Tạp chí Khoa học ĐHQGHN, Ngoại ngữ 27 (2011) 115-123

Chuẩn đánh giá trong dạy và học ngoại ngữ

Nguyễn Quang Thuấn*

Trung tâm Đào tạo từ xa và Bồi dưỡng giáo viên, Trường Đại học Ngoại ngữ, Đại học Quốc gia Hà Nội, Đường Phạm Văn Đồng, Cầu Giấy, Hà Nội, Việt Nam

Nhận ngày 15 tháng 6 năm 2011

Tóm tắt. Ngày nay người ta nói nhiều đến chuẩn: Chuẩn kiến thức và kỹ năng, Chuẩn chương trình, vv… Trong lĩnh vực đánh giá nói chung và đánh giá trong dạy và học ngoại ngữ nói riêng, người ta đặc biệt đề cập đến Chuẩn đánh giá. Tuy nhiên, việc hiểu chuẩn đánh giá về phương diện lý luận và thực tiễn và nhất là cách thức thực hiện Chuẩn đánh giá trong dạy và học ngoại ngữ chưa được giới chuyên môn ở Việt Nam quan tâm và làm sáng tỏ. Trong bài viết này, chúng tôi sẽ cố gắng trình bày những vấn đề cơ bản nhất liên quan đến đánh giá như mục tiêu, nội dung, cách đánh giá và công cụ đánh giá. Đặc biệt, chúng tôi cũng đề cập đến một số loại hình trắc nghiệm cùng với các đặc tính quan trọng của chúng để làm rõ quan niệm cũng như nội dung của Chuẩn đánh giá trong dạy và học ngoại ngữ.

Từ khóa. Đánh giá, chuẩn đánh giá, trắc nghiệm, trắc nghiệm tham chiếu tiêu chí, trắc nghiệm tham chiếu qui chuẩn, độ tin cậy, tính hiệu lực, chỉ số hay độ khó, chỉ số phân loại.

Ngày nay người ta nói nhiều đến các loại chuẩn: chuẩn kiến thức và kỹ năng, chuẩn chương trình, chuẩn học, vv... Trong lĩnh vực đánh giá nói chung và đánh giá trong dạy và học ngoại ngữ nói riêng, người ta đặc biệt đề cập *Chuẩn đánh giá*. Song việc hiểu như thế nào là chuẩn đánh giá về cả lý luận và thực tiễn chưa được làm sáng tỏ và nhất là làm thế nào để thực hiện được chuẩn đánh giá trong dạy và học ngoại ngữ cũng chưa được quan tâm. Bài viết này mong muốn góp phần làm sáng tỏ vấn đề này.

Vậy chuẩn đánh giá là gì? *Chuẩn đánh giá* có thể hiểu một cách đơn giản là *đánh giá được cái cần đánh giá*. Thực tế, *đánh giá* không chỉ là một khái niệm, mà đúng hơn nó vừa là một quá trình và vừa là một sản phẩm. Là *quá trình* vì đánh giá là thu thập một cách hệ thống các thông tin để ra quyết định [1] và như vậy để thu thập được các thông tin chúng ta phải thực hiện một loạt các hoạt động và phải tuân theo các giai đoạn và các bước tiến hành cụ thể. Là *sản phẩm* vì kết quả của các hoạt động này, của quá trình này là có được một công cụ đánh giá chuẩn, đủ khả năng đo cái cần đo, đủ khả năng đánh giá các kiến thức hay kỹ năng cần phải đánh giá và để cuối cùng đưa ra được các quyết định đúng đắn và chính xác.

Thật vậy, nếu như mục đích cuối cùng của đánh giá là ra các quyết định đúng đắn và chính xác, thì điều quan trọng trước hết là phải xác định được chính xác, rõ ràng mục tiêu đánh giá, tức là "tại sao đánh giá?" và muốn thực hiện được mục tiêu đánh giá thì phải xác định được "khi nào đánh giá?", "đánh giá cái gì?" và "đánh giá như thế nào?", "một công cụ đánh giá như thế nào được coi là có độ tin cậy và tính hiệu lực cao?", "một công cụ đánh giá như thế nào được coi là có

* ĐT: 84-912004484.
E-mail: ngquangthuan@yahoo.fr

chỉ số khó và chỉ số phân loại thích hợp?" và "chọn công cụ đánh giá như thế nào?". Trả lời các câu hỏi này cho phép chúng ta trả lời được câu hỏi : « Thế nào là *Chuẩn đánh giá ?* ».

Tại sao đánh giá?

Theo Lussier [2], người ta bao giờ cũng đánh giá theo một mục đích hay một ý định, có nghĩa là theo loại thông tin mà người ta cần để đưa ra các phán quyết hay các quyết định xác đáng. Người chuẩn bị nội dung thi/kiểm tra phải trả lời được câu hỏi sau đây : "*Tại sao đánh giá?*" hay "*Mục tiêu đánh giá là gì?*" Nếu mục tiêu là để phân loại, để xác nhận trình độ, hay để chuyển lên học ở một trình độ cao hơn thì nên dùng trắc nghiệm tham chiếu qui chuẩn (Normed Referenced Assessement). Nếu mục tiêu chỉ là khảo sát trình độ, năng lực của tất cả các sinh viên trong một lớp hay một nhóm để xác định khó khăn, trở ngại của từng sinh viên nhằm giúp cho họ khắc phục và từ đó điều chỉnh quá trình dạy và học thì có thể không cần đến trắc nghiệm mà chỉ cần đến một cuộc điều tra hay phỏng vấn chẳng hạn. Nên đưa ra tất cả các mục tiêu, từ đó chọn lựa ưu tiên theo thứ tự quan trọng của từng mục tiêu. Không nên nhằm quá nhiều mục tiêu trong một lần thi/kiểm tra. Cần phải xác định các mục tiêu này quan trọng và có giá trị như thế nào đối với người học, người dạy, cán bộ quản lý, chỉ đạo, vv. và kết quả nào được coi là chủ yếu. Về vấn đề này, người ta thường dựa vào bảng phân loại mục tiêu giáo dục của Bloom [3]. Về năng lực tư duy nhận thức của con người, tác giả chia làm 6 mức độ sau đây :

1) *Nhận biết* (Knowledge): Ghi nhớ được các sự kiện, thuật ngữ và các nguyên lý dưới hình thức mà người học đã được học.

2) *Hiểu* (Comprehension): Hiểu được các vấn đề đã được học. Người học phải có khả năng diễn giải, mô tả tóm tắt thông tin đã thu nhận được.

3) *Ứng dụng* (Application): Sử dụng được các thông tin, kiến thức, kỹ năng trong các tình huống khác với các tình huống đã được học. Đòi hỏi khả năng khái quát hoá hoặc trừu tượng hoá phù hợp với các tình huống cụ thể.

4) *Phân tích* (Analysis): Biết tách từ tổng thể thành bộ phận và nắm chắc mối liên hệ giữa các thành phần đó với nhau cùng với cấu trúc của chúng.

5) *Tổng hợp* (Synthesis): Biết kết hợp các bộ phận thành một tổng thể mới từ một tổng thể cũ. Đòi hỏi khả năng phân tích đi đến tổng hợp và ở đây bắt đầu thể hiện tính sáng tạo của cá nhân người học.

6) *Đánh giá* (Evaluation): Có khả năng phân tích, phê phán, chọn lọc, quyết định, đánh giá trên cơ sở các tiêu chí và tính hợp lý. Đòi hỏi phải có khả năng tổng hợp để đánh giá.

Khi nào đánh giá?

Người ta đánh giá vào những thời điểm khác nhau của một quá trình học tập hay đào tạo để đáp ứng các nhu cầu khác nhau. Điều này rất quan trọng. Nếu sau một quá trình đào tạo hay học tập như kết thúc một học phần, một môn học hay một chương trình thì người ta dùng đánh giá tổng kết (Summative assessment). Nếu đánh giá trong quá trình đào tạo hay học tập để điều chỉnh dạy và học thì người ta dùng đánh giá quá trình đào tạo (Formative assessment). Nếu đánh giá trước quá trình đào tạo hay học tập để nhằm mục đích phân loại, tổ chức sắp xếp lớp học thì người ta dùng đánh giá chẩn đoán (Diagnostic assessment).

Đánh giá cái gì?

Nói một cách chính xác hơn, người ta không đánh giá người học mà người ta đánh giá cái gì đó ở anh ta qua các hoạt động giáo dục diễn ra trong một hoàn cảnh nào đó. Chính vì vậy, người ta phải xác định trước nội dung cần đánh giá. Cần phải xác định là trong các kiến thức, năng lực và kỹ năng thì kiến thức, năng lực và kỹ năng nào quan trọng hơn, cần được đánh giá hơn.

Thực vậy, nếu đánh giá là tìm kiếm, thu thập một cách hệ thống các thông tin để đưa ra các quyết định thì đánh giá trong dạy và học ngoại ngữ là đánh giá trình độ, năng lực sử dụng ngoại ngữ nào đó trong một hoàn cảnh giao tiếp, trong một hoàn cảnh văn hoá xã hội cụ thể nào đó. Ngày nay, đánh giá trong ngôn ngữ được coi là xác đáng, là chuẩn phải nhằm vào đánh giá trình độ, năng lực sử dụng một ngoại ngữ nào đó để giao tiếp mà không phải nhằm vào đánh giá các

yếu tố ngôn ngữ biệt lập, tách rời, đánh giá hiệu quả sử dụng ngôn ngữ mà không phải là kiến thức ngôn ngữ hay kiến thức lý thuyết ngôn ngữ ấy. Nói một cách khác, đánh giá trong dạy và học ngoại ngữ phải nhằm mục đích đánh giá năng lực giao tiếp chứ không phải nhằm mục đích đánh giá kiến thức ngôn ngữ. Bởi vì mục đích cuối cùng của việc học ngoại ngữ nào đó là để *giao tiếp* bằng ngôn ngữ ấy mà *không phải là biết* ngôn ngữ ấy.

Cũng cần phân biệt *cái được đánh giá* với *phương pháp đánh giá*. Như vừa trình bày ở trên, cái được đánh giá ở đây là trình độ, năng lực sử dụng ngoại ngữ để giao tiếp, còn phương pháp đánh giá ở đây là công cụ được sử dụng để đo trình độ năng lực ấy. Một trắc nghiệm được coi là tốt phải được cấu thành *tối thiểu phương pháp đánh giá* và *tối đa cái được đánh giá* bởi vì cái ta cần đo là cái được đánh giá mà không phải là khả năng làm các trắc nghiệm.

Đánh giá như thế nào?

Một nguyên lý cơ bản trong đánh giá sư phạm là tính tương đẳng (congruence) giữa học và đánh giá. Theo Lussier [2], hoàn cảnh đánh giá, để được chấp nhận, phải bao gồm các đặc tính sau đây :

- Hoàn cảnh đánh giá phải *tương đẳng với mục tiêu học được đánh giá*. Điều này có nghĩa là hoàn cảnh đánh giá phải phản ánh được mục tiêu học; mục tiêu giao tiếp phải được coi trọng và các yếu tố học phải được xác định bằng các thông tin cần phải hiểu hay cần phải diễn đạt.

- Hoàn cảnh đánh giá phải *tương đẳng với phương pháp và kỹ thuật giảng dạy giao tiếp*. Cụ thể là đánh giá phải được đặt vào tình huống giao tiếp có thể chấp nhận được; nhiệm vụ mà người học sẽ thực hiện phải thích hợp; kỹ thuật đánh giá phải thích hợp và các loại tiểu mục được sử dụng cũng phải thích hợp.

- Hoàn cảnh đánh giá phải đầy đủ. Một bài công cụ đánh giá, hay một trắc nghiệm, hay một bài thi/kiểm tra phải được đặt vào một tình huống giao tiếp cụ thể; phải có nhiệm vụ để người học thực hiện ; phải có các chỉ dẫn đầy đủ và rõ ràng, phải xác định ngưỡng đạt, thang đo, đánh giá, v,v...

Trong đánh giá, người ta thường phân biệt *đánh giá tham chiếu tiêu chí* (Criterion Referenced Assessement) và *đánh giá tham chiếu qui chuẩn* (Normed Referenced Assessement).

Đánh giá tham chiếu tiêu chí là đánh giá kết quả học tập của người học so với các tiêu chí đã được xác định trước như mục tiêu hay chuẩn đầu ra của một quá trình đào tạo hoặc căn cứ vào điểm chuẩn đã được xác định trước. Ví dụ tốt nghiệp trường Đại học Ngoại ngữ - ĐHQGHN, sinh viên phải đạt trình độ C1 Khung tham chiếu Châu Âu về tiếng Anh hay tiếng Pháp tuỳ theo ngành học. Kết quả học tập này sau đó được dùng để đánh giá năng lực và khả năng làm chủ của người học.Ví dụ, mục đích của một kỳ thi hay một môn thi là đánh giá người học có khả năng phát âm đúng bằng tiếng Anh hoặc tiếng Pháp 8 màu sắc khác nhau nếu như người ta đưa cho anh ta một chiếc ảnh có mười chiếc túi với mười màu sắc khác nhau (ở đây ngưỡng đạt là 80%) và người học phát âm đúng 8/10 màu khác nhau thì anh ta được đánh giá là đạt. Kiến thức về màu sắc không có liên quan đến cách mà các học sinh khác thực hiện cùng một nhiệm vụ mà nó chỉ liên quan đến mục tiêu được đặt ra. Trong đánh giá tham chiếu tiêu chí dựa vào kỹ năng (in criterion referenced assessment in skill-based programs), người ta quan tâm nhiều hơn đến khả năng của người học có thể thực hiện được các nhiệm vụ phải thực hiện trong cuộc sống hàng ngày hay trong cuộc sống nghề nghiệp [4]. *Đánh giá tham chiếu tiêu chí* cho phép chia nhỏ một chương trình hay một nội dung học và mỗi một mục tiêu này có thể đo được. Người học và người dạy có thể biết được cái đã được dạy và được học như thế nào. Trong đào tạo, người ta ưu tiên và khuyến khích sử dụng *đánh giá tham chiếu tiêu chí* nhằm đánh giá kiến thức và kỹ năng mà người học đạt được so với mục tiêu đã xác định. Từ đó, người ta có thể nhận biết được các điểm mạnh và điểm yếu của người học và vì vậy sẽ giúp cho người học đạt mục tiêu học tập và có khả năng đảm nhiệm các nhiệm vụ trong cuộc sống hàng ngày và trong cuộc sống nghề nghiệp sau này.

Đánh giá tham chiếu qui chuẩn là đánh giá người học theo kết quả học tập hoặc đào tạo so với những người học khác cùng nhóm, hay cùng lớp, hay cùng khoá, vv. Ví dụ, mục đích của một kỳ thi hay một môn thi là đánh giá người học có khả năng phát âm đúng bằng tiếng Anh hoặc tiếng Pháp, khác với trong *Đánh giá tham chiếu tiêu chí*, người học không nhất thiết phải phát âm đúng 8 màu sắc khác nhau thì mới đạt mà anh ta có thể chỉ cần phát âm đúng 4 màu khác nhau hoặc ít hơn, anh ta vẫn đạt nếu những người học khác cùng nhóm, hay cùng lớp, hay cùng khoá, vv. phát âm đúng số âm ít hơn số âm mà anh ta phát âm đúng. Anh ta sẽ không đạt nếu như số âm mà anh ta phát âm đúng ít hơn những người học khác cùng nhóm, hay cùng lớp, hay cùng khoá, vv. *Đánh giá tham chiếu qui chuẩn* cho phép phân biệt các trình độ khác nhau giữa người học, nó đặc biệt phù hợp và có ích cho việc xếp hạng để lựa chọn đối với những trường hợp phải tuyển lựa khắt khe, ví dụ như thi tuyển sinh đại học chẳng hạn.

Tóm lại, nếu mục đích chính của *đánh giá tham chiếu tiêu chí* là mô tả cái mà người học làm được thì *đánh giá tham chiếu qui chuẩn* có mục đích phân loại người học trong cùng một nhóm, hay cùng một lớp, hay cùng một khoá, vv.

Độ tin cậy và tính hiệu lực

Nói đánh giá như thế nào không thể không nói đến công cụ đánh giá trong đó đặc biệt là *trắc nghiệm*. Trắc nghiệm là công cụ đánh giá quan trọng và phổ biến nhất. Bởi vì đánh giá và trắc nghiệm có quan hệ mật thiết hữu cơ với nhau. Tuy nhiên, bản thân trắc nghiệm không có *chức năng đánh giá*, mà chính xác hơn, trắc nghiệm chỉ có *chức năng đo* [5]. Người ta chỉ nói đến đánh giá khi *trắc nghiệm*, chính xác hơn là kết quả của trắc nghiệm, được sử dụng làm cơ sở để đưa ra các quyết định [1]. Vì vậy, để thực hiện được *Chuẩn đánh giá* phải có trắc nghiệm tốt, trắc nghiệm "chuẩn" và phải biết chọn lựa và sử dụng nó một cách thích hợp. Vậy, một trắc nghiệm tốt hay "chuẩn" là một trắc nghiệm như thế nào? Một trắc nghiệm được coi là *tốt* hay "*chuẩn*" phải là một trắc nghiệm *có khả năng đo được cái cần đo*. Để *đo được cái cần đo* và để *đánh giá được cái cái cần đánh giá*, trước hết trắc nghiệm phải có *độ tin cậy* (reliability) và *tính hiệu lực* (validity) cao. Thật vậy, độ *tin cậy* và *tính hiệu lực* là hai đặc tính cơ bản và quan trọng nhất của một trắc nghiệm [6].

Độ tin cậy của một trắc nghiệm được thể hiện ở tính ổn định và không thay đổi của kết quả trắc nghiệm. Một trắc nghiệm được coi là có *độ tin cậy* phải đạt được các tiêu chí sau đây:

- Trong hai lần kiểm tra/thi khác nhau, cùng một người học sẽ đạt điểm xấp xỉ hoặc trùng nhau nếu làm cùng một nội dung kiểm tra/thi và người học này sẽ không được học thêm gì liên quan đến nội dung kiểm tra/thi (Test-retest).

- Hai trắc nghiệm với hình thức khác nhau, nhưng cùng một lĩnh vực sẽ cho các kết quả giống nhau nếu đo cùng một cái định đo (Parall Forms).

- Các câu hỏi hay tiểu mục của một trắc nghiệm phải liên kết chặt chẽ với nhau và đo cùng một bình diện (Internal consistency).

- Hai giám khảo chấm cùng một bài cho hai điểm giống nhau hoặc gần giống nhau (Inter-rater).

- Một giám khảo chấm cùng một bài cho điểm giống nhau hoặc gần giống nhau giữa hai lần chấm khác nhau (Intra-rater).

Tính hiệu lực của một trắc nghiệm được thể hiện ở khả năng đo được cái muốn đo. *Tính hiệu lực* là phẩm chất quan trọng nhất của một trắc nghiệm. Nó cho phép đánh giá hoặc đưa ra các quyết định đúng đắn. Một trắc nghiệm được coi là có tính hiệu lực phải đạt được các tiêu chí sau đây:

- Trắc nghiệm phải là mẫu đại diện cái được dự định đo (Content validity).

- Kết quả của hai trắc nghiệm khác nhau, nhưng có cùng nhiệm vụ đánh giá một kỹ năng hay kiến thức nào đó phải giống nhau hoặc gần giống nhau (Criterion validity).

- Các câu hỏi của một trắc nghiệm phải phản ánh được các nguyên lý của lý luận học ngoại ngữ (Construct validity).

- Trắc nghiệm phải cho cảm giác đo cái cần được đo (Apparent validity).

Độ tin cậy và *tính hiệu lực* là hai đặc tính quan trọng và chủ yếu nhất của một trắc nghiệm. Thiếu một trong hai đặc tính này trắc nghiệm sẽ

không hoàn thành được chức năng của mình và sẽ không có giá trị. Trong mọi hoàn cảnh, hai đặc tính đặc biệt quan trọng này cho phép ta quyết định có sử dụng trắc nghiệm hay không. *Độ tin cậy* bảo đảm *chất lượng* của một trắc nghiệm, trong khi *tính hiệu lực* cho phép khẳng định một trắc nghiệm có được *sử dụng* hay không.

Chỉ số khó và chỉ số phân loại

Một công cụ đánh giá hay một trắc nghiệm được coi là tốt, là chuẩn thì công cụ đánh giá hay trắc nghiệm đó phải có *chỉ số khó* (Index of difficulty hay Degre of difficulty) và *chỉ số phân loại* (Index of discrimination) thích hợp. Hai chỉ số quan trọng này cho phép xác định độ khó hoặc dễ và độ phân loại của một trắc nghiệm, tức là chất lượng và hiệu quả của một trắc nghiệm. Người ta nói nhiều đến hai chỉ số này, song việc xác định và sử dụng hai chỉ số này vào đánh giá không phải bao giờ cũng dễ dàng đối với nhiều giáo viên ngoại ngữ và ngay cả đối với một số người được giao nhiệm vụ thiết kế và xây dựng đề thi/kiểm tra.

Xác định được *chỉ số khó* có một ý nghĩa quan trọng. *Chỉ số khó* chính là tỷ lệ thí sinh hay người học của một nhóm hay một lớp hoàn thành nhiệm vụ do một tiểu mục đòi hỏi. Để xác định được *chỉ số khó* người ta thường dùng công thức tính sau đây:

$$P = \frac{R}{N}$$

P = *chỉ số khó* hay tỷ lệ đạt của một tiểu mục

R = số thí sinh trả lời đúng tiểu mục

N = tổng số thí sinh tham gia trả lời tiểu mục

Ví dụ, trong tổng số 100 thí sinh có 25 thí sinh trả lời đúng tiểu mục, chỉ số khó của tiểu mục này sẽ là :

$$P = \frac{40}{100} = 0,4$$

Độ khó của tiểu mục này là vừa phải. Có nghĩa là tiểu mục này không quá khó và cũng không quá dễ. Một ví dụ khác: nếu trong số 100 thí sinh tham gia trả lời một tiểu mục chỉ có 10 thí sinh trả lời đúng, *chỉ số khó* của tiểu mục sẽ là 0,1. Tiểu mục này là quá khó. Như vậy, *chỉ số khó càng nhỏ thì tiểu mục càng khó* và ngược lại *chỉ số khó càng lớn thì tiểu mục càng dễ*. Trong kiểm tra - đánh giá nói chung, mục tiêu là phân loại các sinh viên giỏi với các sinh viên kém hoặc yếu, giá trị của chỉ số khó của các tiểu mục không nên tiến gần đến hai cực (0 và 1). Khi mà một kỳ thi có mục đích chọn một số ít thí sinh giỏi hoặc rất giỏi trong số rất đông thí sinh, người ta tăng độ khó để giảm *chỉ số khó* của các tiểu mục.

Trong *đánh giá tham chiếu qui chuẩn*, *chỉ số khó* mong muốn của các tiểu mục dao động trong khoảng 0,3 đến 0,7. Tuy nhiên, có thể *chỉ số khó* rất nhỏ, có nghĩa là tiểu mục rất khó, nhưng tiểu mục này vẫn có thể sử dụng được vì việc sử dụng một tiểu mục hay một trắc nghiệm còn tuỳ thuộc vào mục đích của đánh giá. Song lý tưởng nhất trong một bài trắc nghiệm liên quan đến *chỉ số khó* của các tiểu mục là những thí sinh giỏi nhất sẽ trả lời đúng và những thí sinh kém nhất sẽ trả lời sai hoặc không trả lời được. Và trong một bài trắc nghiệm hay một bài thi/kiểm tra phải có cả các tiểu mục dễ, các tiểu mục khó trung bình và các tiểu mục khó. Theo Morisette [7], nếu điểm qua của một môn học là 60%, bài thi/kiểm tra phải có các tiểu mục dễ (85% người học có thể trả lời đúng), các tiểu mục khó trung bình (55 - 85% người học có thể trả lời đúng) và các tiểu mục khó (40 - 55% người học có thể trả lời đúng).

C*hỉ số phân loại* cho phép phân loại các sinh viên đã đạt và các sinh viên chưa đạt được mục tiêu đào tạo. Một tiểu mục được coi là có chất lượng phải có chỉ số phân loại tương ứng với hoàn cảnh *đánh giá tổng kết* (Summative assessment) hay *đánh giá tham chiếu qui chuẩn* (Normed Referenced Assessement). Để kiểm tra được giá trị của chỉ số phân loại, có ba bước sau đây:

Bước một nhằm bảo đảm bài thi/kiểm tra phải tương ứng trong tổng thể với mục tiêu đã được xây dựng trước trong bảng ma trận đề thi (tableau de spécification). Một đề thi có chỉ số phân loại tốt phải là một đề thi chứa các tiểu mục đi theo hướng của bài thi hay kỳ thi : trong một bài thi/kiểm tra, người học hay thí sinh giỏi hơn sẽ đạt được kết quả cao hơn ở các tiểu mục và

ngược lại người học hay thí sinh kém đạt được kết quả thấp hơn hoặc có những tiểu mục không trả lời được.

Trong bước hai, người ta tính chỉ số phân loại của các tiểu mục. Để thiết lập chỉ số phân loại của các tiểu mục, người ta so sánh các câu trả lời của các thí sinh của nhóm giỏi nhất (ví dụ 20% : Ns) và các câu trả lời của nhóm sinh viên kém nhất (ví dụ 20 % : Ni). Người ta tính số thí sinh của nhóm giỏi nhất (Rs) và số thí sinh của nhóm kém nhất (Ri) trả lời đúng. Chỉ số phân loại sẽ được xác định dựa trên kết quả so sánh giữa tỷ lệ trả lời đúng tiểu mục của nhóm giỏi nhất và nhóm kém nhất. Chỉ số phân loại được tính theo công thức sau đây :

$$D = \frac{Rs}{Ns} - \frac{Ri}{Ni} = \frac{Rs - Ri}{Ns}$$

D = chỉ số phân loại

Rs = số thí sinh trong nhóm giỏi nhất (20%) trả lời đúng tiểu mục

Ri = số thí sinh trong nhóm kém nhất (20%)

Ni = Ns = số thí sinh trong 20% giỏi nhất hoặc 20% kém nhất

Ví dụ, trong số 175 thí sinh,

20% số thí sinh đạt điểm cao nhất (Ns) = 20% của 175 = 35

20% số thí sinh đạt điểm thấp nhất (Ni) = 20% của 175 = 35

Trong 36 thí sinh đạt điểm cao nhất của toàn bài, có 30 thí sinh trả lời đúng (Rs = 30) *tiểu mục x* chẳng hạn và trong 36 thí sinh đạt điểm thấp nhất của toàn bài, có 9 thí sinh trả lời đúng (Ri = 9) *tiểu mục x* này, chỉ số phân loại được tính như sau:

$$D = \frac{30 - 9}{35} = 0,6$$

Bước ba là bước đánh giá chỉ số phân loại. Nếu chỉ số phân loại đi theo hướng của toàn bộ bài thi/kiểm tra, giá trị của nó nằm giữa 1 và 0 (1 ≥ D > 0). *Chỉ số phân loại* càng gần 1, thì độ phân loại càng lớn. Nếu chỉ số phân loại D = 0 thì số thí sinh giỏi và số thí sinh kém trả lời đúng là ngang nhau. Lý tưởng là mỗi tiểu mục được xây dựng, điều chỉnh sao cho giá trị của chỉ số phân loại đạt tới 1, tuy nhiên trong thực tế điều này rất khó đạt được, thậm chí không thể đạt được. Theo Morisette [7], nếu chỉ số phân loại của một tiểu mục nằm giữa khoảng +1 và +0,3 (+1 ≥ D > +0,3), độ phân loại của tiểu mục là tích cực, nếu chỉ số phân loại nằm giữa khoảng +0,29 và +0,1 (+0,29 ≥ D > +0,1), độ phân loại của tiểu mục là kém tích cực. Nếu chỉ số phân loại của một tiểu mục nằm giữa khoảng 0 và -1 (0 ≥ D > -1), độ phân loại là tiêu cực. Có nghĩa là số thí sinh giỏi có thể trả lời sai, ngược lại số thí sinh kém có thể trả lời đúng. Như vậy, nhất thiết tiểu mục này phải được xem xét lại.

Tuy nhiên, theo Morisette [7], cũng cần lưu ý một số điểm sau:

- Nếu như mục tiêu của trắc nghiệm không phải là phân loại người học mà là kiểm tra chất lượng học, đào tạo thì tính xác đáng của tiểu mục được ưu tiên hơn.

- Nếu tất cả thí sinh đều trả lời đúng hoặc đều trả lời sai, thì chỉ số phân loại D = 0. Mặt khác, nếu *chỉ số khó* của tiểu mục càng tiến gần đến 0,5 thì khả năng *chỉ số phân loại* sẽ lớn.

- Cùng một tiểu mục, chỉ số phân loại có thể khác nhau ở các kỳ thi khác nhau, nhất là với những kì thi có số thí sinh ít. Do vậy việc giải thích *chỉ số phân loại* cũng phải rất thận trọng.

- Cần phải chú ý, thận trọng xem xét từng tiểu mục, nhất là trước khi quyết định giữ lại hoặc loại bỏ một tiểu mục vì *chỉ số phân loại* của mỗi tiểu mục phải được xem xét và đánh giá trong tổng thể một bài thi/kiểm tra hay một trắc nghiệm.

Chọn công cụ đánh giá thích hợp

Cuối cùng, việc lựa chọn công cụ đánh giá phù hợp cũng đóng vai trò đặc biệt quan trọng trong việc đánh giá chính xác và hiệu quả từng cấp độ nhận thức cũng như từng loại kiến thức và kỹ năng [8]. Thật vậy, các mức độ khó và phức tạp của mục tiêu cũng như nội dung đánh giá cũng đòi hỏi các loại công cụ đánh giá khác nhau.

Ví dụ để đánh giá năng lực nhận biết hay hiểu của người học hay thí sinh thì chỉ cần trắc nghiệm khách quan với loại câu hỏi nhiều lựa chọn chẳng hạn, nhưng nếu muốn đánh giá năng lực phân tích, tổng hợp hay đánh giá của của người học hay thí sinh thì loại trắc nghiệm khách quan sẽ ít hiệu quả. Quan sát *Bảng 1* và *Bảng 2* do Albernot [8] đề nghị dưới đây, ta dễ dàng nhận thấy mục tiêu càng khó và càng phức tạp thì công cụ đánh giá càng "mở" dần. Trả lời của thí sinh sẽ tăng từ nguyên bản đến sáng tạo, tức là từ nhớ, thuộc lòng đến sáng tạo và độc đáo. Tư duy từ đồng nhất đến đa dạng, phong phú. Đánh giá từ đánh giá định lượng chuyển sang đánh giá định tính. Đầu tư của cả người dạy và người học cũng tăng dần theo độ khó và phức tạp của mục tiêu. Về góc độ sư phạm, cũng phát triển từ cái được dạy đến cái được rèn luyện.

Bảng 1[(1)]. Độ phức tạp của mục tiêu và độ mở của công cụ đánh giá

Mục tiêu khó dần	**Công cụ đánh giá mở dần**
1. Biết (nhớ)	- câu hỏi nhiều lựa chọn
2. Hiểu	- câu hỏi truyền thống
3. Ứng dụng	- các loại bài tập khác nhau
4. Phân tích	- vấn đề
5. Tổng hợp	- chủ đề tổng hợp
6. Đánh giá	- nghị luận
	- sáng tạo

Bảng 2. Độ phức tạp của mục tiêu và gia tăng của một số chỉ số

Mục tiêu	**Công cụ**	**Trả lời**	**Tư duy**	**Đánh giá**	**Đầu tư**	**Sư phạm**
Sơ đẳng	Đóng	Nguyên bản	Đồng nhất	Định lượng	Ít	Cái được dạy
Biết (nhớ)						
Hiểu						
Ứng dụng						
Phân tích						
Tổng hợp						
Đánh giá						
Phức tạp	Mở	Sáng tạo, độc đáo	Không đồng nhất	Định tính	Nhiều	Cái được rèn luyện

Mặt khác, để đánh giá hiệu quả mỗi loại kiến thức hay kỹ năng cần lựa chọn công cụ đánh giá hay trắc nghiệm thích hợp [9], ví dụ đánh giá kỹ năng nghe hiểu hay đọc hiểu thì sử dụng *trắc nghiệm khách quan* sẽ phù hợp hơn. Ngược lại, nếu đánh giá kỹ năng diễn đạt viết thì sử dụng *trắc nghiệm tự luận* sẽ tốt hơn. Để chọn được một công cụ đánh giá thích hợp và hiệu quả, cần phải hiểu rõ các ưu điểm và nhược điểm cơ bản của từng công cụ đánh giá. Trong khuôn khổ bài viết này, chúng tôi chỉ đề cập một số ưu điểm và nhược điểm chủ yếu của *trắc nghiệm chấm khách quan* (test to objective correcting) và được gọi tắt là *trắc nghiệm khách quan* (Objective Test) và của *trắc nghiệm chấm chủ quan* (test to subjective correcting) và được gọi tắt là *trắc nghiệm tự luận* (Subjective Test). Đây là hai loại trắc nghiệm chủ yếu và thường dùng nhất trong đánh giá nói chung và trong đánh giá ngoại ngữ nói riêng.

(1) Bảng 1 và 2 do Albernot (1996) đề nghị và có sự điều chỉnh của tác giả.

Trắc nghiệm khách quan được thể hiện dưới 4 dạng sau: trắc nghiệm với câu hỏi nhiều lựa chọn, trắc nghiệm điền khuyết, trắc nghiệm ghép các yếu tố (từ, nhóm từ, câu, đoạn, vv.) và trắc nghiệm đúng hay sai và không thể trả lời được. Trắc nghiệm tự luận bao giờ cũng đi với các câu hỏi mở và được thể hiện dưới 2 dạng sau: trắc nghiệm với câu trả lời ngắn và trắc nghiệm với câu trả lời dài (giống như tiểu luận hay luận văn) do chính thí sinh phải xây dựng bằng ngôn ngữ của riêng mình. Mỗi loại trắc nghiệm đều có những điểm mạnh và những điểm yếu riêng. Thực ra, có thể nói điểm mạnh của loại trắc nghiệm này thường là điểm yếu của loại trắc nghiệm kia và ngược lại. Ưu điểm cơ bản của trắc nghiệm khách quan là có khả năng kiểm tra được một khối lượng kiến thức rộng, nhưng không đi sâu vào từng vấn đề, hiệu quả đánh giá được kiến thức của người học. Loại trắc nghiệm này cho phép đánh giá cấp độ nhận biết hay ghi nhớ (knowledge), hiểu (comprehension) và ứng dụng (application). Do vậy, trắc nghiệm khách quan khuyến khích người học tìm cách thu nhận kiến thức hơn là rèn luyện kỹ năng. Ngược lại, trắc nghiệm tự luận tuy không bao quát được một khối lượng kiến thức rộng nhưng lại đi sâu được vào từng vấn đề. Trắc nghiệm tự luận hiệu quả và thích hợp hơn trong việc đánh giá các trình độ cao và phức tạp hơn như là phân tích (analysis), tổng hợp (synthesis) và đánh giá (evaluation) và đặc biệt hiệu quả trong việc đánh giá năng lực sáng tạo của người học. Do đó, trắc nghiệm tự luận khuyến khích người học kỹ năng tổng hợp, phân tích, đánh giá và diễn đạt [9]. Chính vì vậy, theo các chuyên gia đánh giá, khi đánh giá một số kỹ năng nên kết hợp hai loại trắc nghiệm trên hay hai loại tiểu mục (câu hỏi) chấm khách quan và chấm chủ quan để tận dụng tối đa điểm mạnh và loại bớt tối đa các nhược điểm của mỗi loại công cụ đánh giá, vì chúng sẽ bù trừ cho nhau, ví dụ trong đánh giá kỹ năng đọc hiểu và nghe hiểu chẳng hạn.

Trên đây chúng tôi vừa trình bày một số vấn đề cơ bản nhất liên quan đến *đánh giá* và *chuẩn đánh giá*. Hiểu được *chuẩn đánh giá* vốn không phải dễ dàng, song để thực hiện được *chuẩn đánh giá* còn khó hơn nhiều. Thật vậy, để thực hiện được *chuẩn đánh giá* trong dạy và học ngoại ngữ phải nắm chắc và thực hiện tốt một loạt các vấn đề và nhiệm vụ vừa trình bày trên đây. Nói một cách chính xác hơn là phải trả lời được các câu hỏi sau đây: "Tại sao đánh giá?", "Khi nào đánh giá?", "Đánh giá cái gì?", "Đánh giá như thế nào?" và "Chọn loại công cụ đánh giá hay trắc nghiệm gì?", "Làm thế nào để xây dựng được một trắc nghiệm có độ tin cậy cao và tính hiệu lực cao, có chỉ số khó thích hợp và chỉ số phân loại tích cực?".

Tài liệu tham khảo

[1] C.H. Weiss, Evaluating action programs - Readings in social action and education, Allyn and Bacon Inc, Boston, 1972.

[2] D. Lussier, É*valuer selon une démarche communicative/expérientielle*, Centre Educatif et Culturel Inc, Québec (Canada),1991.

[3] B.S. Bloom, et al., Taxonomy of Educational Objectives. Handbook I, The CognitiveDomain, David Mackay Co, New York, 1956.

[4] G. Scallon, *L'évaluation des apprentissages dans une approche par compétences*, Éditions du Renouveau pédagogique Inc, Québec (Canada), 2004.

[5] Nguyen Quang Thuan, Thiết kế và xây dựng công cụ đánh giá kỹ năng nghe hiểu, *Tạp chí Khoa học, Đại học Quốc gia Hà Nội, T. XXI* 1(2005)47.

[6] L.F. Bachman, & A.-S. Palmer, Language Testing in Practice: Designing and Developing Useful Language Tests, Oxford University Press, Oxford, 1997.

[7] D. Morissette, *Évaluation sommative*, Éditions du Renouveau Pédagogique Inc, Québec (Canada), 1996.

[8] Y. Abernot, *Les méthodes d'évaluation scolaire*, 2è éd., DUNOD, Paris, 1996.

[9] Nguyen Quang Thuan, Xây dựng một công cụ kiểm tra - đánh giá trong dạy và học ngoại ngữ, *Tạp chí Khoa học, Đại học Quốc gia Hà Nội, T. XVIII* 2(2002)23.

Assessment standard in foreign language teaching and learning

Nguyen Quang Thuan

Centre for Distance Education and Teacher Development, University of Languages and International Studies, Vietnam National University, Hanoi, Pham Van Dong street, Cau Giay, Hanoi, Vietnam

The term "standard" has been mentioned these years: Standard of knowledge, standard of curriculum, etc. In assessment in general and in foreign language teaching in particular, *assessment standard* is specially emphasized. However, in language teaching and learning, there has been no clear-cut understanding in *assessment standard* in terms of both theory and practice, particularly how to carry out *assessment standard* is arguable. In this article, we have tried to point out the most basic issues in relation to assessment such as the target, content, methods and devices of assessment. Also, some types of tests of multiple choice as well as their important features are insightfully discussed. As a result, hopefully, the viewpoint and content of *assessment standard* are brought out in teaching and learning language.

Key Words: Assessment, Assessment Standard, Test, Criterion Referenced Assessement, Normed Referenced Assessement, Reliability, Validity, Index of difficulty, Index of discrimination).