Language Research Bulletin, 27, ICU, Tokyo.

Student Self-Assessment of High-Stakes Tests: An Exploratory Study in an English for Liberal Arts Program

Chris Hoskins
Kyoritsu University

Chris Carl Hale
Steve Engler
James Sick
English for Liberal Arts Program
International Christian University

The authors report on research they conducted in which Japanese college students shared responsibility for self-assessing their own written work on two high stakes tests. In this experiment, students served as a second raters on tests that are normally blind rated by two instructors. After engaging in an orientation and norming session identical to that which rating instructors do, students assessed their own tests using the same rubric that instructors use. Following the institutional policy requiring that blind raters must be within two points in their assessment of a piece of writing, students were also required to be within the same range of the first rating instructor in order to satisfy inter-rater reliability, otherwise the test, as is the standard procedure, went to a third rater (in this case, another instructor). While the results showed that students' tended to rate themselves too high, the level of discrepancy between teacher and student ratings was not so extreme as to preclude exploring self-assessment further. In addition, the students generally perceived self-assessment as a positive experience. It is hoped that the lessons learned through this pilot study can lead to the successful establishment of self-assessment for future writing tests.

Assessment procedures in ESL/EFL, as well as the recently emergent English for Liberal Arts (ELA) orientation to language instruction, are undergoing a paradigm shift, notably moving from a traditional teacher-centered process to more student-centered processes with greater emphasis on communication, use of integrated skills and "tests that also teach" (Richards & Renandya, 2002, pg. 335). While the involvement of students in assessment of their own work necessarily raises issues of validity, reliability and objectivity, it is distinctly in students' own interests to learn to address these issues in an effective way when reflecting on the quality of their output. These issues are not adequately addressed by excluding students from active involvement in their own assessment; moreover, the ability to self-assess is an instrumental part of an ELA approach to learning, one that asks students to engage in "a broader, more holistic, more intellectual, and more inquiry-based framework of language learning" (Wadden, Hale, Rush, Punyaratabandhu, Kleindl, Paterson & Engler, 2011, pg. 221). In this brief paper, the authors describe their attempt to implement self-assessment in ICU's ELA program and report preliminary results from their efforts.

The Assessment Process

In the assessment process normally followed for the ELA Program Wide Tests (PWT) each student essay is blind rated by two instructors using the rubric shown in Table 1.

Table 1

Quality of Understanding & Response	**Writing**	**Total**
9-10 Strong understanding of core concepts, and strong, well-developed support for your opinion. **7-8** satisfactory understanding of core concepts, with good **d**evelopment and support for your opinion. **5-6** Basic understanding of core concepts and basic development and support for your opinion. **3-4** Only partial understanding of core concepts and much more development and support needed for your opinion. **0-2** Little or no understanding of core concepts with little or no development or support for your opinion.	**4-5** Strong paragraph unity and coherence with topic sentences (including key words from prompt) and transitions. **3** Minor problems with paragraph unity or coherence such as topic sentences or transitions or minor grammar or word choice errors. **2** Some problems with paragraph unity or coherence such as topic sentences or transitions or minor grammar or word choice errors. **1** Major problems with paragraph unity and coherence lack of topic sentences or transitions and many grammar or word choice errors. **0** Little or no paragraph unity or coherence and pervasive grammar or word choice errors.	
First Rater /10	First Rater /5	
Second Rater /10	Second Rater /5	
Third Rater /10	Third Rater /5	
Final Score /20	Final Score /10	/30

The total scores of the first rater and second rater are required to be within two points of each other. When this is the case, the two scores are summed for the student's final score. If the two instructors' scores differ by three or more, a third instructor serves as an additional third rater, and any two of the three scores that fall within two points difference or less are used in the assessment. The above process is preceded by a "norming session" in which instructors meet and practice rating as many as 10 example essays. After each practice rating is done instructors compare scores and discuss differences in scores greater than two. Through negotiation, instructors adjust the scores falling outside the range of two points difference or less until all are

in agreement on how the piece of writing should be rated. Then the process is repeated for all remaining example essays. In this way, consistency is maintained among instructors doing the rating with data showing inter-rater reliability of 85% or more commonly attained. Following the norming session instructors receive the first of two packets of essays, the scores for which are recorded on a separate document to maintain the blind rating. Upon completion of the first packet of essays each instructor exchanges packets with another instructor and rates those essays, recording scores on the essay form itself. After completion of the second packet a testing coordinator enters the first scores on the test forms and determines which essays require a third rater.

In the process discussed in this paper, the students who wrote the essays served as the second rater for themselves. In the case of a discrepancy of more than two points, the students' own classroom teacher was the third rater. This self-assessment was done for two essay tests that were taken three weeks apart in the fall term with the ELA's Program C students (those who were placed in the top-most level of the program based on ITP TOEFL scores with an average of 580). The rubric for both tests was identical. Five teachers were involved. At the end of the process, the students were given a short questionnaire in order to ascertain their perceptions of self-assessment.

Results

Table 2

	Test 1	Test 2 (3 weeks later)
Number of essays rated within 2 points of agreement	51 out of 90 (57%)	53 out of 90 (59%)
Range of rating differences	0 = 9 1 = 23 2 = 19 3 = 15 4> = 24	0 = 12 1 = 21 2 = 20 3 = 9 4> = 28
Teacher average (out of 15)	10.7	10.7
Student average (out of 15)	11.7	12.0

As can be seen, a majority of students were able to rate within the 2-point allowance, with a slight improvement shown for the second test. The students tended to rate themselves higher than the teachers, as the respective averages show.

Student's reactions

Table 3

Question	Yes	No	Not Sure
Do you like doing self-assessment?	61%	35%	4%
Do you think self-assessment is fair?	72%	13%	15%
Would you like to do self-assessment again?	57%	20%	23%

As Table 3 shows, a majority of the students liked rating their own essays, and a larger majority found it to be fair. A smaller majority would like to do self-assessment again, with a large number of students undecided. As part of the student questionnaire, there was a space for them to write comments. The most common responses mentioned how it caused them to be more engaged and active throughout the whole testing process. Many students also wrote that they didn't like the idea at first, but that after going through the process they found it to be useful. On the negative side, the biggest complaint was that they thought they were not qualified and that assessing was the teacher's job.

Discussion

The teachers and students involved in this self-assessment procedure ended up using a few different approaches when it came to the student norming sessions. In class sections in which the most extensive preparation for self-assessment was done, students were led through a norming session in which they essentially followed the same process used by faculty in preparation for doing the first round of scoring of the essays. First, the process followed by instructors was explained to the class members. Salient points were 1) to read the rubric on the test forms carefully and discuss any questions participants may have about how to interpret the rubric, 2) to refer back to the rubric when assessing the writing, 3) to understand how to negotiate among group members to reach consensus in which total scores given by each of the members of the norming group were adjusted to fall within a range of two points difference or less, and 4) that it is the responsibility of the group to make sure that all group members were given equal opportunity and encouragement to explain their assessment and the reasoning behind it.

Following discussion of the rubric and the process to be followed in norming, students were put into small groups and each provided with copies of two example essays that had already been graded by an instructor in the first round of assessment. The example essays selected for this part of the process were chosen because they were judged to be well-written essays. Neither the names of the students that wrote the essays nor scores for the essays given by instructors in the first round of assessment were made known to the students. Sitting in groups, students read one of the essays and scored it according to their interpretation of the rubric without discussion. When all students in a group had finished scoring by themselves, scores were compared and discussion and negotiation ensued until group members could agree on adjustments that brought all scores given to within a range of difference of two points or less. This process was repeated with the second of the two essay copies, after which the groups were dissolved and students individually assessed their own essay using the same process and rubric as in the norming session. Other sections spent one class period, or part of a class period doing norming as outlined above, while one teacher explained how students' expressions of dislike for self-assessment led to a class discussion that resulted in the instructor encouraging students to just rate themselves highly.

Conclusion

The inclusion of student self-assessment with instructor assessment in the Program Wide Test (PWT) in some class sections in the ELA exemplifies the paradigm shift mentioned above. However, as Wilde, Del Vacchio, and Gustke (in press) stipulate, several conditions exist for reliability in assessment, including "[use of] trained judges, working with clear criteria, from specific anchor papers or performance behaviors," as well as the ability to "ensure that raters use criteria and standards in a consistent manner" (cited in Richards & Renandya, 2002, p. 340). Keeping this in mind, the self-assessment described in this paper has to be regarded as a work in progress due to inconsistency in the way in which individual instructors prepared students to do the self-assessment. While preliminary data tend to validate students' involvement from a developmental standpoint (Hale, Sick, Engler & Hoskins, 2012), literally creating a reality in which the PWTs are "tests that also teach," the preparation for the self-assessment exercise has important implications for the degree to which student self-rating can reach a higher level of consistency that is closer to the rating done by instructors. In this study, the gap between inter-rater reliability for students and inter-rater reliability for instructors was about 22%; that is, 57% to 59% of the student-graded essays did not require a third reader compared to 80% or higher for instructors. However, it is important to remember that instructor training sessions for test grading include as many as ten sample essays compared with only two for students; and that most instructors have undergone the training repeatedly over a number of years. One would therefore expect instructors to have higher inter-rater reliability than students and one could also expect students to improve in inter-rater reliability over the course of a year as they experience more training sessions. Other factors that could help close the gap would be if the student training process could include three or more examples rather than the two that instructors typically use (for instance, an example of a poorly written essay to give students a broader view of the range of writing skill within one class section as well as the opportunity to attempt to apply the rubric to that broader range). In addition, defining a greater-than-2-point difference on a 15-point scale as a discrepancy is quite strict. Using a greater-than-3-point difference as the criteria for a third rater might also be considered reasonable and would have resulted in a much better inter-rater reliability between students and expert raters. Finally, it should be noted that the 22% gap between instructor and student inter-reliability may itself be inflated due to the uncooperative teacher who urged the students just to rate themselves highly.

Of course, these considerations assume that rater-reliability is the primary goal of the activity. While we believe that students will improve rater-reliability the more they do it, we would like to posit that, in fact, the primary goal of implementing self-assessment has very little to do with statistical variance and inter-rater reliability. Students are as integral to the academic community as the teachers who assess them, and there is no greater way to ratify students in a liberal arts community than by trusting them to participate in one of the most salient and lasting referents of academic life: grades. Finally, perhaps if self-assessment is further implemented in a consistent manner at ICU and in other liberally oriented English-language programs, more students will see the value in this process in the same way the following student did, as evinced in this response to the open-ended comments section on the questionnaire:

> The Self-Assessment task made sense after we re-acknowledged our responsibilities in this university academic community. I felt trusted by the instructors, asking us to grade our own paper. Checking my own essay made

me see that the learning process does not stop when we finish our program-wide tests and turn them in. Assessing ourselves is an important task that lets us examine our writing and thinking objectively, a skill that surely will be useful when we write other essays and papers in the future.

References

Hale, C., Sick, J., Engler, S., Hoskins, C. (2012, July). *Self-assessment as an integral part of academic writing instruction.* Paper presented at the second International Conference of the Israel Forum for Academic Writing, Tel Aviv, Israel.

Richards, J. C., & Renandya, W. A. (2002). *Methodology in Language Teaching: An Anthology of Current Practice.* Cambridge: Cambridge UP.

Wadden, P., Hale, C., Rush, E., Punyaratabandhu, D., Kleindl, M., Paterson, R., & Engler, S. (2012). *English for liberal arts: Toward a new paradigm for university language teaching.* In A. Stewart & N. Sonda (Eds.), *JALT 2011 Conference Proceedings,* Tokyo: JALT.

Wilde, J., Del Vecchio, A., Guske, C. (in press). Alternative assessments for Latino students. In M. Gonzales, A. Huerta-Macias, & J. Tinajero (Eds.) *The schooling of Latino students: A guide to quality practice.* Lancaster, PA: Technomic Publishing.