Self Assessment of Oral Performance in the EFL University Classroom
Yuh-Mei Chen
National Chung Cheng University
folcym@ccu.edu.tw

Abstract
This paper reports a study which investigated the effect of self-assessment in the EFL university speaking classroom in terms of its validity and students’ perceptions. Participants were a class of 28 students in a freshman oral training class. Data were collected from an evaluation form, on which marks and written statements were demanded for assessment, and a questionnaire, which elicits students’ opinions about participating in assessment and its impact on learning. The study found that students tended to under-mark themselves, but their marks became more comparable to peer and teacher marks as they gained experience in the assessment process. Their written comments reflected the agreed criteria. Progress in language and delivery was noted in the second self assessment. Students claimed that they felt comfortable about self assessment and performed the task honestly and fairly, and that the assessment task was hard but good, making them independent, think more and learn more. Self assessment was overall considered beneficial for learning and feedback critical and valuable for improvement. Pedagogical implications are provided in the end.

Keywords: Self Assessment, English Classroom Assessment

1. Introduction
In line with theories of self-directed learning and learner autonomy, self assessment is assuming a larger role in language teaching nowadays. The procedure involves students in making judgments about their own learning, particularly about their achievements and learning outcomes. Many researchers and practitioners deem self-assessment as a vital part of learner autonomy and argue that teachers should provide the opportunity for students to assess their language level so as to help them focus on their own learning (Blanche, 1988; Blue, 1994; Dickinson, 1987; Harris, 1997; Henner-Stanchina & Holec, 1985; Oskarsson, 1989). Hunt, Gow & Barnes (1989) even claim that without learner self-evaluation and self-assessment “there can be no real autonomy” (p. 207). Oskarsson (1989) mentions six advantages of using self-assessment in the language classroom: promotion of learning, raised level of awareness, improved goal-orientation, expansion of range of assessment, shared assessment burden, and beneficial post-course effects. Blue (1994) identifies benefits such as encouraging more efforts, boosting self-confidence, and facilitating awareness of the distinction between competence and performance as well as self-consciousness.
of learning strengths and weaknesses.

In addition, self-assessment is considered necessary for effective lifelong learning (Boud, 1995, 2000). There is a common understanding that university education should equip students with the skills and attitudes required throughout their lives. As Boud (2000) argues, “Assessment involves identifying appropriate standards and criteria and making judgments about quality. This is as necessary to lifelong learning as it is to any formal educational experience” (p. 151). Given as such, when asked to assess their own language performance and progress in the classroom, students will identify appropriate standards for the task in hand and seek forms of feedback from the environment (including peers, teachers, written or oral sources), and gradually develop a critical attitude toward their learning, which will in the long run prepare them for their future roles in the workplace.

Research into the reliability and validity of self-assessment among ESL/EFL students has yielded mixed results. Some studies have reported agreement between students’ self-awarded ratings and ratings awarded by their teacher (AlFallay, 2004; Chen, 2005) or scores that students expected to get in a test and those they actually obtained (Bachman and Palmaer, 1989; LeBlanc, 1985) and suggested that students are able to assess their language proficiency accurately. However, discrepancy has also been found between students’ self-ratings and ratings from other sources (Blue, 1988, 1994; Patri, 2002; Wangsotorn, 1981; Yang, 2002). Despite the contradictory findings, literature in this area generally holds the proposition that practice, support, and experience are key factors to increase the accuracy of self-assessment (AlFallay, 2004; Chen, 2005; Orsmond, Merry, & Reiling, 2000; Tara, 2001).

In Taiwan, the use of self-assessment to promote learner autonomy is not novel in the EFL classroom (e.g. Chang, 1999; Chen, 1999, 2000; Chuang, 1997; Fang, 2005). The procedure has been mostly reported as a course-end activity or assignment which demands students to reflect on their learning process or achievement in reading or writing. Self-assessment is mainly resource driven, not a move to share the power inside the classroom. To date we have known little about whether self-assessment is a reliable and valid method of EFL classroom assessment. Chen (2005) observed a significant match between students’ self-assessment and teacher assessment. Yang (2002) reported very low and even negative correlations between students’ self-ratings of English oral proficiency and teachers’ markings in three classes. He attributed the surprising results to students’ lack of rating experience and self-consciousness of their lower proficiency in comparison with peers’ classroom performance.

Literature suggests that student participation in grading is a viable method to encourage learner autonomy because the ability to assess one’s proficiency with a reasonable degree of accuracy and supplement any deficiency therein is essential.
preparation for lifelong learning. The more often students are exposed to self-assessment the more accurate their assessment becomes (Boyd & Cowan, 1985). Given the dearth of research and classroom practice in self-assessment on oral performance, a refined design of assessment procedures with supportive feedback to student assessors and its evaluation are needed.

2. Methodology

The purpose of this study was to investigate the validity of using self-assessment to measure students’ oral performance in an EFL university classroom. The study was conducted within a two-credit course of English Oral Training at a national university in southern Taiwan. The class, meeting 100 minutes each week for a semester, aimed to help students express their ideas in clear and appropriate English, think critically and reflectively, assume their learning responsibilities, collaborate with their peers, and evaluate their performance. Assessment was integrated into learning and teaching in this course. In addition to the speaking tasks such as individual speech, pair talk, and group project, students were demanded to assess their own and peers’ oral performance as well as mid-term and course-end evaluation of participation in the pair talk and group project. The assessment task investigated in the study was only regarding individual oral performance.

Participants were 28 students, with 18 females and 9 males, 22 English majors and 6 non-English majors. Before the study, they had studied English for 6 to 12 years. When asked to compare their English ability with their classmates on a 5-point scale, on average they rated their English listening ability at 3.14, speaking ability at 2.75, and learning attitude at 3.52. Out of all participants, 7 had experience in assessing and making comments on self oral performance; only 2 graded peers’ performance in other classes before.

Instruments used in the study were an evaluation form and a questionnaire. The assessment components were developed by the teacher and students collaboratively through processes of group discussion, class sharing, and teacher summarizing. The assessment criteria included four elements—content (30%), language (30%), delivery (30%), and manner (10%). Also, five-leveled scoring was decided before assessment procedures: excellent, 90% and above, good, 80%-89%, fair, 70%-79%, ok, 60%-69%, and poor, below 60%. The questionnaire was designed by the teacher researcher to elicit students’ opinions of the assessment task. It was administered to students at the end of the assessment task.

Boud (1995) alerts teachers to some potential problems of the practice: students believing they are doing the teacher’s job, self-marking without engagement with criteria, and use of self-assessment activities in isolation. To tackle these problems,
the assessment procedure in the present study embraced three stages: training, observing performances, and sharing comments and suggestions. The training was to facilitate students’ understanding of assessment criteria and standards. At the first class meeting, students were divided into groups and discussed about what to be noted in oral performance. With the teacher’s assistance, the class concluded a set of criteria for good performances and scoring standards. In the second meeting, students first practiced assessing two videotaped performances against the agreed criteria with their group members and shared evaluation results in class. The teacher then commented on the videos to demonstrate her evaluation and scoring of the performances and explain to students how well their evaluations had been made.

After the training, in a cycle of 5 weeks students took turns speaking on self-chosen topics about 3-5 minutes. Using the same evaluation form, both the teacher and peers assessed simultaneously during the performances but students completed self-assessment later. Subsequent to observing performances, students first exchanged comments or suggestions, reflections and observations within groups; then the group speakers and the teacher gave feedback to the performers. This assessment task was undertaken in 10 weeks, that is, each student had two turns talking about something he/she brought to class and practiced self-assessment twice, peer assessment nine times.

Different from some procedures reported in the literature, self-assessment conducted in the present study was asking students to assess their own performance in the classroom setting with the same criteria as peer and teacher assessment. It was not a task for students to rate their language abilities and/or predict their test scores in the questionnaire format as (Bachman & Palmer, 1989; Blue, 1994; LeBlanc & Painchaud, 1985). Nor was it like learning logs practiced as tools for facilitating reflective and critical analysis of learning progress (Baldwin, 2000; Boud, 1995; Schon, 1987). Self-assessment was practiced as a tool to monitor self-performance through critical feedback from peers and the teacher in the classroom. It was not “an isolated or individualistic activity” (Boud, 1999, p. 122) but conducted with peer and teacher assessment. It also became an iterative process when students subconsciously compared peers with themselves in peer assessment.

3. Results and Discussion

3.1 Assessment Statistics

Table 1 summarizes the descriptive statistics of self-, peer, and teacher assessment on oral performance in the first and second cycles. A look at the means of the three different forms of assessment showed that self-awarded marks tended to be lower than those given by peers and the teacher. In the first cycle of assessment,
students under-marked themselves so much that their mean score deviated from the teacher’s mean score at about 5. Another look at the standard deviations observed that scores from self-assessment and teacher assessment had a similar range, but extended wider than those of peer assessment. This suggested that the former two sets of scores spread over an identical range and were internally more discriminating than the latter set of scores. However, the differences among the three forms of assessment decreased in the second cycle. Means of the three sets of scores became quite alike; the differences all fell within one averaged standard deviation. The degree of leniency of self- and peer assessment in the second cycle seemed to bulk large; the mean score increased and the range of standard deviation became narrower than the first cycle. Initial readings of the assessment statistics indicated that mean scores of self-, peer and teacher assessment tended to resemble when students had more practice and gained more experience in assessment.

Table 1 Descriptive statistics of oral assessment in the first and second cycles

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA</td>
<td>28</td>
<td>78.17857</td>
<td>5.436867</td>
<td>21</td>
<td>67</td>
<td>88</td>
</tr>
<tr>
<td>PA</td>
<td>28</td>
<td>81.82143</td>
<td>4.538343</td>
<td>19</td>
<td>75</td>
<td>94</td>
</tr>
<tr>
<td>TA</td>
<td>28</td>
<td>83.07143</td>
<td>5.537271</td>
<td>26</td>
<td>69</td>
<td>95</td>
</tr>
<tr>
<td>2nd</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA</td>
<td>28</td>
<td>83.07143</td>
<td>4.791416</td>
<td>24</td>
<td>67</td>
<td>91</td>
</tr>
<tr>
<td>PA</td>
<td>28</td>
<td>85.82143</td>
<td>3.333929</td>
<td>16</td>
<td>78</td>
<td>94</td>
</tr>
<tr>
<td>TA</td>
<td>28</td>
<td>84.46429</td>
<td>5.587567</td>
<td>26</td>
<td>70</td>
<td>96</td>
</tr>
</tbody>
</table>

AS= self-assessment, PA= peer assessment, TA= teacher assessment

Further statistical tests were performed to make inferential analysis about the scorings derived from different forms of assessment. The ANOVA analysis of the first cycle assessment detected significant difference among the means of self-, peer, and teacher assessment (F= 6.716636, df= 2, p<.01), but no difference was found in ANOVA of the second cycle assessment (F= 2.432418, df= 2, p= 0.094). After significance of the first assessment cycle found by ANOVA, a Tukey HSD test was performed to do a pairwise comparison of the three scoring means. The test result showed that scorings of self- vs. peer assessment (p<.05) as well as those of self- vs. teacher assessment (p<.01) were different. In the second cycle of assessment, t-tests observed that self-assessment differed from peer assessment at p.<05, but appeared comparable to teacher assessment.

Since teachers are always regarded as experienced and more capable assessors,
teacher assessment has been taken as the criterion in the literature of student assessment consistency. This study also adopted this approach to examine the validity of self-assessment. Before doing so, the reliability of both teacher assessment and self-assessment was scrutinized by correlation analysis of their markings given within and without the class. The assessment of the second cycle was then chosen for reliability analysis because it was hypothesized that the assessors were more experienced and accurate after the practice of the first cycle. All the performances in the second cycle were videotaped and assessed again by the teacher and students outside the class at the end of the cycle. The correlation coefficients of two teacher markings and two student self-markings yielded .78 and .73 respectively, indicating significant reliability of both forms of assessment (n= 28, p<.0001).

When the reliability of self- and teacher assessment was ensured, the validity of student self-assessment was examined by comparing students’ self-marking with teacher marking. As was said above, the Tukey test and t-test results revealed that difference between self assessment and teacher assessment was detected in the first cycle, but not in the second. To see whether students made significant progress in assessing themselves in terms of how much their self-marking deviated from teacher marking, another t-test of student-teacher assessment mean difference was also conducted. As shown in Table 2, significantly students’ self-awarded scores differed less from those given by the teacher in the second cycle of assessment. In sum, inferential statistics indicated that student self-marking was consistent and there was a significant match between self-assessment and teacher assessment in the second cycle of assessment. Students’ experience obtained from the first cycle of assessment might contribute to their improvement of assessment accuracy.

Table 2 T-test result of student-teacher assessment difference between the first and second cycle

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
<th>t</th>
<th>p (one-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-T Diff. (1st)</td>
<td>28</td>
<td>-4.89286</td>
<td>5.34</td>
<td>-2.6884</td>
<td>0.00476**</td>
</tr>
<tr>
<td>S-T Diff. (2nd)</td>
<td>28</td>
<td>-1.39286</td>
<td>4.36</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.2 Written Comments

Table 3 summarizes the frequencies of student and teacher written comments on the evaluation in terms of the topic and nature. First, let’s look at self-assessment in both cycles of assessment. It showed that students paid attention to manner and delivery most. A majority identified showing nervousness and giving little eye contact as serious flaws in their performances. In addition to the agreed criteria—content, language, delivery, and manner, other variables were also mentioned in their
comments. Some emphasized their effort, audience amusement, and self-confident, improvement as factors to define success/failure of their oral performance. However, the teacher gave comments on the areas of content, language, and delivery more. She noted students’ language and delivery problems, explained her likes or suggestions for improvement.

In total, students stated fewer written comments in the second cycle of assessment than in the first cycle. An examination of the nature of their comments may explain the divergence. Based on Falchikov’s (1996) classification, in this study positive comments refer to those identifying strengths, negative comments identifying weaknesses, and neutral comments revealing reflection and suggestions for improvement. A close look at frequencies of these three types of comments found that the amount of negative comments was reduced from 40 to 13, and those of positive and neutral comments increased. This might suggest that students were extremely strict when rating their own performances in the first cycle, and the finding supported the result of scoring statistics that students tended to under-mark themselves a lot. After long periods of practice, the students not only grew to notice their strengths and weaknesses, but also learned to analyze their deficiencies and see how to improve them. Their growth may result from the teacher’s guidance and encouragement. Compared with students, the teacher gave much more positive feedback in her evaluation and provided to students the same amount of negative and neutral comments.

Students learn through interaction with the environment. In this oral training class, students were demanded to assess their own and peers’ oral performance for 10 weeks. After each performance, the class shared observations and evaluative feedback to the performers through teacher supervision. Giving and receiving feedback was an essential stage of the assessment procedure. Students’ written comments demonstrated that they were more likely to reflect on their problems and identify areas for improvement in the assessment by the course of time. A close reading of their written remarks also noted the change from a negative tone of “bad, bad, bad performance” to a reflective or suggestive one of “if….it would be better”, the structure used by the teacher when giving suggestions. This finding echoes the conclusions of AlFallay (2004) and Chen (2005) that sufficient practice and critical feedback have a positive effect on the accuracy of assessment.

<table>
<thead>
<tr>
<th>Topic of Comments</th>
<th>1st SA</th>
<th>1st TA</th>
<th>2nd SA</th>
<th>2nd TA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>9</td>
<td>25</td>
<td>8</td>
<td>19</td>
</tr>
</tbody>
</table>

Table 3 Comparison of written comments from self- and teacher assessment in the first and second cycles
<table>
<thead>
<tr>
<th>Language</th>
<th>8</th>
<th>18</th>
<th>4</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delivery</td>
<td>13</td>
<td>20</td>
<td>12</td>
<td>26</td>
</tr>
<tr>
<td>Manner</td>
<td>23</td>
<td>4</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>Others</td>
<td>7</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effort/Preparation</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audience Amusement Response</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lack of Confidence</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Improvement</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>60</td>
<td>67</td>
<td>42</td>
<td>61</td>
</tr>
</tbody>
</table>

**Nature of Comments**

<table>
<thead>
<tr>
<th>Positive</th>
<th>11</th>
<th>45</th>
<th>13</th>
<th>38</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>40</td>
<td>12</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>Neutral</td>
<td>9</td>
<td>10</td>
<td>16</td>
<td>12</td>
</tr>
</tbody>
</table>

### 3.3 Students’ Perceptions of Self-Assessment and Peer Feedback

Student’s t-tests were conducted to examine students’ questionnaire responses on the 5-point scale with a hypothetical population mean of 3. The results showed that significantly students supported participation in self-assessment ($m=4.25$, $t=9.47$, $p<.0001$), felt comfortable about assessing their own oral performance ($m=3.5$, $t=3.33$, $p=0.002503$), and assessed themselves honestly and fairly ($m=3.86$, $t=7.653$, $p<.0001$). With $p$ value smaller than .05, they thought self assessment was fair ($m=3.43$), accurate ($m=3.39$), but hard ($m=3.57$). The task made them think more ($m=4.10$) and learn more ($m=3.61$). Self-assessment was undertaken with peer and teacher assessment. Students also considered evaluative feedback given after their performance ($m=3.51$) was helpful and constructive. Overall, self assessment was regarded as beneficial for learning and peer feedback as critical and valuable for improvement. The results confirmed findings of previous studies (e.g. Chen, 2005; Orsmond et al., 2000; William, 1992) that students enjoyed and perceived benefits in doing self-assessment.

### 4. Conclusion

This study was to investigate whether self assessment was a valid and reliable tool for the EFL oral training classroom. It found that significantly students could assess themselves in a manner comparable to the teacher when they had more practice in assessment and they favored the assessment task and reaped benefits from it. In terms of classroom practice and future research, this study has the following implications.

First, statistical analysis of the study demonstrated that self-assessment is a
reliable and valid tool, if provided with sufficient training and support. It provides students opportunities not only to examine closely their achievement of the desired linguistic and assessment skills or knowledge but also to use the target language in authentic context. The assessment procedure used in the study indeed involved two processes: learning to assess and assessing to learn. The evidence presented in this study and a previous study (Chen, 2005) demonstrated that for university freshman and sophomore students, self-assessment is a viable alternative of English classroom assessment for oral performance. Interested teachers may as well use self-feedback marking as a tool to encourage students to take greater responsibility for their own learning and engender deeper learning.

Second, students’ favorable attitudes toward self-assessment in the study suggested the necessity of the two-week training and evaluative feedback. Assessment is a critical activity. Training is a must for students who are accustomed to the centralized education system and tend to accept the so-called “standard” or authoritative evaluations. In addition, students need to learn to assess and assess to learn through receiving and giving feedback. Self-assessment is best practiced with peer and teacher assessment. Interacting with the environment, students can make judgments about their skills or performance more accurate in relation with others. This self-assessment process is to seek knowledge from the surroundings to make sense of uncertainty, conflict, and doubt, and to draw a clearer self-profile. So, training or orientation of assessment criteria is first and foremost to facilitate the assessment procedure; feedback session is indispensable to monitor assessment accuracy.

Third, students’ personal traits or psychological characteristics may govern students’ assessment behavior. Students in this study were likely to under-mark themselves and give emphasis to affective factors such as confidence, nervousness, and audience feedback in oral assessment. Many have claimed that the degree of leniency increased as the subjects’ level of expertise decreased. For example, Boud and Falchikov (1989) in their extensive review of educational self-assessment studies noted that more able and experienced students tended to under-mark themselves in comparison with teacher marking. Blue (1994) and Orsmond et al., 1997) also reported that lower level of students tended to over-mark while better students tended to err on the side of under estimation. In addition to experience and proficiency, Blue (1994) speculated that nationality can be an important factor in self-assessment, proposing that some nationalities having a tendency to overestimate their level and others tending towards underestimation. AlFallay (2004) looked at the role of some psychological and personality traits and observed that those possessing the positive side of a trait are more accurate than those who have its negative side. To understand
the relation between Chinese students’ personality traits and their self-assessment accuracy, correlation analysis and in-depth interviews are needed.

References


