Feature Article

Are We on the Same Page? Teacher Assessment vs. Student Self-assessment

Mike Guest
University of Miyazaki

This paper reports upon the results of classroom research undertaken by the teacher/researcher regarding qualitative disparities between student self-assessment and teacher assessment. After completing a graded performance class in a communicative English course, first- and second-year medical students were asked to assess their own performance based upon 1) a series of itemized competencies connected to the task, and 2) a holistic score based upon task criteria that had been given to students prior to task performance. It was discovered that student emphasis upon, and self-assessment of, specific competencies differed considerably from the teacher's evaluations of those competencies. It was also noted that self-assessed holistic scores provided by students were inconsistent with their assessment of itemized competencies. It is therefore argued that student notions of competence and the prioritizing of competencies might not match those of teachers, and further, that students may not have a firm grasp of the difference between holistic and itemized grading.

本稿は、医学科１，２年生向けコミュニケーション英語の授業における学生自身の自己評価、および教員による学生評価の両者に見られる差異についてまとめたものである。学生には、段階的なパフォーマンス・タスクを学び終えた段階で、１）既習の個々のタスクに直結した表現能力の項目別評価と、２）事前に渡しているそれぞれのタスク基準に基づいた全体的な自己評価について、それぞれ回答させた。その結果、教員の評価と学生自身による自己評価には、かなりの差が認められた。また、個別項目に設けた学生による個別項目自己評価は、同じく学生自身による全体的評価の結果とは必ずしも一致するものではなかった。このことにより、学生が考える表現能力および能

力評価の優先事項は、教員が求めるものとは合致しない可能性があること、また学生は個別項目評価と全体的評価の違いについての認識が浅い可能性があり、これらの点について議論したい。

Introduction

This research into student self-assessment (SA) versus teacher assessment (TA) was motivated by the fact that the teacher/researcher had become subject to an increasing number of disputes or claims regarding required English test results among first- and second-year medical students. The research hypothesis was twofold: 1) to determine to what degree student perspective of grades given for excellent and mediocre performances were at odds with the teacher's standards and, further, 2) to determine if student notions about prioritizing criteria in assessing language performance also differed from that of the teacher. It is believed that recognizing how and where these differences occur might allow teachers to better establish and explain testing criteria and scoring mechanisms to students in the future as well as allow for a higher standard of critical feedback to students, thus mitigating the potential for misunderstandings and friction.

Background

Student SA has come to be regarded as a valuable component of the classroom evaluation process over the past twenty-plus years. Bachman and Palmer (1989) and Peirce, Swain, and Hart (1993) were among the earliest proponents, arguing that learner self-assessment can be a reliable and valid measure of communication. Other researchers have since ascribed numerous benefits to SA. Among these are what Munoz and Alvaraz (2010) and Dickinson (1997) describe as an enhanced dialogue between teacher and students that can provide useful feedback for both. Butler and Lee (2010) argue that SA enables students to clearly understand task goals and what is required to achieve those goals. O'Malley and Chamot (1993) and Oscarson (1989) report an increased meta-cognitive awareness of the students' own weak and strong points. Munoz and Alvaraz (2010) argue that this meta-cognitive awareness triggers reflection upon the existing target language system of the student, raising consciousness regarding target language shortcomings and needs. Harris (1997) and McNamara (2001) see SA as tied

to an increase in learner autonomy such that learners come to see themselves as active constructors of their own learning and thus develop a greater sense of responsibility for their language skills development. Heilenman (1990) views SA as a means of minimizing teacher-student disputes and subsequent breakdown of mutual trust and understanding.

As a result of the influence of this research, it is now more common for EFL/ESL teachers to have established some type of SA as a standard part of their testing processes. But is there still a gap between teacher and student perceptions regarding both purpose and process?

Methods

The teaching and research setting discussed in this paper involved four first-year English Communication (required course) classes over two semesters (total = 105 students), and three second-year Medical English classes (also required, total = 102 students) over the same period. The test used to undertake the research was a role-play test (course value = 40%) having an emphasis upon oral/aural skills, but also requiring the ability to write medical data accurately and legibly on a medical chart.

The form and content of these tests differed slightly according to the year. A brief description of each follows.

First-year Communication English role-play test

This involved pairs of students, with one playing the role of a doctor carrying out a patient history on another student playing a patient. These roles were reversed thereafter. Students playing doctors were required to be able to gain basic data from the patient and record it on a prepared chart, carry out a basic, orderly patient history, and begin a preliminary inquiry into onset and physical systems to isolate the patient's symptoms (this information, too, was to be charted). The doctor role-playing student had no idea of the patient's condition beforehand and did not even know who their partner would be until they were called in to the examination room two-by-two. Thus, they had to be prepared both as a doctor and as a patient. The language and target forms had been taught and practiced in advance in regular course classes, largely based upon models and tasks found in

the course textbook *English in Medicine* (Glendinning and Holmstrom, 2010).

Second-year Medical English role-play test

This examination involved a three-stage, three-person role-play developed and practiced completely by the team members themselves in advance. The three stages included 1) a General Practitioner (GP)-Patient initial consultation, 2) a GP-Specialist reference telephone call, and 3) a Specialist-Patient consultation. Given the second-year students' presumed superior knowledge of both medicine and medical English, it was expected that these students would delve deeper into medical complexity than could be expected of first-year students. Role-play content was also to be based upon models and tasks previously practiced in class and also found in the same textbook as in first year.

Both courses were represented in this research in order to note any significant differences between first- and second-year medical students, with the expectation that second-year students would be more used to the teacher/researcher's teaching and testing style and thus should have developed slightly more sophisticated academic skills by their second year.

Both the testing format and grading criteria were made explicit to all students in advance using concrete samples and past models of success to illustrate. The overall grading value of all the tests was 40% of the course grade. The criteria (Table 1) were explained verbally and distributed in advance.

Post-test self-assessment forms (Appendices A and B) were given to each student immediately following the test and before they had received any teacher feedback or assessment. Students were given 10 to 15 minutes to complete the form and return it. This was done anonymously. The two forms differed slightly to reflect the slightly differing tasks assigned to first- and second-year students. The first section is divided into eleven different itemized competencies students were to rate on a Likert scale of 1 (poor) to 5 (excellent). The second section refers to the holistic test score, a self-evaluation from 0 to 40, of their overall test performance. It should be noted that the itemized competencies used in the post-exam SA do not correspond directly with the grading criteria given to the students prior to the examinations, and that students were explicitly told that the

Table 1
Student self-testing criteria for role-play evaluation

Test criteria	Student year*
Speed	1,2
Personal style/interactions	1,2
Control of grammar and basic vocabulary	1,2
Use of new medical vocabulary	1,2
Accuracy and professionalism of chart	1,2
Complexity of chart	2
Consistency of medical details	1,2
Logical order and direction of questions	1,2
Critical thinking; ability to "read" the patient's symptoms	1
Creativity and complexity	2
Quality, professionalism, and accuracy of reference letter	2
Ability to separate and use medical vocab. and general vocab. appropriately	2

*First-year students = 1, Second-year students = 2

total SA score was not to be calculated as a mere accumulation of the itemized competencies listed.

Results

Table 2 shows the results of the post-test self-assessment forms from first-year medical students, with each itemized competency score averaged from a total of 105 students (with all numbers rounded off to the nearest 0.1).

Table 3 shows the results from second-year medical students, averaged from a total of 102 students (34 teams of 3).

Table 4 shows the average total score per class year according to students' total score self-assessment (three invalid responses from first-year students were not calculated).

Table 5 shows the average total score per class year as actually assessed by the teacher.

At least two strengths and weaknesses per student performance (first-year) and per team performance (second-year) were cited and provided by the teacher to students as feedback. Table 6 shows the three itemized competencies most

Table 2
Itemized Competencies (first-year)

Competency	Ave. score out of 5
My speed	3.2
My personal style/interactions	3.4
My control of grammar and basic vocabulary	2.3
My use of new or medical vocabulary	2.5
My chart was professional in style	3.3
My chart was easy to follow and understand	3.3
My questions were in a logical order	3.6
I covered all the main points	2.6
I understood the patient's problem well and asked relevant questions	3.0
I asked original and intelligent questions about the affected system	2.8
As a patient my problem made sense and was consistent	3.6
Average per competency	3.05

Table 3
Itemized Competencies (first-year)

Competency	Ave. score out of 5
My speed	3.7
My personal style/interactions	3.4
My control of grammar and basic vocabulary	2.4
Our use of new or medical vocabulary	2.9
Our chart was complex and professional in style	3.5
Our reference letter was professional and complete	3.8
Our role-play questions were in a logical order	3.8
We covered all the main or necessary points	3.8
Our patient's problem was interesting and complex	3.5
We asked original and intelligent questions about the condition	3.9
The patient's problem made sense and was consistent	3.7
Average per competency	3.5

Table 4
Average total SA scores

Year	Ave. total score out of 40
First year	30.3
Second year	30.1

Table 5
Average total TA scores

Year	Ave. total score out of 40
First year	33.6[1]
Second year	31.4

frequently cited by the teacher as weaknesses to first-year students (n = 105).

Table 6
Itemized competencies cited as weakness to 1Y (top 3)

Competency	No. of citations
Questions were not in a logical order	36
I understood the patient's problem well and asked relevant questions	28
As a patient my problem made sense and was consistent	28

Table 7 shows the itemized competencies most frequently cited by the teacher as weaknesses to second-year students (n = 102, resulting in 34 team grades).

Table 7
Itemized competencies cited as weakness to 2Y (top 3)

Competency	No. of citations
We covered all the main or necessary points	14
Personal style/interactions	11
The patient's problem made sense and was consistent	10

The two itemized competencies most frequently noted as strengths by the teacher are noted in Tables 8 (first-year) and 9 (second-year).

Table 8
Itemized competencies cited as strengths (first year)

Competency	No. of citations
My speed	over 65
My control of grammar and basic vocabulary	over 60

Table 9
Itemized competencies cited as strengths (second year)

Competency	No. of citations
Our use of new or medical vocabulary	28
Our chart was complex and professional in style	24

Discussion

Disparities between itemized competency SA and overall score

What is initially most striking in the student SAs is the disparity between the average score of the itemized competencies (3.05/5, or 61%) and the expected overall score (33.6/40 or 84%). While many, if not most, of the itemized competency SA scores were relatively modest, the expected overall score was considerably higher than one would expect. Although the overall grade was understood not to be a mere cumulative totaling of the individual competencies, the paradoxical notion that an admittedly average performance competency-wise should result in a total score that reflects a "very good to excellent" rating is perplexing.

The disparity between SA and TA becomes even more pronounced when it is noted how many individual student responses contained total scores that seem logically untenable or dubious. For example, among the first-year medical students 26 of the 105 respondents (just under 25%) graded themselves on the itemized competencies with an average of under 3.5 per item (moderate) and yet gave their expected overall score as 35 or more out of 40 (excellent).

If this phenomenon had been limited to just a few students (as with three cases in which students mistakenly totaled all the specific skill ratings to calculate an overall score which exceeded 100% and were thus disqualified from the research sample), one might dismiss it as being an anomaly, perhaps a misreading or misunderstanding of the overall scoring value or related criteria. Instead, there were a significant number of students who rated themselves as average in performance per itemized competency yet somehow believed they had achieved excellence in terms of an overall grade. Why the inconsistency?

Possible explanations and interpretations

One possible explanation for these phenomena might be that respondents are rating itemized skills in comparison to some ideal, perhaps in comparison to a native English speaker, and were thus modest in self-assessments of their discrete skills. Saito and Fujita (2004) raise the possibility of cultural factors such as modesty affecting SA outcomes, although Chen (2008) reports TA-SA equilibrium in a similar Taiwanese setting.

In grading their overall test result, it is also possible that students held the belief that if the role-play was successfully completed and that both content and meaning were conveyed, a high score is warranted; in other words, if the task was completed successfully, even if flawed, full marks or thereabouts are expected. A process-based notion that "we tried our best and managed to complete it" thus may weigh more heavily as a criterion for praise and/or reward in the students' minds than in that of the product-focused teacher.

It is also possible that when assessing, teachers may be more concerned with noting specific flaws in the task than maintaining a more holistic sense of the students having successfully completed the task. In short, for a number of students, the criteria used in providing an overall SA score may differ from that of the teacher, even if the specific criteria are given to students in advance (as it was in this assessment). Similar phenomena, resulting in low correlations between student SA and TA, have been noted by Blue (1988), Oscarson (1997) and Patri (2002), even when strict set criteria have been provided.

Grammar/Vocabulary as salient factor; meta-cognition as insignificant

These results may also be partially explained when noting exactly which itemized competencies were rated high or low by the teacher and the students. Interestingly, the language control competency (grammar and vocabulary) was given a uniformly low rating by both first- and second-year students (2.3 and 2.4, respectively, the lowest SA of all itemized competencies), contrasting widely with the teacher's assessments regarding performance strengths and weaknesses. In fact, language control was the item cited second most frequently as a strength

by the teacher for the first-year students.

How might we explain this disparity? One possibility is that students, based on teaching methodologies they have been previously exposed to, believe that grammar plus vocabulary alone constitutes the essence of language. If they feel that they lack general English competency, they may apply a harsher assessment of their own grammar and vocabulary levels simply because these categories loom large in their minds as critical determiners of overall English competency. Therefore, if one's English task performance is in any way flawed, then, it is possible that such a learner will believe that grammar and vocabulary must be the culprits.

This point is augmented by the fact that the results seem to indicate a lack of meta-cognitive awareness exhibited by students regarding interpersonal and strategic functions of language. The SA of itemized competencies such as "logical order" and "relevance" of questions fell at the opposite end of the scale from "language control." They tended to rate themselves somewhat highly (3.6, the highest, and 3.0, respectively) on these strategic competencies, while the teacher felt that these competencies were in fact the two most prominent weak spots.

It is quite possible that students are simply unaware of, or underestimate, the importance of these features in performance and not only tend to dismiss them not only in terms of SA but also as a part of their role-play study or test preparation, despite the explicitly stated criteria established by the teacher. One may therefore suggest that students need to be better informed of the importance of developing strategic skills and that their meta-cognitive grasp of language acquisition, especially at the university level, needs to be addressed.

As a response, the researcher has since initiated an orientation session focusing upon developing English communication skills for incoming university students in which these qualities were communicated. It will be interesting to compare if these newer students differ from their teacher in terms of assessment as much their predecessors did.

Possible design flaws

The discrepancies between SA and TA may also indicate that a minor design flaw in the SA form has affected the outcomes. Even though all SA itemized competencies were explicitly stated to be related to this test performance only, it is possible that some respondents instead rated their general skill levels when responding and did not limit their self-evaluation to this single performance. Care must be taken to avoid this possibility when distributing future SA forms perhaps by clearly stating on the form itself that the SA is limited only to the single performance.

Furthermore, it is quite possible that, despite explicit teacher explanations in both English and Japanese, several competencies overlap in meaning and/or function and thus may confuse students. When delineating and categorizing specific competencies, teachers should take into consideration Jansen-van Dieten's (1989) claim that more concrete and descriptive scales are better than global or holistic scales, coupled with the Munoz and Alvaraz (2010) claim that the more abstract concepts are simply less graspable to students.

First- vs. second-year student results

As for differences between the first-year and second-year student SA, it is significant that the disparity between SA and TA total score was slightly more pronounced among first-year medical students.

Why might this be the case? First, the higher correlation between SA and TA total scores may be because second-year students will be more familiar with the teacher's standards even without having completed an SA under that teacher previously. Second-year SAs also indicated a more realistic balance between itemized competency scores and their overall score, which may show that the connection between the competencies and overall performance was clearer to these slightly more mature students. Moreover, the TA results indicate that while second-year students had developed some new skills (chart writing and medical English usage), they still did not view their performance in more holistic or meta-cognitive competencies as being as weak as the teacher did.

Conclusions

These results seem to indicate that teachers and students may hold a different understanding as to how a holistic communicative task score should be calculated, as well as a different understanding of the grading criteria, even when the criteria is made explicit to them in advance. It would not be sufficient to simply provide students with a list of the testing criteria before an exam. Nunan and Carter (2001) argue that students need training and scaffolding in the SA process, and therefore that these criteria should be emphasized and highlighted throughout the entire course and not just before the examination.

Boud (1995), Brown (2004), and Nicol and MacFarlane-Dick (2006) all maintain that regular and consistent feedback of this sort should not only lead to more consistent SA results but also improve learning outcomes by raising the meta-cognitive awareness of students, allowing them to consciously access a more sophisticated and complex language system and thereby understand a language as being more than a simple grammar-vocabulary slot and filler model.

At the university level in particular, incoming students need to raise consciousness regarding the important role of meta-cognitive English skills, such as interpersonal features, creative and critical thinking, and rhetorical organization, but can only develop these if teachers consistently and regularly emphasize their importance.

Teachers must also help students become conscious of the distinction between process and product and be explicit as to where on this continuum test or course grades will be focused. Pre-course orientation seminars or extended introductory lessons could address these issues and thereby help avert possible teacher-student misunderstandings and disputes.

Note

1 The fact that the combined scores of all the itemized competencies (11 items rated on a scale of 1 to 5) equaled the average total score (full score =40) for 1st year students is a statistical coincidence. Care was immediately taken in the research process to make sure that students were not simply totaling their itemized competency scores to produce the self-evaluated test score.

References

Bachman, L., & Palmer, A. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing, 6*, 14-25.

Blue, G. M. (1988). Self-assessment: The limits of learner independence. In A. Brookes & P. Grundy (Eds.), *Individualisation and autonomy in language learning*. ELT Documents 131 (pp. 100-118). London, England: Modern English Publications/British Council.

Boud, D. (1995). *Enhancing learning through self-assessment*. London, England: Kogan Page.

Brown, S. (2004). Assessment for learning. *Learning and Teaching in Higher Education, 1*, 81-89.

Butler, Y. G., & Lee, J. (2010). The effect of self-assessment among young learners of English. *Language Testing, 27*, 5-31.

Chen, Y. M. (2008). Learning to self-assess oral performance in English: A longitudinal case study. *Language Teaching Research, 12*(2), 235-262.

Dickinson, L. (1987). *Self-instruction in language learning*. Cambridge, England: Cambridge University Press.

Glendinning, B., & Holmstrom, E. (2010). *English in Medicine*. Cambridge, England: Cambridge University Press.

Harris, M. (1997). Self-assessment of language learning in formal settings. *ELT Journal, 51*(1), 12-20.

Heilenman, L. K. (1990). Self-assessment of second language ability: The role of response effects. *Language Testing, 7*(2), 174-201.

Janssen-van Dieten, A. M. (1989). The development of a test of Dutch as a foreign language: The validity of self-assessment by inexperienced subjects. *Language Testing, 6*(1), 30-46.

McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language Testing, 18*, 333-349.

Munoz, A., & Alvarez, M. (2010). Washback of an oral assessment system in the EFL classroom. *Language Testing, 27*(1), 33-49.

Nicol, D., & MacFarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice.

Studies in Higher Education, 31(2), 199-218.

Nunan, D., & Carter R. (2001). *The Cambridge guide to teaching English to speakers of other languages*. Cambridge, England: Cambridge University Press.

O'Malley, J. M., & Chamot, A. U. (1990). *Learning strategies in second language acquisition*. Cambridge, England: Cambridge University Press.

Oscarson, M. (1989). Self-assessment of language proficiency: Rationale and applications. *Language Testing, 6*(1), 2-13.

Oscarson, M. (1997). Self-assessment of foreign and second language proficiency. In C. Clapham & D. Corson (Eds.), *The encyclopedia of language and education: Vol. 7* (pp.175-187). Dordrecht, The Netherlands: Kluwer.

Patri, M. (2002). The influence of peer feedback on self-and peer assessment of oral skills. *Language Testing*, *19*(2), 109-131.

Peirce, B. N., Swain, M., & Hart D. (1993). Self-assessment, French immersion and locus of control. *Applied Linguistics, 14*(1), 25-42.

Saito, H., & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research, 8*(1), 31-54.

Author's bio

Michael (Mike) Guest *is Associate Professor of English at the University of Miyazaki. He teaches medical and nursing students and is particularly interested in English for Specific Purposes and testing. He writes a regular monthly column on ELT for* The Japan News *plus a novel,* The Little Suicides*. mikeguest59@yahoo.ca*

Received: November 24, 2011

Accepted: March 14, 2014

Appendix A

First-year medical role-play test self-assessment

1 means "not at all/very poor" and 5 means "very much so/excellent"

1. My speed 1 2 3 4 5
2. My personal style/interactions 1 2 3 4 5
3. My control of grammar and basic vocabulary 1 2 3 4 5
4. My use of new or medical vocabulary 1 2 3 4 5
5. My chart was professional in style 1 2 3 4 5
6. My chart was easy to follow and understand 1 2 3 4 5
7. My questions were in a logical order 1 2 3 4 5
8. I covered all the main points 1 2 3 4 5
9. I understood the patient's problem well & asked relevant questions 1 2 3 4 5
10. I asked original & intelligent questions about the affected system 1 2 3 4 5
11. As a patient my problem made sense and was consistent 1 2 3 4 5

Total score meanings:

Under 24- You really couldn't take a basic patient history

25-28 You were OK but had many problems doing it

29-32 You were pretty good. A few problems but generally in control.

33-36 You were very good. Only a few small problems.

37-40 You were almost perfect.

What do you think your score should be? _____/40

Appendix B

Second-year medical 3-part role-play test self-assessment

1 means "not at all/very poor" and 5 means "very much so/excellent"

1. My speed 1 2 3 4 5
2. My personal style/interactions 1 2 3 4 5
3. My control of grammar and basic vocabulary 1 2 3 4 5
4. Our use of new or medical vocabulary 1 2 3 4 5
5. Our chart was complex and professional in style 1 2 3 4 5
6. Our reference letter was professional and complete 1 2 3 4 5
7. Our role-play questions were in a logical order 1 2 3 4 5
8. We covered all the main or necessary points 1 2 3 4 5
9. Our patient's problem was interesting and complex 1 2 3 4 5
10. We asked original & intelligent questions about the condition 1 2 3 4 5
11. The patient's problem made sense and was consistent 1 2 3 4 5

Total score meanings:

Under 24- Not enough value for 4 weeks of work and practice

25-28 Many content problems and/or not well-performed

29-32 A few problems but some complexity, professionalism, and ok performance

33-36 Very good. Only a few small problems.

37-40 Original, complex, professional, well-performed

What do you think your score should be? _____/40