenszel (MH): Đây là một trong những phương pháp phổ biến nhất để phát hiện DIF. Phương pháp này so sánh tỷ lệ trả lời đúng giữa hai nhóm, và phân tích Logistic Regression: Phương pháp này mô hình hóa xác suất trả lời đúng của một câu hỏi dựa trên năng lực của người tham gia. Một câu TN được xác định có DIF nếu $p<0.05$. Độ mạnh hiện tượng được đánh giá qua tỷ số chênh Odds Ratio (OR).

Tóm lại, sự hiện diện của DIF có thể ảnh hưởng đáng kể khả năng thiên vị trong đánh giá dẫn đến sự không công bằng. Do đó, nghiên cứu này nhằm đánh giá tình trạng thiên lệch chức năng của câu TN trong bài thi môn Sinh lý trong kỳ thi tuyển sinh CKI năm 2023 giữa các nhóm thí sinh thuộc 2 nhóm giới tính, xếp loại tốt nghiệp và nhóm năm tốt nghiệp khác nhau.

2.2. Phương pháp nghiên cứu

Nghiên cứu cắt ngang thực hiện trên 1.058 thí sinh dự thi tuyển sinh SĐH CKI môn Sinh lý học tại Đại học Y Dược TP.HCM năm 2023. Các biến nhân khẩu học bao gồm: giới tính, cơ sở đào tạo (ĐHYD TP.HCM và các trường khác), xếp loại tốt nghiệp và năm tốt nghiệp.

- Công cụ đo lường: Bộ đề thi gồm 120 câu hỏi trắc nghiệm (MCQ) 4 lựa chọn, tập trung vào kiến thức Sinh lý học nâng cao.

- Phân tích thống kê: Dữ liệu được xử lý bằng phần mềm phân tích câu nội bộ Basicstat và SPSS 22.0. Chức năng mục hỏi khác biệt (DIF) được xác định khi các nhóm có cùng mức năng lực nhưng xác suất trả lời đúng khác nhau.

2.3. Kết quả

Bảng 1. Sự khác biệt câu trả lời đúng giữa nam nữ

	Câu số
	Giới tính*
OR>1	54, 57, 58, 59, 67, 89, 91, 114
OR<1	3, 8, 14, 29
Tổng số câu có DIF	12

Ghi chú: *quy ước: nhóm Nam là nhóm tham chiếu

Trong cùng mức năng lực làm đúng từ 48 câu trở lên, giữa hai phái có khác biệt về khả năng làm đúng các câu số 3, 8, 14, 29, 50, 57, 59, 62, 64, 67, 70, 72, 89, 91, 96, 98, 114, 119 ($p<.05$), nam giới có khả năng làm đúng nhiều hơn ở đa số các câu này. Trong khi đó, học viên nữ có khả năng làm đúng cao hơn ở 1 số câu gồm 54, 57, 58, 59, 67, 89, 91, 114.

Bảng 2: Thiên lệch chức năng trong câu hỏi giữa các nhóm năm tốt nghiệp và xếp loại tốt nghiệp

	Năm tốt nghiệp		**Loại tốt nghiệp**	
	OR	**P**	**OR**	**p**
1	1.08	0.126	1.39	0.003*
2	1.12	0.02*	1.27	0.03*
3	0.82	0.04*	0.84	0.435
4	0.95	0.505	0.58	0.001*
5	1.25	p<.001*	1.27	0.04*
6	1.08	0.15	1.39	0.01*
7	1.02	0.70	1.28	0.03*

8	1.12	0.118	0.63	0.01*
9	0.89	0.02*	0.79	0.03*
10	0.98	0.849	0.76	0.04*
11	1.02	0.759	0.61	0.003*
12	1.19	0.005*	1.01	0.94
13	0.64	p<.001*	0.90	0.72
14	0.96	0.457	0.76	0.02
Tổng số câu có DIF		**6**		**11**

Nhóm năm tốt nghiệp: 1=Từ 2013 về trước; 2=Từ 2014 đến 2015; 3=Năm 2016; 4=Năm 2017; 5=Năm 2018; 6=Từ 2019 về sau Loại tốt nghiệp: 1=Giỏi; 2=Khá; 3=Trung bình khá; 4=Trung bình

Yếu tố số năm tốt nghiệp ảnh hưởng đến khả năng làm đúng của 6 trên 120 câu của đề thi. Cụ thể, những học viên có thời gian tốt nghiệp càng ngắn có xu hướng trả lời đúng các câu số 14, 49, và 95, trong khi đó học viên có thời gian tốt nghiệp dài hơn có xu hướng làm đúng các câu còn lại. Yếu tố xếp loại tốt nghiệp ảnh hưởng đến khả năng làm đúng của 11 trên 120 câu hỏi của đề thi gồm 6, 14, 32, 54, 57, 66, 69, 79, 87, 108. Cụ thể, các học viên có xếp loại TN thấp có khả năng trả lời đúng đa số các câu hỏi trên gồm 32, 66, 69, 79, 87, 108.

2.4. Bàn luận và đề xuất biện pháp

Kết quả nghiên cứu đã cung cấp một cái nhìn chi tiết và khách quan về đặc tính đo lường của đề thi thông qua việc phân tích thiên lệch chức năng. Việc xuất hiện DIF không đồng nghĩa với việc câu hỏi đó kém chất lượng, nhưng nó chỉ ra rằng các nhóm thí sinh có cùng mức năng lực thực tế lại có xác suất trả lời đúng khác nhau do ảnh hưởng của các biến số nhân khẩu học hoặc nền tảng đào tạo.

Sự khác biệt theo giới tính (Gender-based DIF)

Một phát hiện quan trọng là có đến 18 câu hỏi xuất hiện DIF theo giới tính, trong đó nam giới chiếm ưu thế ở đa số các câu, trong khi nữ giới chỉ vượt trội ở một số câu. Nam giới thường có xu hướng quyết đoán hơn trong các câu hỏi đòi hỏi tư duy không gian hoặc xử lý tình huống cấp cứu nhanh, trong khi nữ giới thường thể hiện tốt hơn ở các câu hỏi liên quan đến kỹ năng giao tiếp, đạo đức y khoa hoặc các nội dung đòi hỏi sự tỉ mỉ, chi tiết. Sự tồn tại của DIF giới tính đòi hỏi hội đồng khảo thí cần phân tích nội dung (content analysis) để đảm bảo ngôn ngữ và ngữ cảnh trong câu hỏi không chứa đựng định kiến giới (gender bias) tiềm ẩn.

Tác động của thâm niên tốt nghiệp và xếp loại học tập

Sự xuất hiện DIF ở biến "số năm tốt nghiệp" (6/120 câu) và "xếp loại tốt nghiệp" (10/120 câu) là một tín hiệu tích cực cho thấy bộ đề có tính ổn định cao. Những thí sinh mới tốt nghiệp thường có ưu thế ở các câu hỏi lý thuyết hàn lâm hoặc kiến thức cập nhật mới), trong khi những người tốt nghiệp lâu năm lại dựa vào kinh nghiệm thực hành lâm sàng tích lũy. Đáng chú ý, kết quả ghi nhận nhóm có xếp loại tốt nghiệp thấp lại có khả năng trả lời đúng cao hơn ở một số câu hỏi. Đây là một nghịch lý cần được giải mã: có thể những câu hỏi này rơi vào các mảng kiến thức bổ trợ hoặc kỹ năng thực hành mà sinh viên có học lực trung bình/khá tập trung rèn luyện nhiều hơn, hoặc các câu hỏi này đang đo lường một khía cạnh năng lực khác nằm ngoài khung đánh giá học thuật truyền thống tại trường đại học.

III. KẾT LUẬN

Nghiên cứu cho thấy bài thi tuyển sinh CKI môn Sinh lý học nhìn chung có độ ổn định và tính công bằng tương đối tốt. Phần lớn các câu hỏi trong đề thi không bị ảnh hưởng bởi các yếu tố thâm niên công tác hay xếp loại tốt nghiệp, điều này chứng tỏ bộ đề đã tập trung đo lường đúng năng lực chuyên môn của thí sinh. Tuy nhiên, phân tích sâu về kỹ thuật DIF (Chức năng mục hỏi khác biệt) đã chỉ ra một số điểm đáng lưu ý. Về giới tính: Có khoảng 10% (12/120 câu) mục hỏi xuất hiện thiên lệch giới tính. Trong đó, thí sinh nữ có ưu thế ở 8 câu và nam giới có ưu thế ở 4 câu. Điều này gợi ý rằng một số ngữ cảnh hoặc cách đặt vấn đề trong câu hỏi có thể đang "ưu ái" lối tư duy hoặc trải nghiệm đặc thù của một giới. Tác động tích lũy: Mặc dù số lượng câu hỏi có DIF ở từng hạng mục không cao, nhưng nếu không được điều chỉnh, các sai lệch nhỏ này có thể tích lũy lại làm ảnh hưởng đến tính chính xác của tổng điểm và lỗi chuẩn đo lường.

Để nâng cao chất lượng và tính công bằng cho các kỳ thi tuyển sinh y khoa trong tương lai, nghiên cứu đưa ra các khuyến nghị sau: Rà soát nội dung câu hỏi: Hội đồng chuyên môn cần kiểm

tra lại nội dung của 12 câu hỏi có DIF giới tính và 11 câu có DIF trường học. Mục đích là để xác định xem ngôn ngữ, thuật ngữ hoặc tình huống lâm sàng có chứa đựng định kiến vùng miền hay giới tính nào không, từ đó điều chỉnh lại cách đặt câu hỏi trung tính hơn. Xây dựng ngân hàng câu hỏi đa dạng: Cần sự tham gia biên soạn từ giảng viên của nhiều cơ sở đào tạo khác nhau để đảm bảo ngân hàng câu hỏi bao quát được các phương pháp giảng dạy đa dạng, tránh việc bộ đề chỉ tập trung vào thế mạnh của một trường duy nhất. Mở rộng nghiên cứu: Trong tương lai, cần phân tích thêm các yếu tố "giao thoa" (ví dụ: một thí sinh nữ, tốt nghiệp loại giỏi, ở vùng sâu vùng xa) để có cái nhìn toàn diện hơn về sự công bằng trong đánh giá giáo dục y khoa.

Lời cảm ơn: Nghiên cứu này được tài trợ kinh phí bởi Đại học Y Dược Thành phố Hồ Chí Minh theo Hợp đồng số 329/2025/HĐ-ĐHYD, ngày 25/09/2025.

TÀI LIỆU THAM KHẢO

Bond, T. G., & Fox, C. M. (2015). Applying the Rasch Model: Fundamental Measurement in the Human Sciences.

Hope, D., Adamson, K., McManus, I. C., Chis, L., & Elder, A. (2018). Using differential item functioning to evaluate potential bias in a high-stakes postgraduate knowledge based assessment. *BMC Medical Education,* 18, 1-7.

Phạm Dương Uyển Bình và cộng sự (2024). Mối quan hệ giữa độ khó năng và độ khó phân cách của các câu hỏi trắc nghiệm sinh lý học trong đề thi tuyển sinh sau đại học chuyên khoa I từ năm 2018-2022. *Tạp chí Y học TP.HCM*, 27(1), 170-176.

O'Neill, T. R., Wang, T., & Newton, W. P. (2022). The American Board of Family Medicine's 8 years of experience with differential item functioning. *The Journal of the American Board of Family Medicine,* 35(1), 18-25.

Holland, P. W., & Thayer, D. T. (1988). Differential Item Functioning and the Mantel-Haenszel Procedure. In H. Wainer & H. I. Braun (Eds.), Test Validity (pp. 129-145). *Lawrence Erlbaum Associates.*

Zumbo, B. D. (1999). A Handbook on the Theory and Methods of Differential Item Functioning (DIF). *National Defense Headquarters, Department of National Defense.*

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of Item Response Theory. *Sage Publications.*

Camilli, G., & Shepard, L. A. (1994). Methods for Identifying Biased Test Items. *SAGE Publications.*

Haist, S. A., Wilson, J. F., Elam, C. L., Blue, A. V., & Fosson, S. E. (2000). The Effect of Gender and Age on Medical School Performance: An Important Interaction. *Advances in health sciences education : theory and practice,* 5(3), 197–205. https://doi.org/10.1023/A:1009829611335

Rubright JD, Jodoin M, Woodward S, Barone MA. Differential Item Functioning Analysis of United States Medical Licensing Examination Step 1 Items. Acad Med. 2022 May 1;97(5):718-722. doi: 10.1097/ACM.0000000000004567. Epub 2022 Apr 27. PMID: 34907964.