Tạp chí Khoa học ĐHQGHN, Ngoại ngữ 26 (2010) 234-238

Khảo sát sự thống nhất của giảng viên trong việc đánh giá bài thi nói

Trần Thị Thanh Phúc*

Khoa Sư phạm tiếng Anh, Trường Đại học Ngoại ngữ, Đại học Quốc gia Hà Nội, Đường Phạm Văn Đồng, Cầu Giấy, Hà Nội, Việt Nam

Nhận ngày 20 tháng 12 năm 2010

Tóm tắt. Vai trò của kiểm tra - đánh giá đối với quá trình học ngoại ngữ là hết sức quan trọng. Tuy nhiên, với đặc thù của các bài thi nói, tính chủ quan của giám khảo chấm thi có thể ảnh hưởng tới độ chính xác trong việc cho điểm thí sinh. Nghiên cứu này khảo sát sự thống nhất trong cách chấm điểm của các giảng viên trẻ đối với các bài thi nói. Kết quả nghiên cứu cho thấy có sự chênh lệch khá lớn giữa điểm số của các giảng viên cho cùng một thí sinh. Đồng thời, những giảng viên có kinh nghiệm lâu năm hơn có cách chấm điểm nhất quán hơn. Thực tế này đặt ra yêu cầu quá trình tập huấn kỹ năng chấm thi cho các giảng viên nói chung và giảng viên trẻ nói riêng cần được tiến hành thường xuyên và hiệu quả hơn.

1. Đặt vấn đề

Trong quá trình dạy và học ngoại ngữ, kiểm tra - đánh giá có vai trò hết sức quan trọng. Nhờ quá trình này, người học có thể nhận thức được những điểm mạnh, điểm yếu của mình và nhờ đó có các điều chỉnh phù hợp nhằm đạt được tiến bộ trong học tập. Các kết quả kiểm tra - đánh giá ảnh hưởng tới từng cá thể, và cả cộng đồng [1].

Đối với các bài kiểm tra khách quan (objective tests), học viên lựa chọn hoặc điền những thông tin cần thiết vào một bản cho sẵn. Hình thức này thường được tiến hành đối với kỹ năng nghe và đọc. Kết quả là việc đánh giá khá nhất quán với nhau, nếu một bài thi được chấm bởi hai giáo viên, hoặc một giáo viên chấm cùng bài thi trong những khoảng thời gian khác nhau thì kết quả vẫn vậy.

Trái lại, với các bài kiểm tra mang tính chủ quan như các bài thi viết và nói, hai người chấm có thể đưa ra hai điểm số khác nhau đối với cùng một bài. Thậm chí một người chấm có thể cho điểm một bài nói khác nhau khi chấm vào những thời điểm khác nhau. "Điều này khiến cho ta khó có thể tin rằng những điểm số được cho trong một kỳ kiểm tra nói là chính xác và đáng tin cậy" [2].

Nắm bắt được đặc tính chủ quan cao trong việc đánh giá các bài kiểm tra nói, nghiên cứu này được tiến hành nhằm khảo sát năng lực kiểm tra - đánh giá của đội ngũ giảng viên tiếng Anh Trường Đại học Ngoại ngữ, Đại học Quốc gia Hà Nội.

2. Phương pháp nghiên cứu

2.1 Câu hỏi nghiên cứu

1. Đánh giá của giảng viên đối với năng lực

* ĐT: 84-982913669.
E-mail: thanhphuc0705@gmail.com

nói của sinh viên có thống nhất không?

2. Kinh nghiệm giảng dạy và việc được tập huấn chấm thi có ảnh hưởng đến sự chính xác trong đánh giá của giảng viên hay không?

2.2. Khách thể nghiên cứu

Khách thể nghiên cứu là 15 giảng viên đang trực tiếp giảng dạy cho sinh viên khoa Tiếng Anh Sư phạm tại tổ tiếng Anh 1. Các giảng viên đều dưới 30 tuổi. Về kinh nghiệm, 8 giảng viên có thời gian công tác dưới 6 tháng và 7 giảng viên còn lại có thời gian giảng dạy từ 2.5 năm trở lên.

2.3. Công cụ nghiên cứu

Nghiên cứu sử dụng dạng bài thi nói theo chuẩn PET (Preliminary English Test) theo khung Trình độ chung châu Âu (European Common Framework). Một bài thi nói được đánh giá theo bốn tiêu chí sau đây:

* Ngữ pháp và từ vựng

* Diễn ngôn

* Phát âm

* Giao tiếp tương tác

Các thông số được sử dụng để phân tích kết quả gồm:

- Mean (điểm trung bình): được tính bằng cách cộng tất cả các điểm của giám khảo chia cho tổng số giám khảo (15), viết tắt là M.

- Standard deviation: được dùng để xác định độ lệch của các điểm chấm so với điểm trung bình, viết tắt là SD.

Ngoài ra, nghiên cứu sử dụng kết quả đánh giá cuối cùng của giảng viên sau khi tất cả giảng viên đã cho điểm và cùng thảo luận về điểm số của sinh viên. Điểm kết luận này được coi là điểm chuẩn và được viết tắt là C.

2.4. Các bước tiến hành nghiên cứu

Nghiên cứu được tiến hành trong buổi tập huấn giám khảo nói của Bộ môn Tiếng Anh 1, Khoa Sư phạm tiếng Anh trường Đại học Ngoại ngữ, ĐHQG Hà Nội. Trong buổi tập huấn này, các giảng viên được phát tài liệu và thảo luận về phương thức chấm các bài thi nói học phần 1. Trong phần tiếp theo, các giảng viên tiến hành chấm thử bài thi nói của một cặp thí sinh A và B (đã được quay video từ trước). Kết quả chấm được thảo luận để cùng thống nhất điểm số đối với từng tiêu chí chấm thi. Sau đó các giảng viên tiếp tục chấm bài thi nói của cặp thí sinh thứ hai C và D. Kết quả chấm thi của mọi người được lưu lại để phục vụ nghiên cứu.

3. Kết quả nghiên cứu

3.1. Câu hỏi nghiên cứu 1: Đánh giá của giảng viên đối với năng lực nói của sinh viên có thống nhất không?

* Thể hiện qua thông số Mean (điểm trung bình)

Bảng 1: Sự thể hiện qua thông số Mean

	Ngữ pháp - từ vựng		Diễn ngôn		Phát âm		Giao tiếp tương tác	
SV	C	M	C	M	C	M	C	M
A	7.5	7.46	7	7.67	7	7.33	9	8.27
B	8	8.06	8	7.67	9	8	9	8.53
C	8	7.4	6	6.93	7	6.67	8	7.87
D	8	7.93	7	7.73	8	7.8	9	8.67

Qua biểu đồ trên, ta có thể thấy rằng độ chênh lệch giữa điểm chuẩn so với điểm trung bình của 15 giảng viên là rất thấp, đa số thấp hơn 1.0. Tuy nhiên, các biểu đồ cũng thể hiện một xu hướng khá rõ. Đó là đối với những sinh viên có điểm chuẩn tương đối ở mức trung bình và khá (6-7 điểm) thì điểm thực tế giảng viên cho là cao hơn điểm chuẩn. Ngược lại, với

những sinh viên có điểm chuẩn ở mức tốt và rất tốt (8,9 điểm) thì điểm thực tế giảng viên cho lại thấp hơn điểm chuẩn.

* Thể hiện qua thông số Standard Deviation

Nếu chỉ nhìn vào thông số Mean (điểm trung bình), ta có thể thấy việc giảng viên đánh giá năng lực nói của sinh viên có sự chính xác tương đối cao. Tuy nhiên, tiêu chí điểm trung bình này không phản ánh được thực tế là có sự chênh lệch giữa điểm số của từng cá nhân giảng viên. Ví dụ: Nếu một giảng viên cho sinh viên 5 điểm, một giảng viên cho 9 điểm, thì điểm trung bình của sinh viên là 7 điểm. Điểm này có thể trùng với điểm chuẩn nhưng rõ ràng cách đánh giá của hai giảng viên là hoàn toàn khác nhau. Bởi vậy, để có kết luận chính xác hơn, chúng ta sử dụng thông số Standard Deviation. Thông số này xét đến mức độ chênh lệch trung bình thực tế giữa điểm chấm của các giảng viên. Với mỗi tiêu chí và mỗi sinh viên, thông số này được xác định như sau:

Bảng 2: Sự thể hiện qua thông số Standard Deviation

Sinh viên	Ngữ pháp và từ vựng	Diễn ngôn	Phát âm	Giao tiếp tương tác
A	0.83	0.97	1.17	0.96
B	0.59	0.72	0.75	0.63
C	0.82	0.96	0.61	0.74
D	0.59	0.45	0.41	0.48

Từ bảng dữ liệu trên, ta thấy rõ độ chênh lệch trung bình thấp nhất là 0.45, cao nhất là 0.97 (trừ tiêu chí phát âm của sinh viên A - 1.17). Điều này phản ánh thực tế rằng với một sinh viên, nếu được hỏi bởi hai giám khảo khác nhau thì điểm thi nói có thể vênh trong khoảng từ 0.9 đến 2 điểm.

3.2. Kinh nghiệm giảng dạy và việc được tập huấn chấm thi có ảnh hưởng đến sự chính xác trong đánh giá của giảng viên hay không?

* Kinh nghiệm giảng dạy

Các khách thể nghiên cứu được chia thành 2 nhóm giảng viên, một nhóm có kinh nghiệm dưới 6 tháng (là các giảng viên trẻ mới ra trường, gồm 8 giảng viên), và một nhóm có kinh nghiệm giảng dạy từ 2.5 năm trở lên. Nhóm các giảng viên mới ra trường được gọi là nhóm 1, nhóm còn lại là nhóm 2.

- Thể hiện qua thông số Mean (Bảng 4):

Qua biểu đồ, nhìn chung sự chênh lệch giữa điểm của hai nhóm so với điểm chuẩn là không lớn. Tuy nhiên, nhóm 2 có điểm trung bình gần với điểm chuẩn hơn so với nhóm 1.

- Thể hiện qua thông số Standard Deviation (Bảng 5):

Bảng 4: So sánh thông số Mean

SV	Ngữ pháp - từ vựng		Diễn ngôn		Phát âm		Giao tiếp tương tác	
	Nhóm 1	Nhóm 2	Nhóm 1	Nhóm 2	Nhóm 1	Nhóm 2	Nhóm 1	Nhóm 2
A	7.63	7.14	7.75	7.43	7.00	7.43	8.38	8.14
B	8.13	8.00	7.50	7.71	8.00	8.00	8.50	8.57
C	7.25	7.71	7.13	6.86	6.50	6.86	7.75	7.86
D	8.00	7.86	7.50	7.86	8.00	7.57	8.63	8.71

Bảng 5: So sánh thông số Standard Deviation

	Ngữ pháp - từ vựng		Diễn ngôn		Phát âm		Giao tiếp tương tác	
SV	Nhóm 1	Nhóm 2	Nhóm 1	Nhóm 2	Nhóm 1	Nhóm 2	Nhóm 1	Nhóm 2
A	0.92	0.69	0.89	**0.98**	1.41	0.98	0.74	**1.21**

B	0.64	0.58	0.76	0.49	0.76	**0.82**	0.76	0.53
C	0.71	**0.76**	0.99	0.90	0.53	**0.69**	0.71	**0.90**
D	0.76	0.38	0.53	0.38	0.00	**0.53**	0.52	0.49

(Những thông số ở nhóm 2 lớn hơn nhóm 1 được bôi đậm)

Số liệu trên phản ánh thực tế là các giảng viên ở nhóm 2 (nhóm gồm các giảng viên nhiều kinh nghiệm hơn) có cách chấm điểm nhất quán hơn so với các giảng viên ở nhóm 1.

* Tập huấn chấm thi

Thông số Standard Deviation của các giảng viên đối với từng sinh viên cả về tổng thể (xem bảng 2) và đối với từng nhóm giảng viên (xem bảng 5) cho thấy: đối với sinh viên C và D, thông số này nhỏ hơn hẳn hai thông số của sinh viên A và B. Điều này chứng tỏ sau khi được tập huấn và thảo luận, giảng viên có cách cho điểm thống nhất hơn.

4. Kết luận chung

Việc phân tích số liệu cho thấy rõ đánh giá của giảng viên đối với năng lực nói của sinh viên chưa có tính thống nhất cao. Bên cạnh đó, giảng viên có xu hướng nâng điểm cho những sinh viên có năng lực thuộc loại trung bình và khá, nhưng lại có xu hướng hạ thấp điểm đối với những sinh viên có năng lực tốt và giỏi. Điều này phần nào phản ánh tâm lý "cào bằng" của giảng viên. Đây là một đặc điểm tâm lý cần phải được khắc phục. Kết quả nghiên cứu cũng phản ánh nhóm giáo viên có kinh nghiệm giảng dạy nhiều hơn (từ 2.5 năm trở lên) có sự đánh giá chính xác hơn và nhất quán hơn so với nhóm giáo viên trẻ có kinh nghiệm giảng dạy dưới 6 tháng. Đồng thời, việc được tập huấn giúp các giảng viên có sự đánh giá thống nhất hơn.

5. Đề xuất

Dựa trên kết quả nghiên cứu, chúng tôi xin đưa ra một số đề xuất cụ thể. Thứ nhất, đội ngũ giảng viên trẻ cần thường xuyên được bồi dưỡng và tập huấn nhằm nâng cao khả năng giảng dạy và đánh giá. Các buổi tập huấn cần được tiến hành với mật độ nhiều (khoảng 1 lần/1 tháng) tại tất cả các tổ bộ môn trong trường nhằm giúp giảng viên có nhiều điều kiện thực tập và điều chỉnh cách cho điểm sao cho phù hợp nhất.

Nghiên cứu này cho thấy chưa có sự thống nhất cao trong việc chấm thi nói đối với các giảng viên dạy trong cùng một tổ, có cùng đối tượng giảng dạy và sử dụng cùng các tiêu chí chấm điểm. Trong khi đó, các kỳ thi nói được tiến hành tại khoa Sư phạm tiếng Anh được tổ chức với lực lượng giám khảo gồm giảng viên từ nhiều tổ bộ môn khác nhau. Để kết quả thi có độ tin cậy cao nhất, việc tập huấn giám khảo nói cần được tiến hành trước khi các kỳ thi nói diễn ra và chỉ những giảng viên có cách chấm nhất quán và tương đối chính xác mới được lựa chọn làm giám khảo nói.

APPENDIX

PHIẾU ĐÁNH GIÁ CỦA GIÁM KHẢO

Speaking Test Assessment Scales (PET – B1 level)

Analytical Scales

Name	Grammar and Vocabulary	Discourse Management	Pronunciation	Interactive Communication
A				
B				
C				
D				

Tài liệu tham khảo

[1] A. Davies (ed.), *Language Testing 14/3* (special issue on ethics in language testing), 1997.

[2] N. Underhill, *Testing Spoken Language,* Cambridge University Press, 1987.

[3] R. Burns, *Introduction to Reseach Methods*, SAGE Publications, 2000.

A case study into oral examiners' consistency in assessing students' performance in a speaking test

Tran Thi Thanh Phuc

Faculty of English Language Teacher Education, College of Foreign Languages, Vietnam National University, Hanoi, Pham Van Dong Street, Cau Giay, Hanoi, Vietnam

Testing and assessment play a very important role in language teaching and learning. However, the low reliability in the assessment of oral examiners in speaking tests makes it hard to trust in the scores awarded to test takers. This research was carried out in order to investigate the consistency of oral examiners, who are also lectures of English, in their assessment of students' speaking performance. The findings suggested that there was a big difference between the scores that the examiners gave to a single student's performance. In addition, examiners who had more teaching experience tended to have more consistent scores. Therefore, more training programs for lecturers need to be carried out, and only those who are qualified enough should be selected to be oral examiners.