PHÂN TÍCH TÍNH GIÁ TRỊ HAI BÀI KIỂM TRA NGỮ PHÁP – TỪ VỰNG HỌC PHẦN 2A + 2B TẠI KHOA NGÔN NGỮ VÀ VĂN HÓA PHÁP, TRƯỜNG ĐẠI HỌC NGOẠI NGỮ - ĐẠI HỌC QUỐC GIA HÀ NỘI

Đỗ Thị Bích Thủy*

Khoa Ngôn ngữ và Văn hóa Pháp, Trường Đại học Ngoại ngữ, ĐHQGHN,
Phạm Văn Đồng, Cầu Giấy, Hà Nội, Việt Nam

Nhận bài ngày 07 tháng 09 năm 2018
Chỉnh sửa ngày 23 tháng 11 năm 2018; Chấp nhận đăng ngày 27 tháng 11 năm 2018

Tóm tắt: Bài viết này trình bày kết quả của một nghiên cứu về tính giá trị của hai bài kiểm tra Ngữ pháp - Từ vựng học phần 2A + 2B năm học 2016-2017 tại Khoa Ngôn ngữ và Văn hóa Pháp, Trường Đại học Ngoại ngữ, Đại học Quốc gia Hà Nội (ĐHNN-ĐHQGHN). Nghiên cứu đánh giá độ tương thích giữa hai bài kiểm tra này với bản mô tả kĩ thuật bài kiểm tra; đo chỉ số độ khó của từng tiểu mục trong bài kiểm tra; và đo một số thông số chung của toàn bài kiểm tra. Kết quả cho thấy hai bài kiểm tra đều đảm bảo tính giá trị, bài kiểm tra số 2 có tính giá trị cao hơn bài số 1. Tuy nhiên, cần điều chỉnh cấp độ ngôn ngữ bài kiểm tra cho phù hợp hơn với bản mô tả kĩ thuật và chỉnh sửa lại những tiểu mục có chỉ số độ khó chưa phù hợp.**

Từ khóa: kiểm tra đánh giá, bài kiểm tra Ngữ pháp - Từ vựng, tính giá trị, bản mô tả kĩ thuật, chỉ số độ khó

1. Dẫn nhập

Song song với công tác giảng dạy, kiểm tra đánh giá năng lực của sinh viên/ học sinh là một hoạt động quan trọng của giảng viên/ giáo viên (Dương Thu Mai, Nguyễn Thị Chi & Phạm Thị Thu Hà, 2017). Những nghiên cứu gần đây nhất về công tác kiểm tra đánh giá có thể được chia thành bốn mảng chính. Mảng thứ nhất bao gồm những nghiên cứu tập trung đo lường hiểu biết và quan điểm của giáo viên các cấp về kiểm tra đánh giá (Crombs, DeLuca, LaPointe-McEwan & Chalas, 2018; Issaieva & Crahay, 2010; Lê Thị Huyền Trang & Trần Thị Tuyết, 2015; Remesal, 2011). Mảng thứ hai là những nghiên cứu về các phương pháp kiểm tra đánh giá và tự đánh giá kết hợp ứng dụng công nghệ như e-portfolio hay dùng các phần mềm để soạn thảo các đề thi trực tuyến (Hooker, 2017; Nguyễn Văn Long, 2017). Mảng thứ ba tập trung vào xây dựng khung năng lực kiểm tra đánh giá cho giáo viên hay phát triển các bản mô tả kĩ thuật đề thi (Dương Thu Mai et al., 2017; Herppich et al., 2017; Hoàng Hồng Trang, Nguyễn Thị Chi & Dương Thu Mai, 2016). Cuối cùng là những nghiên cứu phân tích các bài kiểm tra nhằm có những điều chỉnh phục vụ cho quá trình dạy, học và khảo thí (Đỗ Quang Việt, 2014; El Allaoui, Rhazi Filali, El Hadri, Fetteh & Bouhadi, 2016; Nguyễn Thị Ngọc Quỳnh, 2018; Nguyễn Thị Phương Thảo, 2018; Nguyễn Thị Quỳnh Yến, 2016). Bài viết này nằm trong nhóm các nghiên cứu cuối cùng, với mục tiêu đo tính giá trị của hai bài kiểm tra Ngữ pháp - Từ vựng (KTNPTV) học

* ĐT.: 84-976062007
Email: dbthuy2003@gmail.com

** Nghiên cứu này được hoàn thành với sự hỗ trợ của Trường Đại học Ngoại ngữ, Đại học Quốc gia Hà Nội trong đề tài mã số N.17.04.

phần 2A + 2B năm học 2016-2017 tại Khoa Ngôn ngữ và Văn hóa Pháp, Trường ĐHNN-ĐHQGHN.

1.1. Tính giá trị

Tính giá trị (validity) được coi là thuộc tính quan trọng nhất để đánh giá chất lượng một bài kiểm tra (Bachman, 1990). Theo quan điểm truyền thống trước đây, một bài kiểm tra có giá trị khi nó đo được cái cần đo (Hughs, 2003; Lissitz, 2009). Tuy nhiên, cách định nghĩa này đã bị nhiều nhà nghiên cứu phê bình vì chủ yếu chỉ tập trung vào bài kiểm tra mà không tính tới nhiều yếu tố nhận thức-xã hội khác. Sau này, Messick (1989) đã đưa ra một khái niệm về tính giá trị hợp nhất như sau: "*tính giá trị là một nhận xét tổng hợp về mức độ những căn cứ khoa học và cơ sở lý thuyết có thể chứng minh sự đúng đắn và phù hợp của các nhận định đánh giá về năng lực và của các hành động có liên quan tới kết quả đánh giá*" (Messick, 1989, tr.13). Quan điểm thứ ba là quan điểm của các nhà lý thuyết hiện đại, cho rằng tính giá trị là một khái niệm tương đối, gồm nhiều mặt và cần thu thập nhiều loại bằng chứng khác nhau để chứng minh tính giá trị của một bài kiểm tra (Weir, 2005). Ví dụ, một bài kiểm tra có thể có tính giá trị nội dung cao vì bám sát bản mô tả kĩ thuật, nhưng tính giá trị trong việc chấm thi lại thấp.

Dưới đây là các khía cạnh của tính giá trị theo Messick (1990).

Bảng 1. Các khía cạnh của tính giá trị (Messick, 1990 - Bản dịch của Dương Thu Mai et al., 2017)

Khía cạnh	**Căn cứ xác định giá trị**
Nội dung (content)	Sự liên quan và tính đại diện của nội dung dùng để đo lường so với khái niệm cấu trúc đang được đo.
Kết cấu (structural)	Mối quan hệ tương quan của các phần hoặc nội hàm: cấu trúc trong. Mối quan hệ tương quan giữa các thang chấm và khung năng lực cần đo. Mối quan hệ tương quan của mức năng lực đo được với các kết quả đánh giá (ĐG) khác hoặc với các đặc điểm khác của người học: cấu trúc ngoài.
Quá trình (substantive)	Sự hợp lý và đầy đủ của quá trình thực hiện năng lực của người học.
Khái quát hóa (generalizability)	Những điểm giống và khác nhau trong quá trình thực hiện năng lực trong những lần ĐG khác nhau. Độ dao động của kết quả sau khi người học được hướng dẫn thêm.
Ngoại vi (external)	Mối quan hệ giữa các điểm số ĐG năng lực này và kết quả ĐG các năng lực tương tự hoặc năng lực khác.
Hệ quả (consequence)	Ý nghĩa sử dụng của các nhận định về điểm số, có xét tới các mục đích ĐG ban đầu.

Các nhà nghiên cứu quan tâm tới tính giá trị của các bài kiểm tra năng lực giao tiếp (Nghe, Nói, Đọc, Viết) thường áp dụng khung đo tính giá trị theo đường hướng nhận thức-xã hội (socio-cognitive framework) của Weir (2005). Mô hình lý thuyết này có tên gọi là nhận thức-xã hội vì: (1) năng lực cần đánh giá phải gắn với quá trình xử lý thông tin tại não bộ thí sinh, và (2) tập trung vào việc đánh giá khả năng sử dụng ngôn ngữ trong một tình huống có thực trong xã hội chứ không chỉ tập trung vào mặt ngôn ngữ. Khung đo tính giá trị theo Weir (2005) lấy thí sinh là trung tâm và bao gồm 6 yếu tố:

- Đặc điểm của thí sinh (test taker characteristics): đặc điểm sinh lý, tâm lý, trải nghiệm sống.

- Tính giá trị ngữ cảnh (context validity): tình huống đưa vào bài thi phải là một tình huống gần nhất có thể với tình huống giao tiếp thực; tình huống này phải cho phép đánh giá được năng lực cần kiểm tra.

- Tính giá trị nhận thức (theory-based or cognitive validity): gắn với quá trình xử lý thông tin tại não bộ của người dự thi.

- Tính giá trị khi chấm thi (scoring validity): bao gồm các yếu tố ảnh hưởng tới việc chấm điểm.

- Tính giá trị ảnh hưởng (consequential validity): ảnh hưởng của bài thi và điểm số tới thí sinh, hệ thống giáo dục và xã hội, hay nói cách khác "tác động dội ngược" của bài thi (Bailey, 1996; Nguyễn Thúy Lan, 2017; Đỗ Thị Bích Thủy, 2018).

- Tính giá trị tiêu chuẩn (criterion-related validity): đo độ tin cậy của điểm số ở những lần thi khác nhau của cùng một kì thi; so sánh kết quả một đề thi khi cho thi ở những lần khác nhau; so sánh với những bài thi khác.

Theo Weir (2005), để đánh giá tính giá trị, có thể thu thập minh chứng trước khi kiểm tra (so sánh độ tương thích giữa bài kiểm tra với bản mô tả kĩ thuật) và sau khi kiểm tra (phân tích kết quả điểm số). Đánh giá tính giá trị của bài kiểm tra/ thi là trách nhiệm của những tác giả biên soạn đề (Messick, 1995); tuy nhiên vì nhiều lí do, rất ít khi các tác giả soạn đề đánh giá tính giá trị của các đề thi (Nguyễn Thị Ngọc Quỳnh, 2018). Ở Việt Nam, việc đo lường tính giá trị bài thi được làm ở một số đề thi cấp quốc gia.

Tùy thuộc vào tính chất của bài kiểm tra, các nhà nghiên cứu sẽ chọn một số thông số khác nhau để đo tính giá trị. Ví dụ, khi đo tính giá trị một bài thi cuối kì Sinh học, nhà nghiên cứu đã chọn các minh chứng sau: so sánh với bản mô tả kĩ thuật, tính chỉ số độ khó, tính chỉ số phân loại, tính điểm trung bình, độ lệch chuẩn (El Allaoui et al., 2016). Còn khi đo tính giá trị bài thi Viết VSTEP, tác giả đo điểm trung bình, độ lệch chuẩn, độ xiên, độ tương quan giữa các lần chấm thi, tính nhất quán của người chấm thi (Nguyễn Thị Ngọc Quỳnh, 2018; Nguyễn Thị Quỳnh Yến, 2016). Với bài Đọc hiểu VSTEP, nhà nghiên cứu lại chọn so sánh với bản mô tả kĩ thuật, chỉ số độ khó, điểm trung bình, độ lệch chuẩn, độ xiên (Nguyễn Thị Phương Thảo, 2018).

Theo *Hướng dẫn quy trình biên soạn và phát triển ngân hàng câu hỏi chuẩn hóa thi kết thúc các học phần thuộc khối kiến thức chung trong chương trình đào tạo đại học* số 3289/HD-ĐHQGHN do Đại học Quốc gia Hà Nội ban hành ngày 28/8/2017, quy trình biên soạn ngân hàng câu hỏi thi gồm 10 bước trong đó bước 7 là thử nghiệm câu hỏi, bước 8 là phân tích câu hỏi sau khi thử nghiệm, bước 9 là chỉnh sửa câu hỏi sau khi thử nghiệm (chỉnh sửa, tăng giảm độ khó, loại bỏ câu hỏi). Thực trạng hiện nay là quy trình biên soạn đề thi ở phần lớn các trường chỉ dừng đến khâu cho sinh viên làm bài thi, còn phần phân tích kết quả thi để có những điều chỉnh nhằm cải thiện quá trình dạy và học rất ít khi được thực hiện ở quy mô tổ bộ môn, khoa.

1.2. Câu hỏi nghiên cứu

Bài viết này có mục tiêu là đo tính giá trị hai bài KTNPTV học phần 2A + 2B tại Khoa Ngôn ngữ và Văn hóa Pháp, ĐHNN-ĐHQGHN. Câu hỏi nghiên cứu đặt ra là:

- Hai bài KTNPTV học phần 2A + 2B có tuân thủ các yêu cầu trong bản mô tả kĩ thuật hay không?

- Những tiểu mục nào cần chỉnh sửa độ khó/ độ dễ?

- Kết quả thống kê mô tả cho thấy điểm trung bình, độ lệch chuẩn và độ xiên của hai bài KTNPTV đã phù hợp với một bài kiểm tra ngoại ngữ thường xuyên hay chưa?

Hai bài KTNPTV này đã được cho kiểm

tra tại Khoa Ngôn ngữ và Văn hóa Pháp năm học 2016-2017 với 89 em sinh viên (bài 1) và 88 em sinh viên (bài 2). Để trả lời câu hỏi nghiên cứu, chúng tôi lựa chọn đo tính giá trị bài KTNPTV bằng các thông số sau: so sánh với bản mô tả kĩ thuật (minh chứng trước khi kiểm tra), đo chỉ số độ khó, điểm trung bình, độ lệch chuẩn, độ xiên (minh chứng sau khi kiểm tra). Cả hai bài KTNPTV này đều ở dưới dạng trắc nghiệm khách quan bốn lựa chọn với ưu điểm là chấm bài nhanh, chính xác, khách quan, không phụ thuộc vào người đánh giá và thời điểm đánh giá (Nguyễn Văn Long, 2017; Đỗ Quang Việt, 2014). Chính vì vậy, chúng tôi không cần xem xét tới tính giá trị trong việc chấm thi trong nghiên cứu này.

Kết quả nghiên cứu cho thấy cả hai bài KTNPTV đều đảm bảo tính giá trị tương đối tốt, đặc biệt là bài số 2. Tuy nhiên cần chỉnh sửa một số tiểu mục quá dễ hoặc quá khó và giảm cấp độ ngôn ngữ của bài kiểm tra. Đây là nghiên cứu đầu tiên của Khoa Ngôn ngữ và Văn hóa Pháp, ĐHNN-ĐHQGHN đo tính giá trị của một bài kiểm tra/ bài thi do Khoa quản lý.

2. Phương pháp nghiên cứu

2.1. Thu thập dữ liệu

Dữ liệu nghiên cứu gồm hai đề bài KTNPTV số 1 và số 2 học phần 2A + 2B và kết quả thi theo từng tiểu mục của 89 sinh viên cho bài số 1 và 88 sinh viên cho bài số 2. Tất cả các em đều là sinh viên đã học tiếng Anh ở cấp 3 và bắt đầu học tiếng Pháp ở bậc đại học. Số sinh viên bắt đầu học tiếng Pháp từ bậc phổ thông không tham gia nghiên cứu này.

Mỗi bài KTNPTV có cấu trúc giống nhau, gồm 30 câu hỏi trong đó có 20 câu Ngữ pháp và 10 câu Từ vựng. Tất cả các câu hỏi đều dưới hình thức trắc nghiệm khách quan bốn lựa chọn. Bài thứ nhất được kiểm tra vào tuần 5 và bài thứ hai vào tuần 12 học kì hai của năm thứ nhất. Hai bài này được tính điểm thường xuyên cho học phần thực hành tiếng 2A và 2B. Giáo viên được giao bản mô tả kĩ thuật khi soạn bài kiểm tra.

2.2. Xử lý dữ liệu

Nghiên cứu này sử dụng cả hai phương pháp nghiên cứu định lượng và định tính. Trước hết, tính giá trị nội dung của bài KTNPTV sẽ được kiểm chứng bằng cách so sánh độ tương thích giữa nội dung bài kiểm tra với bản mô tả kỹ thuật đề thi. Cụ thể các tiêu chí so sánh là: dạng thức câu hỏi, số lượng câu hỏi, nội dung kiến thức cần kiểm tra, cấp độ ngôn ngữ của các câu hỏi. Hai phần mềm được sử dụng để đo cấp độ ngôn ngữ của các tiểu mục là Readability Formulas (www.readabilityformulas.com) và Compleat Lexical Tutor software version 6.2 (http://www.lextutor.ca/).

Sau đó, chúng tôi đo độ khó của từng tiểu mục bằng cách tính số sinh viên trả lời đúng tiểu mục đó trên tổng số sinh viên tham gia bài kiểm tra (El Allaoui et al., 2016; Morissette, 1996). Một tiểu mục quá dễ (hầu hết sinh viên đều trả lời đúng) hay quá khó (rất ít sinh viên trả lời đúng) đều không đạt yêu cầu và phải được chỉnh sửa biên tập lại. Theo Morissette (1996), những tiểu mục đạt yêu cầu phải có độ khó $> 0,4$ và $< 0,9$. Còn theo El Allaoui et al., (2016), chỉ số độ khó phù hợp là $> 0,2$ và $< 0,8$. Nghiên cứu của El Allaoui et al. (2016) được tiến hành trên một bài thi cuối kì ngành Sinh học, có yêu cầu phân loại thí sinh cao hơn bài KTNPTV của chúng tôi. Chính vì vậy, chúng tôi quyết định sử dụng ngưỡng chỉ số độ khó của Morissette (1996) phù hợp hơn với bài kiểm tra Ngoại ngữ và bài kiểm tra với mục đích điều chỉnh quá trình học tập.

Cuối cùng, chúng tôi sử dụng phương pháp thống kê mô tả trên Excel để xác định một số đặc tính cơ bản của bài kiểm tra bao gồm điểm trung bình (mean), độ xiên (skewness) và độ lệch chuẩn (standard deviation). Đây là một công cụ thống kê có sẵn trong Excel và tương đối dễ sử dụng.

3. Kết quả

3.1. So sánh độ tương thích với bản mô tả kĩ thuật

Bảng 2. So sánh độ tương thích của bài KTNPTV số 1 với bản mô tả kĩ thuật

Tiêu chí so sánh	Bản mô tả kĩ thuật	Bài KTNPTV số 1
Dạng thức câu hỏi	Trắc nghiệm khách quan 4 lựa chọn	Trắc nghiệm khách quan 4 lựa chọn
Số lượng câu hỏi	20 câu Ngữ pháp và 10 câu Từ vựng	20 câu Ngữ pháp và 10 câu Từ vựng
Nội dung kiến thức Ngữ pháp	Imparfait PC 4 tiểu mục So sánh 3 tiểu mục Il y a - Depuis 2 tiểu mục COD/ COI 3 tiểu mục Qui, que, à qui 2 tiểu mục Hợp giống số phân từ quá khứ 2 tiểu mục .....[1] Tổng 20 tiểu mục	Imparfait PC 4 tiểu mục So sánh 3 tiểu mục Il y a - Depuis 2 tiểu mục COD/ COI 3 tiểu mục Qui, que, à qui 2 tiểu mục Hợp giống số phân từ quá khứ 2 tiểu mục ..... Tổng 20 tiểu mục
Nội dung kiến thức Từ vựng	Nông thôn và thành thị 1 tiểu mục Nhà ở, sửa nhà 3 tiểu mục Mô tả tính cách 2 tiểu mục ..... Tổng 10 tiểu mục	Nông thôn và thành thị 1 tiểu mục Nhà ở, sửa nhà 3 tiểu mục Mô tả tính cách 2 tiểu mục ..... Tổng 10 tiểu mục
Cấp độ ngôn ngữ	Cấp độ ngôn ngữ A2	Reading ease: 87,6% Flesch-Kincaid Grade Level: 1.8 K1+K2: 90,25 % K3: 94,04 % Trên K3: 5,96 %

Bảng số 2 cho thấy có độ tương thích rất cao giữa bài KTNPTV số 1 với bản mô tả kĩ thuật ở ba tiêu chí dạng thức câu hỏi, số lượng câu hỏi và nội dung kiến thức cần kiểm tra. Ở tiêu chí cấp độ ngôn ngữ các tiểu mục, bản mô tả kĩ thuật quy định cấp độ ngôn ngữ A2 nhưng lượng từ vựng trên mức độ K1+K2 là 9,75%.

Bảng 3. So sánh độ tương thích của bài KTNPTV số 2 với bản mô tả kĩ thuật

Tiêu chí so sánh	Bản mô tả kĩ thuật	Bài KTNPTV số 2
Dạng thức câu hỏi	Trắc nghiệm khách quan 4 lựa chọn	Trắc nghiệm khách quan 4 lựa chọn
Số lượng câu hỏi	20 câu Ngữ pháp và 10 câu Từ vựng	20 câu Ngữ pháp và 10 câu Từ vựng
Nội dung kiến thức Ngữ pháp	Danh từ hóa 1 tiểu mục Dạng bị động 2 tiểu mục Các thời quá khứ 2 tiểu mục Đại từ y/ en 1 tiểu mục ..... Tổng 20 tiểu mục	Danh từ hóa 1 tiểu mục Dạng bị động 2 tiểu mục Các thời quá khứ 2 tiểu mục Đại từ y/ en 1 tiểu mục ..... Tổng 20 tiểu mục

[1]Vì lí do bảo mật, chúng tôi không đưa toàn bộ nội dung kiến thức cần kiểm tra vào trong bảng thống kê.

Nội dung kiến thức Từ vựng	Kể lại một vụ trộm 1 tiểu mục Phim ảnh 2 tiểu mục Kể lại một kì nghỉ 1 tiểu mục ..... Tổng 10 tiểu mục	Kể lại một vụ trộm 1 tiểu mục Phim ảnh 2 tiểu mục Kể lại một kì nghỉ 1 tiểu mục .... Tổng 10 tiểu mục
Cấp độ ngôn ngữ	Cấp độ ngôn ngữ B1-	Reading ease: 82,9% Flesch-Kincaid Grade Level: 2.5 K1+K2: 89,65 % K3: 92,63 % Trên K3: 7,37 %

Với bài KTNPTV số 2, ba tiêu chí đầu về dạng thức câu hỏi, số lượng câu hỏi và nội dung kiến thức cũng cho thấy bài kiểm tra bám rất sát bản mô tả kĩ thuật. Ở tiêu chí cấp độ ngôn ngữ, các chỉ số Reading ease và Flesch-Kincaid Grade Level cho thấy bài số 2 khó hơn bài số 1, tuy nhiên chỉ số này không cao ở mức B1. Về từ vựng, cấp độ từ vựng trên K3 của bài số 2 là 7,37% cũng khó hơn bài số 1. Theo bản mô tả kĩ thuật đề thi Nghe trình độ B1 của Aptis General Technical Manual, 95% lượng từ vựng của bài Nghe phải nằm trong K3, số từ trên K3 không được vượt quá 5%, (O'Sullivan & Dunlea, 2105).

3.2. Độ khó của từng tiểu mục

Độ khó của từng tiểu mục được đo bằng tổng số sinh viên làm đúng trên tổng số sinh viên tham gia bài kiểm tra (Morissette, 1996).

Bảng 4. Chỉ số độ khó của từng tiểu mục trong hai bài kiểm tra

Tiểu mục	Bài KTNPTV số 1	Bài KTNPTV số 2
1	**0,39**	0,67
2	0,71	0,58
3	**0,36**	0,85
4	0,78	0,82
5	0,59	0,74
6	***0,97***	0,46
7	0,58	0,68
8	0,51	0,59
9	**0,39**	0,68
10	0,80	0,53
11	0,45	0,60
12	0,54	0,68
13	0,70	0,62
14	0,64	0,67
15	0,67	0,65
16	0,84	0,40
17	0,55	0,88
18	0,74	0,64
19	0,68	**0,31**
20	0,84	0,53
21	**0,19**	0,69
22	0,84	0,70
23	0,51	0,60
24	0,48	0,62
25	**0,29**	0,47
26	0,74	0,48
27	0,47	0,83
28	**0,10**	0,64
29	0,55	***0,91***
30	***0,97***	0,87

Theo Morissette (1996), với một bài kiểm tra có ngưỡng điểm đạt là 6/10 thì độ khó của các tiểu mục nên ở giữa 0,4 và 0,9. Như vậy, những tiểu mục có độ khó không đạt ngưỡng này ở bài kiểm tra số 1 là: 1, 3, 9, 21, 25, 28 (< 0,4 quá khó) và tiểu mục 6, 30 (> 0,9 quá dễ). Kết quả này cũng trùng khớp với phản ánh của một số giáo viên sau khi chấm thi có đề nghị điều chỉnh lại một số tiểu mục, đặc biệt phần Từ vựng có nhiều câu khó. Ở bài kiểm tra số 2, chỉ có câu số 19 là có độ khó < 0,4 và câu 29 có độ khó > 0,9.

3.3. Thống kê mô tả

Cuối cùng, chúng tôi sử dụng phương pháp thống kê mô tả trên Excel để xác

định một số đặc tính cơ bản của bài kiểm tra bao gồm điểm trung bình (mean), độ xiên (skewness) và độ lệch chuẩn (standard deviation).

Bảng 5. Kết quả thống kê mô tả cho bài KTNPTV số 1

Kết quả bài KTNPTV số 1	
Mean	6,05
Standard Error	0,14
Median	6,00
Mode	5,70
Standard Deviation	1,36
Sample Variance	1,86
Kurtosis	3,51
Skewness	-0,95
Range	9,00
Minimum	0,00
Maximum	9,00
Sum	538,80
Count	89,00

Thống kê mô tả cho thấy điểm trung bình chung bài KTNPTV số 1 là 6.05/10, một điểm số ở mức trung bình cho toàn khối. Theo Morissette (1996), điểm trung bình của bài kiểm tra quá trình học tập nên ở mức 6,5-7,5/10, như vậy là điểm trung bài của bài KTNPTV số 1 là hơi thấp. Độ xiên là -0,95 < 0, có độ xiên âm lệch trái nghĩa là các giá trị cực nhỏ (điểm thấp) hơn giá trị trung bình sẽ nhiều hơn so với giá trị cực lớn (điểm cao) hơn giá trị trung bình. Độ lệch chuẩn là 1.36 tương đối thấp, trong tình huống kiểm tra quá trình học tập thì độ lệch chuẩn thấp này là phù hợp, chứng tỏ mức độ nắm bài của sinh viên là tương đối như nhau, không có độ khác biệt quá lớn giữa các sinh viên. Độ nhọn là 3,51 cho thấy phân phối mẫu tương đối tập trung, và cho kết quả giống độ lệch chuẩn là không có độ khác biệt quá lớn giữa các sinh viên.

Bảng 6. Kết quả thống kê mô tả cho bài KTNPTV số 2

Kết quả bài KTNPTV số 2	
Mean	6,80
Standard Error	0,20
Median	7,00
Mode	8,00
Standard Deviation	1,90
Sample Variance	3,60
Kurtosis	2,47
Skewness	-1,23
Range	9,60
Minimum	0,00
Maximum	9,60
Sum	598,03
Count	88,00

Ở bài KTNPTV số 2 này, điểm trung bình chung toàn khối là 6,80, cao hơn bài số 1 và nằm trong ngưỡng điểm hợp lý theo Morissette (1996). Độ xiên là -1,23, vẫn là độ xiên âm lệch trái, lí do là có hai em bị điểm 0 và một vài em điểm dưới trung bình. Độ lệch chuẩn là 1,90, cao hơn độ lệch chuẩn bài số 1, chứng tỏ sự chênh lệch giữa nhóm sinh viên đạt điểm cao và nhóm sinh viên đạt điểm thấp là lớn hơn bài trước. Độ nhọn là 2,47 cho thấy các mẫu phân tán hơn, điểm số các sinh viên trong khối khác nhau nhiều hơn. Có thể là bài kiểm tra số 2 này được tiến hành vào cuối học kì 2 của năm thứ nhất, mức độ phân hóa sinh viên đã lớn hơn đầu học kì.

4. Diễn giải

4.1. So sánh độ tương thích với bản mô tả kĩ thuật

Nghiên cứu cho thấy là nhìn chung, hai bài KTNPTV có độ tương thích cao với bản mô tả kĩ thuật và đảm bảo tính giá trị nội dung cho bài kiểm tra. Một bài kiểm tra có tính giá trị nội dung khi nội dung đưa vào bài thi liên quan chặt chẽ với nội dung của môn học và mang tính đại diện cho toàn bộ nội dung môn

học (Messick, 1990). Ví dụ, một bài kiểm tra ngữ pháp phải bao gồm các tiểu mục liên quan chặt chẽ tới kiến thức ngữ pháp và mang tính đại diện cho các cấu trúc ngữ pháp đã học. Nghiên cứu của Nguyễn Thị Phương Thảo (2018) cũng cho kết quả tương tự: các tác giả biên soạn đề thi Đọc VSTEP đã bám rất sát bản mô tả kĩ thuật đề thi, đảm bảo tính giá trị nội dung cho đề thi.

Các giáo viên thực hành tiếng tổ 1, Khoa Ngôn ngữ và Văn hóa Pháp, ĐHNN-ĐHQGHN, đa phần là giáo viên trẻ, chưa có nhiều kinh nghiệm về kiểm tra đánh giá nhưng nhờ có bản mô tả kĩ thuật rõ ràng và sự chỉ đạo chuyên môn sát sao, họ đã áp dụng tương đối tốt bản mô tả kĩ thuật khi soạn bài KTNPTV cho năm học 2016-2017.

Tuy nhiên, về cấp độ ngôn ngữ của các tiểu mục thì cần có sự điều chỉnh dễ hơn và phù hợp hơn về mặt từ vựng, đặc biệt là ở bài kiểm tra số 1. Cụ thể cần giảm số lượng từ vựng trên K3 xuống dưới 5% cho bài kiểm tra số 2 và giảm số lượng từ vựng trên K2 xuống dưới 5% cho bài số kiểm tra số 1. Về độ dễ đọc của các tiểu mục, các chỉ số Reading ease và Flesch-Kincaid Grade Level đều ở cấp độ dễ vì hai bài KTNPTV đều là trắc nghiệm 4 lựa chọn với đa số là câu đơn và nhiều phương án chỉ có một từ đơn lẻ. Kết quả này cũng đi cùng hướng với nghiên cứu của Nguyễn Thị Phương Thảo (2018) cho thấy cấp độ ngôn ngữ sử dụng trong đề Đọc VSTEP khó hơn yêu cầu trong bản mô tả kĩ thuật, do vậy phần nào ảnh hưởng tới kết quả thi của thí sinh.

Như vậy, giáo viên cần ý thức hơn nữa về độ khó của cấp độ ngôn ngữ sử dụng trong bài kiểm tra. Thường thì khi soạn đề giáo viên quan tâm nhiều đến dạng thức câu hỏi và nội dung kiến thức cần kiểm tra, còn ngữ cảnh đặt nội dung đó chưa được nhiều giáo viên thực sự quan tâm. Chính vì vậy, trong quá trình duyệt đề, tác giả bài báo này đôi khi gặp những tiểu mục đánh giá kiến thức ngữ pháp cấp độ A2 nhưng đặt trong câu dẫn có cấp độ ngôn ngữ B1+. Theo chúng tôi, các công cụ giúp giáo viên đo cấp độ ngôn ngữ đầu vào như Readability Formulas hay Compleat Lexical Tutor nên được giới thiệu rộng rãi tới các giáo viên thực hành tiếng.

4.2. Độ khó của từng tiểu mục

Về độ khó của tiểu mục, kết quả phân tích cho thấy bài KTNPTV số 2 có độ khó phù hợp hơn bài số 1. Ở bài số 2 chỉ có 2/30 tiểu mục quá khó hoặc quá dễ, trong khi ở bài số 1, số lượng tiểu mục có độ khó chưa phù hợp là 8/30 câu. Rút kinh nghiệm từ bài KTNPTV số 1, nhóm soạn đề đã có một số điều chỉnh hiệu quả ở bài số 2. Trong nghiên cứu của El Allaoui et al. (2106), trên 26 tiểu mục có 2 tiểu mục quá khó và 2 tiểu mục quá dễ. Còn Nguyễn Thị Phương Thảo (2018) đã tìm ra 5 tiểu mục khó hơn mức độ yêu cầu và 2 tiểu mục dễ hơn mức độ yêu cầu trong tổng số 40 tiểu mục.

Khi xem xét lại các tiểu mục quá khó, chúng tôi nhận thấy có các nguyên nhân sau đây: câu hỏi rơi vào trường hợp đặc biệt, câu hỏi kiểm tra hai mảng kiến thức trong cùng một câu, câu hỏi kiểm tra mảng kiến thức dễ nhầm giữa tiếng Pháp và tiếng Anh, câu hỏi có câu dẫn ở cấp độ ngôn ngữ khó, câu hỏi kiểm tra kiến thức từ vựng cao hơn cấp độ yêu cầu (ví dụ từ *caution* là một từ cao hơn cấp độ A2), câu hỏi kiểm tra kiến thức từ vựng sinh viên chưa học kĩ trên lớp.

Ở các tiểu mục có số sinh viên làm đúng trên 90%, các câu dẫn và phương án trả lời rất rõ ràng, cấp độ ngôn ngữ A2, từ vựng rất ấn tượng và dễ nhớ (ví dụ *coup de foudre - tiếng sét ái tình*). Đa số các tiểu mục cần xem xét lại, rơi vào phần Từ vựng (5/10 câu cho cả hai bài), cho dù trọng số của phần Từ vựng chỉ là 33% tổng điểm bài kiểm tra. Có vẻ như giáo viên khi làm đề mới chỉ quan tâm xem từ đó

đã xuất hiện trong bài học trên lớp chưa, chứ chưa thực sự tính đến việc từ đó có phù hợp với cấp độ ngôn ngữ chuẩn đầu ra hay chưa. Hơn nữa, hầu như các từ mới sinh viên mới gặp một đôi lần trong sách giáo khoa hay sách bài tập, tần suất chưa đủ để những kiến thức từ vựng được khắc sâu trong trí nhớ của các em.

4.3. Thống kê mô tả

Kết quả thống kê mô tả cũng cho thấy bài KTNPTV số 2 có điểm trung bình chung toàn khối (6,80/10) cao hơn bài số 1 (6,05/10) và phù hợp hơn với yêu cầu của một bài kiểm tra quá trình học tập (Morissette, 1996). Như vậy những điều chỉnh về độ khó của từ vựng trong phần câu dẫn, độ khó của các tiểu mục trong cả bài thi đã giúp cho sinh viên có kết quả thi phù hợp hơn với nỗ lực học tập của các em. Độ lệch chuẩn của cả hai bài kiểm tra là 1,36 và 1,90 chứng tỏ sự phân bố điểm không quá chênh lệch giữa các sinh viên nhóm điểm cao và sinh viên nhóm điểm thấp. Điểm trung bình chung trong nghiên cứu của El Allaoui et al. (2016) là 10,10/20; của Nguyễn Thị Phương Thảo (2018) là 15,08/40; tuy nhiên, cả hai nghiên cứu này đều đo các bài thi có mục đích cấp chứng chỉ nên sẽ có mức điểm trung bình thấp hơn bài kiểm tra quá trình học tập thường xuyên.

5. Kết luận

Nghiên cứu này đã chỉ ra rằng hai bài KTNPTV học phần 2A + 2B số 1 và số 2 có tính giá trị nội tương đối phù hợp với các công cụ phân tích, trong đó bài số 2 có tính giá trị nội dung cao hơn bài số 1. Cả hai bài KTNPTV đều tuân thủ bản mô tả kĩ thuật về dạng thức câu hỏi, số lượng câu hỏi, nội dung kiến thức cần kiểm tra. Đa số các câu hỏi có độ khó phù hợp từ 0,4-0,9, điểm trung bình của bài KTNPTV số 2 cũng như độ lệch chuẩn của cả hai bài đều ở trong mức phù hợp (Morissette, 1996). Tuy nhiên, kết quả phân tích cũng chỉ ra rằng ngôn ngữ sử dụng ở cả hai bài kiểm tra còn ở cấp độ cao hơn với yêu cầu của bản mô tả kĩ thuật, tỉ lệ các tiểu mục có chỉ số độ khó chưa phù hợp còn tương đối cao ở bài KTNPTV số 1 (27%), điểm trung bình chung của bài số 1 còn hơi thấp so với yêu cầu của một bài kiểm tra quá trình học tập.

Như vậy, hai bài kiểm tra cần được rà soát lại dựa trên những kết quả trên. Nếu không tìm thấy vấn đề ở khâu biên soạn đề, thì cần xem xét lại quá trình dạy học mảng kiến thức ngữ pháp từ vựng đó, tại lớp có kết quả chưa cao (El Allaoui et al., 2016). Trong tương lai, bản mô tả kĩ thuật đề thi cần mô tả chi tiết hơn yêu cầu về cấp độ ngôn ngữ sử dụng trong đề thi (cấp độ từ vựng – ngữ pháp, số lượng từ trong câu dẫn, số lượng từ trong phương án trả lời...).

Đây là một trong những nghiên cứu đầu tiên đặt mục tiêu đo tính giá trị nội dung của một bài kiểm tra/ thi tại Khoa Ngôn ngữ và Văn hóa Pháp. Nghiên cứu tính giá trị nội dung của bài kiểm tra thi cấp bộ môn này là một bước tiệm cận dù ở quy mô rất nhỏ tới quy trình kiểm tra đánh giá quốc tế và cũng thể hiện mong muốn của đội ngũ giảng viên ĐHNN-ĐHQGHN là nâng cao chất lượng khảo thí tiến tới hội nhập quốc tế (Nguyễn Thị Ngọc Quỳnh, 2018). Nghiên cứu cũng tập trung sử dụng các công cụ đo tương đối đơn giản được cung cấp miễn phí mà giáo viên không cần có trình độ công nghệ cao có thể sử dụng được. Tuy nhiên, một số chuyên gia khảo thí Việt Nam cũng khuyến cáo cần "*tham khảo và áp dụng hệ thống chuẩn nước ngoài một cách chọn lọc, linh hoạt*" (Đỗ Quang Việt, 2014, tr. 52).

Nghiên cứu này cũng cho thấy giáo viên cần được tập huấn thêm về năng lực khảo thí (El Allaoui et al., 2016; Combs et al., 2018). Người biên soạn đề thi/ kiểm tra cần ý thức rõ ràng hơn về việc chọn cấp độ ngôn ngữ phù hợp với trình độ yêu cầu trong bản mô tả kĩ

thuật, kể cả với đề Ngữ pháp – Từ vựng. Cần giới thiệu tới các giáo viên các phần mềm hỗ trợ đo cấp độ ngôn ngữ của bài kiểm tra. Tuy nhiên, các công cụ này chỉ là một kênh tham khảo, trình độ chuyên môn của người soạn đề vẫn đóng vai trò quan trọng nhất khi xác định cấp độ ngôn ngữ phù hợp của một bài kiểm tra (Nguyễn Thị Phương Thảo, 2018, tr. 135). Trong báo cáo tổng kết cuối năm học 2017, nhiều giáo viên trẻ của Tổ Thực hành tiếng năm thứ nhất, Khoa Ngôn ngữ và Văn hóa Pháp, ĐHNN-ĐHQGHN, đều bày tỏ nhu cầu được tập huấn thêm về kĩ năng kiểm tra đánh giá. Tập huấn về Kiểm tra đánh giá do Đề án Quốc gia về Ngoại ngữ 2020 tổ chức cho đối tượng giáo viên các ngoại ngữ khác tiếng Anh vào tháng 8 năm 2018 tại Hà Nội cho thấy rất ít giảng viên đại học sử dụng bản mô tả kĩ thuật đề thi một cách bài bản khi soạn đề thi/ kiểm tra.

Lời cảm ơn

Xin trân trọng cảm ơn Trường Đại học Ngoại ngữ - Đại học Quốc gia Hà Nội, Ban chủ nhiệm Khoa Ngôn ngữ và Văn hóa Pháp, toàn thể các thầy cô dạy Thực hành tiếng năm thứ nhất QH2016 và các em sinh viên năm thứ nhất QH2016 đã giúp chúng tôi hoàn thành nghiên cứu này.

Tài liệu tham khảo

Tiếng Việt

Nguyễn Thúy Lan (2017). Một số tác động của bài thi đánh giá năng lực tiếng Anh theo chuẩn đầu ra đối với việc dạy tiếng Anh tại Trường Đại học Ngoại ngữ - Đại học Quốc gia Hà Nội. *Nghiên cứu Nước ngoài, 33*(4), 122-136.

Nguyễn Văn Long (2017). Thiết lập đề thi trắc nghiệm trực tuyến trên cơ sở các tiêu chí ngôn ngữ của khung tham chiếu năng lực ngoại ngữ chung châu Âu (CEFR). *Nghiên cứu Nước ngoài, 34*(3), 153-163.

Dương Thu Mai, Nguyễn Thị Chi, Phạm Thị Thu Hà (2017). Xây dựng năng lực đánh giá cho giáo sinh ngành sư phạm tiếng Anh tại Đại học Quốc gia Hà Nội dựa trên nguyên tắc về tính giá trị. *Nghiên cứu Nước ngoài, 33*(1), 60-72.

Đỗ Thị Bích Thủy (2018). Ý kiến phản hồi của người dạy và người học về công tác kiểm tra đánh giá các học phần thực hành tiếng 1A + 1B, 2A + 2B tại Khoa Ngôn ngữ và Văn hóa Pháp, Trường Đại học Ngoại ngữ - Đại học Quốc gia Hà Nội. *Nghiên cứu Nước ngoài, 34*(3), 125-137.

Lê Thị Huyền Trang & Trần Thị Tuyết (2015). Đổi mới kiểm tra đánh giá: Từ thực tế của các lớp bồi dưỡng tiếng Anh cho giáo viên tiểu học. *Tạp chí Khoa học ĐHQGHN: Nghiên cứu Nước ngoài, 31*(2), 51-60.

Đỗ Quang Việt (2014). Khảo sát thực trạng việc sử dụng dạng thức Trắc nghiệm khách quan và Trắc nghiệm tự luận trong kiểm tra tiếng Pháp ở trung học phổ thông khu vực phía Bắc Việt Nam. *Tạp chí Khoa học ĐHQGHN: Nghiên cứu Nước ngoài, 30*(1), 42-54.

Tiếng Anh

Bachman, L. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Coombs A., DeLuca C., LaPointe-McEwan D., Chalas A. (2018). Changing approaches to classroom assessment: An empirical study across teacher career stages. *Teaching and Teacher Education, 71*, 134-144. https://doi.org/10.1016/j.tate.2017.12.010.

Bachman, L. & Palmer, A. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

Bailey, K. M. (1996). Working for Washback: A Review of the Washback Concept in Language Testing. *Language Testing*, *13*(3), 257-279.

Hoang Hong Trang, Nguyen Thi Chi, Duong Thu Mai (2016). Specifications Framework for Tests in an Outcome-based Language Program. *VNU Journal of Science: Foreign Studies, 32*(4), 64-73.

Herppich, S., Praetorius, A. K., Förster, N., Glogger-Fre, I., Karst, K., Leutner, D., Behrmann, L., Böhmer, M., Ufer, S., Klug, J., Hetmanek, A., Ohle, A., Böhmer, I., Karing, C., Kaiser, J., Südkamp, A. (2017). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education xxx,* 1-13. https://doi.org/10.1016/j.tate.2017.12.001.

Hooker, T. (2017). Transforming teachers' formative assessment practices through ePortfolios. *Teaching and Teacher Education, 67*, 440-453. http://dx.doi.org/10.1016/j.tate.2017.07.004.

Hughes, C. (2003). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Lissitz, R. W. (2009) (ed.) *The concept of validity: revisions, new directions, and applications*. Charlotte, NC: Information Age Publishing Inc.

Messick, S. (1989). Validity. In R. L. Linn (Ed.) *Educational Measurement* 3rd ed. (pp. 13-103). New

York: American Council on Education/Macmillan.

Messick, S. (1990). *Validity of Test Interpretation and Use*. Princeton, New Jersey: Educational Testing Service.

Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. *American Psychologist, 50*(9), 741-749.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45*(1-3), 35–44.

Remesal, A. (2011). Primary and secondary teachers' conceptions of assessment: A qualitative study. *Teaching and Teacher Education, 27*, 472-482. doi: 10.1016/j.tate.2010.09.017.

Nguyen Thi Ngoc Quynh (2018). A study on the validity of VSTEP writing tests for the sake of regional and international integration. *VNU Journal of Foreign Studies, 34*(4), 115-128.

Nguyen Thi Phuong Thao (2018). An investigation into the content validity of a Vietnamese standardized test of English proficiency (VSTEP.3-5) reading test. *VNU Journal of Foreign Studies 34*(4), 129-143.

Nguyen Thi Quynh Yen (2016). Rater Consistency in Rating L2 Learners' Writing Task. *VNU Journal of Science: Foreign Studies, 32*(2), 75-84.

O' Sullivan, B. & Dunlea, J. (2015). *Aptis General Technical Manual: Version 1.0*. London: British Council.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Tiếng Pháp

El Allaoui, A., Rhazi Filali, F., El Hadri, E. M., Fetteh, K., Bouhadi, M. (2016). Étude évaluative d'examen normalisé de sciences de la vie et de la terre au cycle secondaire collegial. *European Scientific Journal, 12*(1), 283-299. doi: 10.19044/esj.2016.v12n1p283.

Issaieva, E. & Crahay, M. (2010). Conceptions de l'évaluation scolaire des élèves et des enseignants: Validation d'échelles et étude de leurs relations. *Mesure et évaluation en éducation*, *33*(1), 31–61. doi:10.7202/1024925ar.

Morissette, D. C. (1996). *Guide pratique de l'évaluation sommative: gestion des épreuves et des examens.* Montréal: Éd. du Renouveau Pédagogique Inc.

VALIDITY ANALYSIS OF 2A + 2B GRAMMAR - VOCABULARY TESTS AT THE FACULTY OF FRENCH LANGUAGE AND CULTURE, UNIVERSITY OF LANGUAGES AND INTERNATIONAL STUDIES - VIETNAM NATIONAL UNIVERSITY, HANOI

Do Thi Bich Thuy

Faculty of French Language and Culture, VNU University of Languages and International Studies, Pham Van Dong, Cau Giay, Hanoi, Vietnam

Abstract: This paper investigated the validity of two grammar - vocabulary tests (2A + 2B proficiency units in the academic year 2016-2017) at the Faculty of French Language and Culture, University of Languages and International Studies - Vietnam National University, Hanoi. The study aimed to evaluate the relevance of the content in these tests in comparison with the test specifications, to measure the difficulty index of each item and some indicators of the tests. The study results showed a relatively high validity of the investigated tests, with a better validity for the second test. However, the language level and items with irrelevant difficulty index should be reviewed to better fit the test specifications.

Keywords: language testing, grammar - vocabulary tests, validity, test specification, difficulty index