PAC3
at
JALT
2001
Conference
Proceedings

International
Conference
Centre
Kitakyushu
JAPAN
November
22–25, 2001

Assessing L2 Achievement In e-Based Programs

Gerry Lassche
Ajou University

The paper discusses the relationship between language epistemology and the construct validity of language achievement procedures. Influenced by the work of Bachman's (1990) and Bachman and Palmer's (1996) work on the evaluation of test usefulness, the exploration of the procedures used for gauging student achievement will be presented in case study format on three aspects of assessment with regard to construct validity. These include a characterization of the entrance test, within-course exercises, and tutorial tasks. Examples provided about these aspects suggest that the program evaluates student performance in terms of a discrete knowledge domain. Since this view conflicts with the program's stated language ability construct, the procedure for the assessment of student achievement is invalid.

Network technologies, information systems, the Internet – all these technological affordances are impacting educational domains, including foreign languages, to a greater and greater degree. To that end, there have been many efforts to develop interactive multimedia programs (IMM), and advances in computer technology (multi-media and video-conferencing) are making those efforts more tangible day by day (Brett, 1998). This integration has not been

occurring without some debate concerning the degree of substantiating research (White House Papers, 1997; Daugherty and Funke, 1998), which asks the question "Is IT more conducive to learning than traditional classroom-based methods?" Hara and Kling (2000) suggest that the vast majority of research that is being conducted and published tends to highlight only positive aspects of the use of network technology.

In Lassche (2000), I described some of the outstanding pedagogical issues with reference to language learning. In that paper, I proposed several language teaching pedagogical principles, and from that platform criticized IT efforts in terms of cognitive and social drawbacks in the web-based learning context. In this paper, I present in case study format an evaluation of one popular, online web-based language learning program currently in use in Europe. The case study will focus on the assessment procedures used by this program, following the recommendations proposed by Bachman (1990) and expanded by Bachman and Palmer (1996) for determining the usefulness of testing procedures. Bachman's approach to construct validity has become seminal in the language-testing field, having received widespread endorsement (Douglas, 2000; Chappelle, 1997; Hegelheimer and Chappelle, 2000; Gunn, 1995; Clarksen and Jensen, 1995, among others.) Bachman and Palmer's entire test usefulness framework is too comprehensive to deal with in this paper, and so the evaluation will focus on one aspect of the framework: defining the language construct for construct validity.

Construct validity, in terms of language epistemology, will be discussed. The influence of language epistemology on the construction of language tests follows. As a result of this discussion, validity will be presented as the degree of correspondence between language epistemology and test use. Finally, an application of this view of test design issues will be provided in the case study format described above.

What is a construct?

When researchers are trying to understand some phenomenon, they try to compile a set of characteristics which describe their perceptions about it. This representation is their deliberate attempt to organize those perceptions in a way that helps to quantify and qualify that phenomenon (after Bachman, 1990, 255ff). These constructs, then, typify one's epistemological views of that phenomenon (after Cohen et al, 2000). In the TESOL field, language is viewed as either serving:

1. a communicative function, in a "social semiotic" sense (Martin, 1993); or
2. a knowledge function, as a set of discrete language bits.

In the first view, language is viewed as a system of meaning-making, negotiated between individuals (Long and Robinson, 1996), a system which is used by real people in real situations for real purposes (Widdowson, 1976), as influenced by cultural, psycho-social, physical and temporal characteristics of the interactants, the setting, and the particular communicative functions at issue (Martin, 2000).

Another epistemological position views language as a set of discrete informational elements that, if accumulated by the student in particular sequences, would eventually give rise to an ability to understand a foreign language. This is best exemplified by the *Grammar Translation* approach (Richards and Rogers, 2000, 6). Language in this view exists as a body of knowledge independent of communicative purpose. The elements are discrete in that they are taken to have absolute meanings, which are not influenced through interaction with other elements.

What is validity?

Validity is seen by Bachman and Palmer (1996) as involving three broad aspects: construct definition, in terms of the way the test/program developer views the nature of language; construct realization, in the way that characteristics of the test/program exercises reflect the construct definition; and construct interpretation, in the way that the language construct definition is perceived tangibly by the stakeholders, and the way the resulting test-taker/student performance on the exercises is used for scoring, providing feedback, and making other administrative decisions.

What is the relationship between test purpose and language epistemology?

In practice, test developers and TESOL practitioners view language in different ways at different times, reflecting a flexible perspective which suits a context-specific purpose. Brindley (1989, 31ff) describes several types of achievement tests, each type corresponding to a different purpose each test aspires to (see Table 1).

Table 1: *Levels of achievement testing*

Type	Measures	Purpose	Nature
Level 1 = proficiency	Overall language ability	Placement, screening, etc.	Communicative
Level 2 = achievement	Syllabus objectives	Skill / content areas	Comm. / Discrete
Level 3 = diagnostic	Pre-communicative sub-skills	In situ, process feedback	Discrete

Level 1 tests generally measure overall language proficiency: how much general L2 facility does the test-taker have? Bachman (1990, 71) sees proficiency tests as

directly measuring the degree of language facility in real-life contexts or the "real-life domain", that is, language in real use. A Level 2 test is concerned with the functional gains in student abilities; that is, what new skills can the student use as a result of program intervention? Programs may try to train students in particular skills or components of language ability. Achievement tests assess the students' successful use of these skills. Level 3 testing is an informal assessment of the "enabling [sub]skills and knowledge which are part of a particular course of study," and focusing on these sub-skills renders the assessments generally unrelated to communicative performance (Brindley, 1989, 17).

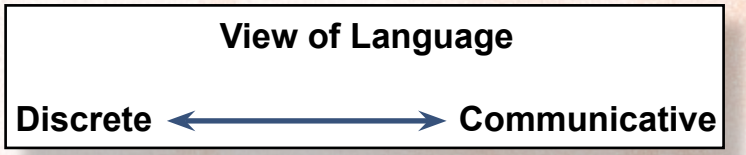

Figure 1: *A continuum of language nature*

Figure 1 above shows that discrete versus communicative epistemologies are more accurately seen as a continuum rather than distinct entities. Level 1 and level 2 achievement tests are more toward the communicative end, and level 3 is at the discrete end. In testing, then, construct validity would be demonstrated when one's view of language matched the purpose and use made of specific test contents and results. For example, if the one's epistemology saw language as communicative, then a valid test that reflected this position would be either Level 1 or 2. If a level 3 test were used to judge proficiency, however, such test use would be invalid.

Case study evaluation

To solicit the participation of the case study program, the researcher provided detailed feedback on the usefulness of test procedures in exchange for free access to the site for research purposes. The selection of programs was restricted to a program which offered distance language education online, required tuition, and involved students who would never have F2F meetings with either fellow students or their instructors. Once consent had been obtained from the appropriate authorities, online examination of the programs ensued, and interviews were conducted via email and telephone contacts. Three aspects of the assessment procedures used by online program will be outlined, as shown in Table 2 below. Then, these aspects will be evaluated in terms of construct validity.

Table 2: *Assessment categories*

Categories of assessment		Purpose
1	Entrance test	Placement Level 1: Proficiency-based
2	Within-course exercises, activities & mid-level test	Level 3: Diagnostic
3	Tutorials	
	Email	Level 2: Progress achievement
	Writing	Level 2: Progress achievement
	Telephone interview	NA
4	Exit test	NA

As indicated in the table 2, the program has four types of assessment. Access was provided to the first five units of the level, and since the exit test is part of the last five units, I was not able to inspect it. In addition, criteria or archived scripts of the telephone interviews were also not obtained.

Language ability construct

The informant at the site stated that his working definition of mid-intermediate proficiency was based on the *Common European Framework of Reference for Languages* (CUP 2001) descriptor for the B2 band (independent user):

> Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.

Three aspects of the above construct definition are important for the present analysis:

1. the emphasis on whole text;
2. the range of topics from concrete to abstract; and
3. the emphasis on fluency and spontaneity.

First component: The Entrance test

The entrance test has two parts. The first section is a 70-item word-fill exercise (see table 3), and the second section asks the users to answer a series of questions in open-ended fashion While the ability construct identified by the course designer suggests an emphasis on whole text, none of the test components described in table (2) tap the skills necessary for constructing whole texts. The 70 questions on the entry test purport to measure proficiencies up to and including advanced levels, yet the test items themselves are discrete sentences. For each item, the test taker needs to insert

the word into the correct place (click on the word and drag it into place):

Table 3: *Example items from section 1*

21.	have	If I'd had the money I would gone to Italy.
31.	was	The plane crashed and everyone on it killed.
41.	by	The problem was caused the computer breaking down.

There is no feedback on these exercises, and the criteria required for a given level of performance not stated or explained. Whole text features, such as mood, transitivity, theme, conjunctive relations, reference, lexical relations, cohesion, and schematic structure (see Eggins, 1994), are not dealt with. The four written response questions also require a two or three sentence response for each. Examples are shown below in Table 4:

Table 4: Example items from section 2

1.	*How do you spend your free time?*
4.	What did you do last weekend?
7.	What are your plans for the next few months?
10.	How important is the Internet to you?

The entry test focuses on topics of general, daily events. No specific genres are readily apparent in the examples. The written component also focuses on themes of a concrete nature (schedule, family, education, job), but two imply a more abstract discussion ("ideal journey" and "importance of the internet").

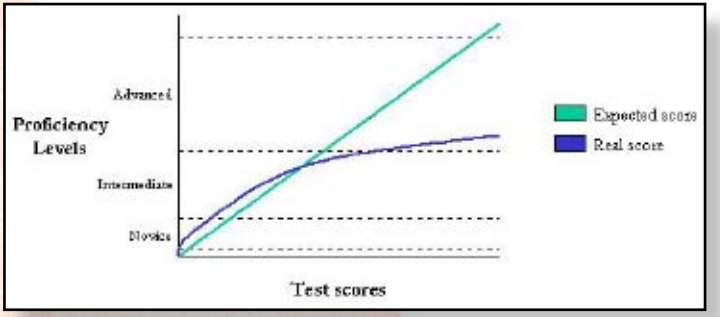

Figure 2: *Relationship between proficiency level and test performance*

The real question is whether the expected response actually discriminates among proficiency levels, corresponding with construct (after Rovinelli and Hambleton, 1976). My informant maintains that the test is "doing its job", since the administration re-assigns the student level only once in roughly every 20 responses on the basis of the written task (personal communication). My suspicion is that most students who take this course are intermediate, with the novice level test-takers guessing at the correct answers.

My reason for this guess is based in part on the findings of Jones (1998), who found that most

consistent and successful users of online programs tend to have intermediate language proficiency. In these cases, their writing, in terms of sentence structure, would give them away. In such a case, a true novice learner probably scores higher than true ability on the first part of the test. Advanced learners score lower than ability, since they may be operating on a "get it wrong for the right reason." These two groups are lower in frequency, hence the perception that the test is "working", when in fact it does not discriminate higher and lower levels accurately, as suggested in Fig. 2.

Second component: within-course exercises

Table 5: *Unit 1 - Task Text*

Unit 1. Happiness
The flowers of spring Birds and bees and all those things. Hot coffee and fresh bread, These words in my head.

The program's language ability construct supports the idea of whole text discourse. Within-course exercises show a greater attention to text level construction, but inconsistently. The text shown in Table 5, culled from Unit 1, is representative of this. It presents a song text (listening mode) in its entirety—a full-length text. Inspection of some of the textual features show characteristics of song genre which diverge from the more typical academic genres in written or spoken discourse, such as lack of a main predicate and non-standard use of punctuation for sentence fragments. Instead of drawing attention to these features, the program has the users do a gap-fill exercise, where they are to click and drag blanked-out words into the correct location in the text.

The written exercises have the students write phrases of favorite things ("my father's cooking", "the taste of beer on a summer's day"), and then construct single sentences using a modal auxiliary (ie can). Students are also encouraged to use a notice board (similar to a discussion post on WebCT) to post their ideas and comment on other students' work. Inspection of the notice board showed only 5 postings in the last 4 months for the mid-intermediate level, and none of these items bore relation to course work. Students that did post are not provided feedback by tutors on this effort. Feedback obtained at random from one of the students gives credence to this assertion: this particular student said he did not finish all the exercises because there was "no one to chat with".

This tendency for students to not fully utilize IMM available in websites is not uncommon to web-based learning. Smith and Salam (2000), in their review of

web-based sites, and Jung (2000), in her review of web-based learning in university environments, found that little or no interaction took place between students either.

Third component: Tutorials

In these exercises, students are asked to compose a text on a topic specified by the program, trying to incorporate or recycle the grammar and vocabulary they had learned in the unit. Teachers use a scoring protocol for these kinds of tutorials which typically follow five dimensions:

- Task achievement - does the content match the specifications? (4 marks)
- Range, complexity and accuracy of grammar (4 marks)
- Vocabulary range and accuracy (4 marks)
- Style and layout (3 marks)
- Overall effect on the reader - is it coherent and communicative? (5 marks)

Table 6: Tutorial task

Task
Describe an experience you had while traveling. Write at least 6 sentences, and try and use some of these expressions: I have to be able to; I don't have to have; I have to have; I should be able to; I don't have to be able to; I don't need to be able to; I need to have; I need to be able to

Of all the exercises in the program, these written tutorials correspond most strongly to the construct definition. In the example tutorial task, found in Table 6, the student is writing a composition for an audience. One student's response is found in appendix 1. Although the tutor mentions in appendix 1 that this student has made "some tense and agreement mistakes that [she] shouldn't be making at her level!" this is more likely an indication of her reliance on context rules rather than code rules (Widdowson, 1979, 194), demonstrating that the task is tapping her communicative competencies (see Lassche, unpublished manuscript). The scoring protocol also corresponds with the definition: i.e. fluency, creativity, etc. The extent that the student has gone beyond the specified length of the task (it calls for 8 sentences) is perhaps an indication of the student's approval of the task and her eagerness to comply with it.

The tutor's response on her work is perhaps the clearest indication of the program's working construct definition for assessing language performance. The diagnostic feedback provided an analysis of semantics and syntactical errors without reference to context, bearing little relevance to the content of the chapter, or to the overall communicative intent of the student-writer.

Conclusion

The participating website, echoing the findings of Gatton (1999), seems more concerned with the financial bottom-line, developing products with high face validity but little construct validity. Testing practices are concerned with assessing discrete language knowledge domains, without regard to issues of stakeholder accountability, but possessing high reliability in many aspects due to the frequency of items with dichotomized test responses. This, in turn, has severely curtailed the ability of assessment outcomes to demonstrate that acquisition in any communicative sense has taken place.

References

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: OUP.

Bachman, L. and Palmer, A. (1996). *Language Testing and Practice*. Hong Kong: OUP.

Brett, Paul. (1998) An Intuitive, Theoretical and Empirical Perspective on the Effectiveness Question for Multimedia. In Cameron, K. (ed.) *Multimedia CALL: Theory and Practice*, 81-93. Exeter: ElmBank Publications. Available URL: http://pers-www.wlv.ac.uk/~le1969/

Brindley, G. (1989). Assessing achievement in the learner-centered curriculum**.** Sydney: NCELTR.

Chapelle, C. (1997). CALL in the year 2000: Still in search of research paradigms? *Language Learning and Technology, 1* (1), pp. 19-43.

Clarkson, R. and Jensen, M. (1995). Assessing achievement in English for professional employment programs. Ch. 7 (pp 165 – 194) in Brindley, G. (ed.) *Languageassessment in action.* Sydney: NCELTR.

Cohen, A. (1998). Strategies and Processes in test-taking and SLA. Ch 4 (90 – 111) in Bachman, L. and Cohen, A. (Eds.), *Second language acquisition and languagetesting interfaces*. Cambridge: CUP.

Daugherty, M. & Funke, B. (1998). University faculty and student perceptions of web-based instruction. *Journal of Distance Education*, *13* (1). Available URL: http://cade.athabascau.ca/vol13.1/daugherty.html

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: CUP.

Eggins, S. (1994). *An introduction to systemic functional linguistics*. Continuum: New York.

Gatton, W. (1999). *Call trends: A post-TESOL view*. Available URL: http://www.dyned.com/dyned/japan/htm/trend.htm

Gunn, M. (1995). Criterion-based assessment: a classroom teacher's perspective. Ch. 9 (239 – 270) in Brindley, G. (ed.) *Language assessment in action*. Sydney: NCELTR.

Hara, N. and Kling, R. (2000). Students' distress with a web-based distance education course. In press. Available URL: http://www.slis.indiana.edu/CSI/wp00-01.html

Hegelhiemer, V. and Chapelle, C. (2000). Methodological issues in research on learner-computer interactions in CALL. *Language Learning and Technology, 4* (1), 41-59. Available URL: http://llt.msu.edu/vol4num1/hegchap/default.html

Jones, F. (1998). Self-instruction and success: A learner profile study. *Applied Linguistics, 19* (3), 378 – 406.

Jung, I., & Rha, I. (2000). *The impact of information and communication technology in higher education: Experiences in Korea's virtual university*. Available URL: http://www.com.unisa.edu.au/cccc/pa-pers/non_refereed/jung.htm

Lassche, G. (2000). Web-based Language Learning in Korea: A pedagogical critique. *Korea TESOL Journal*, *3* (1), 55-76. Seoul: KOTESOL.

Lassche, G. (Unpublished manuscript). *Dimensions of authenticity for language testing and teaching: Issues surrounding text, interlocuters, and context.*

Long, M. and Robinson, P. (1996). Focus on form: theory, research and practice. Ch 2 (pp 15 – 41) in Doughty, C. and Williams, J. (eds.) *Focus on form inSLA*. Cambridge: CUP.

Martin, J. (1993). Genre and literacy - modelling context in educational linguistics. *Annual Review of Applied Linguistics, 13*, 141-172.

Martin, J. (2000). Design and practice: enacting functional linguistics in Australia. *Annual Review of Applied Linguistics, 20* (20th Anniversary Volume 'Applied Linguistics as an Emerging Discipline'), 116-126.

Richards, J. and Rodgers, T. (2001). *Approaches and methods in language teaching*. (2nd ed.). Cambridge: CUP.

Rovinelli, R., and Hambleton, R. (1976). On the use of content specialists in the assessment of criterion-referenced test item validity. Paper presented at the annual meeting of *AERA*, San Francisco. Eric Document # ED121845.

Smith, M. and Salam, U. (2000). Web-based courses: a search for industry standards. *CALL-EJ Online, 2* (1). Available URL http://www.lerc.ritsumei.ac.jp/callej/5/msmith&salam.html

White House Papers. (1997). *Report to the President on the Use of Technology to Strengthen K-12 Education in the United States.* Available URL: http://www.whitehouse.gov/WH/EOP/OSTP/NSTC/PCAST/k-12ed.html

Widdowson, H. (1979). *Explorations in applied linguistics*. Oxford: OUP.

Appendix 1: Tutorial Task

Describe an experience you had while traveling.

Write at least 6 sentences.

<archival info> Last summer I travelled around Peru with my friends, Alicia and Eva. We were in Nazca and we wanted to go to Arequipa (south of Peru). This journey took about eight hours so, we decided to do it at night. The people in our hotel made the reservation (1) of our bus tickets. We always do it by ourselves, but they were so kind that we entrust (2) it to them. During the day we heard some things that made us to think (3) there were (4) something wrong but we really couldn't imagine what was in store. We paid a ticket for a comfortable coach, "Royal Class", as they said.We were supposed to leave (5) at 10.00 p.m. but the bus arrived at 11.15 p.m. The first impression wasn't very good. It was a very dark and old bus. (6) We left our backpackes in the boot and got into the bus. There was a door which separated the bus driver from the rest of the seats. I was the first person who openned the door. I was literally knocked out by a terrible smell. It could be said that all livestock of Peru (7) had been there.
<snip half the text>

<tutor's response>

(1) Correct!
(2) entrusted
(3) made us think
(4) was
(5) Correct!
(6) very dark old bus ("and" not necessary in English)
(7) I would re-word this "It was as if all the livestock of Peru..."

This was a very dramatic account of a disastrous journey - I can't help thinking that it was based on the truth? There are a number of minor errors, and some tense and agreement mistakes that you shouldn't be making at your level! (e.g. 4, 8, 12), but the more interesting errors are syntactical and idiomatic (7, 13, 14, 17). On the whole, the story comes across clearly despite these slips - I will give it 14 out of 20.