Student responses to alternative EFL evaluation

Michael Guest
University of Miyazaki

Reference data:

Guest, M. (2011). Student responses to alternative EFL evaluation. In A. Stewart (Ed.), *JALT2010 Conference Proceedings*. Tokyo: JALT.

Using a combination of open-ended classroom surveys (*n*=593) and personal interviews (*n*=56) over a period of six years, first and second year medical students at a national university in Japan were asked to express their feelings regarding the modes and methods of alternative assessment the teacher had been using in a Communication English class. The survey and interviews served as an action research response to the teacher's sense that first year students were not responding to these methods of evaluation as positively or effectively as had been expected. After completing, and subsequently analyzing both the surveys and interviews, it was concluded that many forms of alternative assessment may lack face validity for students, making it difficult for students to choose effective study or performance strategies. It was also noted that while the negative responses were more frequent among first year students, these negative comments decreased among second year students.

国立大学の1、2年生の医学生に対して、筆者がコミュニケーション英語の授業で　取り入れてきた「代替評価」の手段・手法について、クラス調査 (593件) と個 人面接 (56人)を過去6年間に渡って実施し、感想を求めた。調査と面接は、1年次はこの代替評価に関して肯定的、効果的な反応を示すのは困難であろうという筆者の認識に呼応したアクション・リサーチとして機能している。実施と分析を重ねた結果、多用な代替評価の様式は、学生側にとって「表面的妥当性」に欠ける可能性があることが判明した。すなわち、効果的な学習や表現ストラテジーの選択を困難にする可能性がある。また、1年次のほうがより否定的な反応を示す一方で、これらの否定的コメントは、2年次になると、減少する ことも指摘している。

There seems to be a widespread, popular belief among EFL teachers at the tertiary level that Japanese students would like to be liberated from drudgery that is allegedly based upon rote memorization, the mechanical inculcation of grammatical minutiae, and an emphasis upon teacher-centered receptive activities such as de-contextualized drills and choral repetition (Mulvey, 1999; Yoshida, 2001; Murphey, 2009). One might expect then, that after having endured the long period of preparation for and undertaking of university entrance examinations, university students would respond positively to less traditional, more productive and interactive tasks and activities. This would, of course, include the manner in which course evaluation is conducted. The problem is that I had been carrying out various types of progressive or alternative evaluations—that is, evaluations not utilizing receptive, discrete-item paper-based tests, for first and second year university medical students for seven years but sensed that students were not happy, and in fact were often visibly frustrated, with the forms and methods of evaluation being used. The purpose of this research was to respond to this classroom dilemma which appeared to contradict my expectations about how students would respond to new or alternative forms of assessment. It was, in short, action research.

Oller (1979) was one the earliest to distinguish between discrete point and integrative testing in the field of languages. Since that time, alternative assessment has become widely applied in ESL/EFL circles. But what does the term imply? Huerta-Macias (1995) referred to alternative assessment as consisting of evaluating students based upon "…what they produce and integrate rather than what they can recall or reproduce" as well as "…how students are approaching, processing, and completing real-life tasks in a particular domain" (p. 9). Hamayan (1995) offers the following as five salient characteristics of alternative assessment:

1. Proximity to actual language use and performance
2. A holistic view of language
3. An integrative view of language
4. Developmental appropriateness (in terms of the cognitive, academic, or social needs of the learner)
5. Multiple referencing (the measuring of multiple competencies)

Brown and Hudson (1998) divided assessment into three categories with two of them, constructed-response and personal-response, falling under the rubric of alternative assessment as used in this research.

All of these qualities are incorporated into the definition of alternative assessment used here. More specifically, I am referring to the following forms and features, all of which my students faced during first year English classes that I conducted:

1. Ongoing assessment of multiple skills/competencies, as opposed to one final test focusing upon only one competency, namely memorization of discrete points. Hancock (1994) distinguishes between the notion of assessment, as a cumulative practice tied into a curriculum, versus the notion of periodic self-contained, achievement-oriented testing.
2. Open-book, open-note formats, including the distribution of successful samples of past tests.
3. Open-ended (not forced choice/discrete item) holistic tasks predominating. The testing locus was upon the consolidation of numerous language skills and cognitive levels towards a communicatively meaningful purpose.
4. Dynamic forms (role-play; real-time interviews, etc.) predominating. These were often student-generated, carried out in real-time, being neither scripted nor mere teacher/text produced repetition.
5. Focus upon the cognitive levels of recall and (re)production—not merely recognition. This provides an emphasis upon active, productive testing.
6. Student-generated content, themes, or questions being incorporated into tasks—Students always being given choices and license to develop and manipulate contexts. In some cases, students even created the content of the class tests themselves, by committee
7. Student-selected summaries/reviews of content—Students being asked to list and explain main points of interest and value that they had learned during the course period. This could be understood as a type of summary portfolio.
8. Personal interview (including self-reflection)—This involved a summarization of students' own strengths, weaknesses, and creating a plan for action regarding future English study one-on-one with the teacher.
9. Cooperative development; peer checking and sharing—Many evaluated projects involved teamwork plus both pre- and post-task checking/explanation with other groups

of students. The centrality of cooperative activities to alternative assessment has been explicated in particular by Calderon and Hertz-Lazarowitz (1992).

10. Peer/self assessment—Students were encouraged to comment on others' performances and reflect upon their own performances on various evaluated tasks. Carrying these acts out was included as a part of the total course evaluation.
11. Diagnostic feedback post-test, reviews, re-tests, and model answers—all designed with the purpose of using evaluation as a curriculum-driven pedagogical tool by making students more conscious of weaknesses and strengths, plus offering opportunities to address these, in accordance with principles laid out by both Spolsky (1992) and Mitchell (1992).

Several compendiums of activities connected to alternative assessment (sometimes referred to as authentic and/or performance assessment) also exist, outlining approaches and suggested methods of implementation. Most prominent among these are Herman, Aschbacher, and Winters (1992), Brown (1998), and Law and Eckes (2007), all of whom have informed what I practice in the classroom.

Methods

This research focuses upon first and second year medical students at a national university in the required courses Communication English 1 and 2. As all medical students are required to take this course, it would seem to provide a sound base from which to secure a comprehensive, well-rounded representation of student views and attitudes. However, it should be noted that although this means I teach all medical students at least once, a number of them entered my class two, or even three times, over the their first two years in the Communication English course. Thus, some student responses were duplicated in the surveys and/or interviews (although this does not necessarily mean that the actual comments and opinions expressed were duplicated, as students' perspectives and opinions can and do change even over just a short time).

Surveys

Standardized classroom surveys used at my university contain twelve set questions based on a Likert scale, followed by a space for extra comments on any topic deemed necessary by the teacher. It was this open section that I focused upon. I gave students a full 30 minutes to write responses regarding their feelings about the various types of evaluation they had experienced in my class (in addition to the standardized twelve Likert-scale questions which were not connected to the additional comments which provide the foundation for this research). Instructions were provided in both Japanese and English and responses were also acceptable in either.

This response section was guided by six explicit questions I explained in Japanese and English and wrote on the board in English (students were not required to respond to all of the questions—they could choose any number upon which they wanted to comment). The six questions were:

1. Have you ever experienced this type of English evaluation before?
2. Did you find the evaluation tasks and methods 1) helpful 2) interesting 3) challenging? Explain why or why not.
3. How do you feel about these methods, as compared to traditional tests?
4. Do you think that the tests helped you improve your English? Explain how or why.

5. Do you think the tests were a fair assessment of your English skills?
6. Do you have any suggestions for future test types?

These surveys were conducted once each semester towards the end of the regular class sessions (twice per academic year), over a period of six years.

Interviews

Over the same number of years I invited 56 first- and second-year students (generally those involved in English clubs and specialist English courses with whom I was familiar) to discuss the Communication English course evaluation methods in a small tutorial room. Students could respond in either Japanese or English and the conversation often drifted between the two languages. On eight occasions two students attended as a pair but most (40) were solo interviews. During these interviews I asked several questions distinct from those asked on the surveys. Most common among these were: Which test types were you familiar with before entering university? Were you surprised by any test type? If yes, by what aspect? Do you think each test type helped to reflect your actual English ability? Did the different test types help you improve your English skills? How did you prepare for the tests? In general, please tell me any positives or negatives you have regarding the various tests you've taken in these courses.

Students did not always answer all of these questions, nor was there always time to pose all these questions, or in this particular order. Sometimes the conversation meandered. I made quick notes during these interviews. The students interviewed had also completed the surveys, so it should be noted that these do not comprise two distinct sets of subjects.

Results

Survey results

Widespread responses to both survey and interview questions addressed to students were treated as significant but no rigid statistical analysis was applied, since the open-ended nature of the survey and interview responses did not allow for simple numerical classification. The twelve Likert questions contained on the surveys mentioned earlier were standardized university questions that did not address the researcher's interest and thus played no role in the subsequent research.

Previous experience with alternative assessment in English

Based upon the first interview question, "*Which test types were you familiar with before entering university?*", my students' previous experiences of English evaluation, seem to have consisted largely of the following types: Discrete-item based paper tests, which were non-interactive (one student, one paper sheet), and which were inevitably carried out in either the course final class or post-class testing season, meaning students received only a numerical grade or pass/fail and little diagnostic feedback. A focus upon memorization (receptive, recognition-based). Although some students did productive English tasks in class, very few mentioned production in assessment. Open-book tests were virtually unknown and the mode was generally receptive, testing memory, and pitched at the lowest cognitive level—that of recognition. Tasks consisting primarily of multiple (forced) choice or fill-in-the-blanks format (passive), or sentence restructuring/reordering. Language was invariably de-contextualized and almost never generated by the students. Some assessment of essays and "reports," although further inquiry (in the interviews) revealed very few, if any, of these instances to be process writing, involving revisions and schema development, but rather one-off paragraphs.

This establishes the fact that very few first-year students had hitherto been exposed to alternative methods of assessment, whereas second-year students had gained exposure during their first year. Comparing first- and second-year student responses allows us to measure those who had no background with alternative assessment with those who had one year's exposure.

Over the six-year period, a total of 593 surveys were returned. Of these, 261 did not address the questions adequately or comment meaningfully (containing glib or superficial comments such as, "I liked the teacher's tests"). Of the remaining 332 surveys that were considered significant for this study, 208 were from first year students who were thus unfamiliar with my teaching and evaluation habits and standards. The other 124 were second-year students who were at least somewhat familiar with my teaching and evaluations.

Of these 208 first-year students, only 23-27 (the imprecise number is due to subjective definitions) had experienced any type of alternative evaluation previously, regardless of previous institution. This means that there were 181-185 students (just under 90%) who were new to such methods of assessment.

First year students

In regard to the question, "*Did you find the evaluation tasks and methods 1) helpful 2) interesting 3) challenging? Explain why or why not*", there were a large number of negative or otherwise critical responses but almost exclusively among first-year students, and particularly from their first semester surveys. Among the most common of these were as follows (sample quotations below represent a summarization of common or widespread comments or have been altered syntactically to conform to accepted norms):

> "I wasn't familiar with the format"
>
> "I don't like depending on other people"
>
> "I didn't know exactly what the teacher wanted me to show"
>
> "I was nervous about not knowing exactly what the teacher wanted"
>
> "Real-time English is very stressful"
>
> "I couldn't think of any good ideas for the role-play"

The number of comments expressing the sentiments above led me to realize that being in a new situation that involved something as stressful as an evaluation increased the sense of frustration for students. Not being familiar with formats, expectations, and being unused to performance or production-based testing did not motivate or excite many students, who perhaps felt more comfortable within the familiar confines of more traditional exams.

However, in regard to the question, "*Do you think that the tests helped you improve your English? Explain how or why" many interesting comments emerge, even from students who, in the previous question, had expressed anxiety or frustration with the evaluations:*

> "I had to think about making use of my new English"
>
> "I had to think actively about content"
>
> "I could learn from working with other students"

Here we can see some positive academic habits emerging. Keeping good notes and reviewing are not only widely regarded as good study habits but are also important steps in terms of developing learner autonomy (Allwright, 1988; Yoshida 2001). Cognitive engagement with the content can also be seen to be emerging.

Second year students

The number of positive comments was noticeably higher among second-year students than first. For example, in regard to the question, *"Do you think the tests were a fair assessment of your English skills?"* almost all second year students responded positively. Another frequently expressed opinion was:

> "Because we do many kinds of tests everybody can have a chance to do well".

Among the most common responses from second year students when asked if the tests had helped them improve their English skills, common responses included:

> "I had to keep good notes all year (which) is good for me"

> "The (summary) test made me review everything I had studied"

Therefore, it seems there exists a recognition that a variety of skills are being tested and that this is perceived as beneficial to more students in the second year group. That students are gradually being weaned off the notion that the memorization of discrete points is the central, or only, skill necessary for second language acquisition is apparent.

Interview results

First year students

Responses from interviews yielded results similar to those noted in the first year students' survey responses, in which students expressed elements of frustration or confusion. For example, note the common response to the question: *"Do you think each test type helped to reflect your actual English ability?"*:

> "On the speaking tests I was very nervous and couldn't think clearly so I couldn't say anything"

Unfamiliarity with the testing locus seems to lead to ineffective performance, at least from the students' own perspective. A second common comment expressed: *"Students who didn't study hard could see the textbook, so it's not fair to students who worked hard to remember and study"*

Again, students expressed the notion that open-book or open-note testing is not "real" testing, as it lacks face validity, or does not meet student expectations of what a test should look like, given that the notion of what constitutes testing for many students had hitherto involved the memorization of textbook contents. A further question on the same topic, *"Were you surprised by any test type? If yes, by what aspect?" reveals a similar element of surprise:"I was surprised that we could look at our textbook and notes".*

Second year students

By the second year a more positive attitude towards this testing format emerges as seen in the widespread comment: *"I was happy that you showed us previous successful tests. Teachers usually don't show us that kind of thing"*

The interview responses also reaffirmed the fact that second year students had begun to appreciate and adapt themselves and their study habits to the new test formats as seen in the responses to the question: *"Did the different test types help you improve your English skills?"*

> "I had to think about everything deeply"

> "We had to use all our English skills. We couldn't rely on just one or two things, like memory or vocabulary power"

The same phenomena can be noted in the differences between first- and second-year student responses to the question, *"How did you prepare for the tests?" In the first year, "We let the strongest member of our team prepare most of it"* was not an uncommon response, as were responses along the lines of "*memorizing or re-reading the textbook*". But among second year students the following responses, indicating the inculcation of more academically fruitful skills, were far more common:

> "I highlighted things in the textbook and on the prints that the teacher talked most about and studied those"

> "I looked at the model from last year"

> "I used what we had practiced before as a basis"

In fact, a change in response from memorizing textbooks or depending upon strong team members (first year) to these more viable academic study methods (second year) was directly noted by three of students who were interviewed in both years.

Discussion

As we have seen, the 1st year students offered mixed reactions regarding the alternative evaluation styles. The most common negative reactions/responses were that they:

- Did not understand task targets/purposes well
- Did not understand grading criteria well
- Did not understand test format/administration
- Had studied inappropriately or focused on unhelpful 'skills'
- Held an overdependence upon strong partners

Lack of familiarity with testing locus, formats, and criteria leads to anxiety and frustration, even if the testing formats are pedagogically sound. This brings us to the question of face validity. If students do not see a test as living up to their expectation as to what a test should be, it might affect performance (noting the difficulties that many first year students, who showed communicative English competency outside the testing situation, displayed on the assessments), regardless of actual student skill or ability. Newfields (2002) has argued that such face validity is merely a cosmetic construct that should have no bearing on deeper issues of test validity but Roberts (2000) claims that a lack of consideration for face validity can affect evaluation outcomes, since students will not be psychologically in tune with the type of measurement being employed. This view appears to be initially consistent with my own findings. Teachers must find ways of increasing face validity by providing students with sufficient explanation as to the role, criteria, and function of the particular evaluation they are undertaking without resorting to the familiar test formats, as a lack of a background using productive study habits or interactive tasks seemed to hamper students.

Survey and interview results indicated that those first year students who had already experienced alternative testing before entering university enjoyed and knew how to prepare for (as well as manage) the tests. Not surprisingly, twelve of the fourteen first year students interviewed, who reported experiencing various forms of alternative assessment for English in the past, ranked as the top three students in the English Communication 1 course in their respective first years. This raises the issue of test-wiseness—whether some students are gaining higher grades simply because of a familiarity with the test format or character issues that allow them to negotiate testing procedures and formats better than others. Brown and Yamashita (1995) see an emphasis upon measuring test-wiseness and reducing the validity of university entrance examinations in Japan. Given that entrance exams are highly procedural we might assume that most testing strategies that require fairly detailed procedures and formatting—such as many of the alternative methods we have discussed in this paper—may also be easier for students familiar with their formats. This offers a preliminary answer to the original classroom action research question as to why students who were obviously quite fluent in English in practice often performed poorly on the actual tests, particularly in first year courses.

On the other hand, second year students were uniformly far more positive regarding alternative testing than first year students. From the responses I noted that:

- More second year students had adapted their study habits to suit the test. More second year students saw the greater long-term educational value of alternative testing.
- More second year students had adapted diagnostic feedback into their subsequent studies.
- More second year students embraced the autonomous and productive elements (cognitive engagement) of alternative testing

Familiarity with the format and criteria seems to have lead to the inculcation of better academic skills, productive study habits, a greater sense of learner autonomy, and cognitive engagement of content. Meta-cognitive skills which will undoubtedly be useful in future study and research also seem to have developed.

Five implications for teachers and course/materials designers and conclusion

1. Evaluation/task content must be made very clear (previous successful models and samples, detailed explanations and outlines should be provided). This will reduce anxieties regarding expectations and standards as well as any claims of unfairness based upon catering to students already familiar with the testing system.
2. Criteria and focus must be made clear (text/study references, grading focus, level of expectation should be made explicit). Whole classes can and should be allotted to detailed preparation. Specific study referents, skills, and grading formula should be explained, preferably in both written and spoken text. This echoes advice given by Wiggins (1994).
3. Feedback and chances for correction and/or redemption are crucial to ensure fairness and skill development. This would favor ongoing assessment over one-time course-ending evaluation.
4. A large number of test types should be used. A well-rounded evaluation should measure a number of skills and skill-types in order to provide for a holistic framework. Certain personalities or students with limited skills (such as good memories for detail) should not be favored. Although having a large number of test types will likely mean that students have to adapt to and become familiar with a wider variety of testing forms, skills, and evaluation methods, addressing a wide variety of skills and types would allow for a more well-rounded, accurate assessment of student performance than a single type. It would also help to inculcate a wider range of useful academic skills, making it more productive pedagogically.
5. A good test should encourage improved academic/study skills and learner autonomy but it takes a considerable

amount of time for these to develop. Tests should be empowering and enabling, and their diagnostic and pedagogical functions should be made manifest in future study habits.

This study indicates that teachers shouldn't expect the benefits of alternative testing to be immediately apparent among students who are not yet used to these types of evaluation, a point that is consistent with the recommendations made by Worthen (1993) when implementing forms of alternative assessment. Rather, as maturing learners students must grow into this different kind of testing, and by implementing the suggestions listed above, the teacher can guide them in a more beneficial direction.

References

Allwright, R. (1988). Autonomy and individuation in whole class instruction. In A. Brooks & P. Grundy (Eds.), *Individuation and autonomy in language learning* (pp. 35-44). London: Modern English Publications and the British Council.

Brown, J. D. (Ed.). (1998). *New ways of classroom assessment*. Alexandria: Teachers of English to Speakers of Other Languages.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, *32* (4), 653-675.

Brown, J.D., & Yamashita, S. (1995). English language entrance examinations at Japanese universities: What do we know about them? JALT Journal, 17 (1), 7-30.

Calderon, M., & Hertz-Lazarowitz, R. (1992). Dynamic assessment of teachers and language minority students through cooperative learning. *Cooperative Learning*, *13* (1), 27-29.

Hamayan, E.V. (1995). Approaches to alternative assessment. *Annual Review of Applied Linguistics*, *15*, 212-226.

Hancock, C. R. (1994). Alternative assessment and second language study: What and why? *The Ohio State University Online Digests*. Retrieved from <http://www.cal.org/resources/digest/hancoc01.html>.

Herman, J., Aschbacher, P., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria: Association for Supervision and Curriculum Development.

Huerta-Macias, A. (1995). Alternative assessment: Responses to commonly asked questions. *TESOL Journal*, *5*, 8-10.

Law, B., & Eckes, M. (2007). *Assessment and ESL: An alternative approach*. Winnipeg: Portage & Main Press.

Mitchell, R. (1992). *Testing for learning*. New York: Free Press/Macmillan.

Murphey, T. (2009). Innovative school-based oral testing in Asia. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *13* (1), 14-21.

Mulvey, B. (1999). A myth of influence: Japanese university entrance exams and their effect on junior and senior high school reading pedagogy. *JALT Journal*, *21* (1), 125-142.

Newfields, T. (2002). Challenging the notion of face validity. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *6* (3),14.

Oller, J.W., Jr. (1979). *Language tests at school*. London: Longman.

Roberts, D. M. (2000). Face validity: Is there a place for this in measurement? *Shiken: JALT Testing & Evaluation SIG Newsletter*, *4*(2), 6-7.

Spolsky, B. (1992). Diagnostic testing revisited. In E. Shohamy & R.A. Walton (Eds.), *Language assessment and feedback: Testing and other strategies* (pp. 29-39). Dubuque: Kendall/Hunt Publishing Co.

Wiggins, G. (1994). Toward more authentic assessment of language performances. In C. R. Hancock (Ed.), *Teaching, testing, and assessment: Making the connection. Northeast conference reports.* Lincolnwood: National Textbook Co.

Worthen, B. (1993). Is your school ready for alternative assessment? *Phi Delta Kappan*, *74* (6), 455-456.

Yoshida, Y. (2001). Authentic progress assessment of oral language: Oral portfolios. Retrieved from National Library of Education, ERIC database, Item ED453674.