
LU Jun & FAN Jinsong
Shanghai Jiao Tong University

Abstract

Low rating reliability has long been the primary concern in school-based oral English achievement tests. In this study, a computer-aided rating system (CARS) was developed to improve inter- and intra-rater reliability through the instantiation of rating criteria, task division and random distribution, on-line training, reliability verification and sound wave “reading” and “writing.” A rating experiment was conducted among six raters to compare intra- and inter-rater reliability between traditional rating and rating with CARS. At the end of each round of rating, a conference was held. Both quantitative and qualitative analyses show that CARS can significantly improve inter- and intra-rater reliability, mainly through helping raters use criteria more accurately and focus more attention on rating. In addition, the research has also shed light upon further study on improving rating reliability.

Key words: school-based oral English achievement test; rating; reliability

1. Introduction

The College English Curriculum Requirement (CECR) issued by the Ministry of Education stresses that “the objective of College English teaching in China is to develop students’ ability to use English in a well-rounded way, especially in listening and speaking” (CECR, 2007: 18). CECR sets three levels of requirements for undergraduate English teaching and stipulates that the evaluation at each level should be focused on “the assessment of students’ ability to use English in communication, particularly their ability
Research on Improving Rating Reliability for School-based Oral English Achievement Tests

to listen and speak in English” (CECR, 2007: 27).

An achievement test, normally administered at the end of an instruction unit or a group of instructional units, is intended to measure what a student has learned within a period or up to a given time, based on the content of the course (See, e.g., Bachman, 1990: 60; Davies, Brown, Elder, Hill, Lumley & McNamara, 1999: 2; Heaton, 1988: 172). If properly designed and administered, an achievement test can provide useful feedback about teaching and learning, thus forming part of the formative assessment in the classroom or school context and, at the same time, exert positive washback effects (Heaton, 1988: 172).

2. Literature Review and the Significance of the Present Study

Speaking tests usually falls into two types, OPI (Oral Proficiency Interview) and SOPI (Simulated Oral Proficiency Interview). The former is normally a face-to-face interview, conducted and rated by trained interviewers, while the latter elicits and records oral discourse from test-takers through the use of recorded and visual task stimuli. Many researchers have set out to compare the two types of speaking tests. For example, Stansfield (1991: 199-209) carried out a correlation analysis between two types of rating and found that the scores were highly correlated, with coefficients ranging between 0.89 and 0.95. In 1992, Stansfield and Kenyon (1992: 347-64) conducted a rating error analysis and found that there was no significant difference between two types of speaking tests. However, as warned by Shohamy (1982: 161-71, 1994: 99-123), correlations per se cannot provide sufficient evidence that the two tests measure the same language abilities. Rather, there is a need to examine the two tests from additional perspectives. In the study conducted by Shohamy (1994: 99-123), the two types of speaking tests, OPI and SOPI, were investigated from a number of perspectives, adopting Weir’s (1988: 15-25) model of a priori and a posteriori validation. The study brought the researcher to the conclusion that there was a need to examine validity from multiple perspectives.

In China, Jin and Guo (2002: 72-79) conducted a study on 75 subjects, who took both CET-SOPI (College English Test—Simulated Oral Proficiency Interview) and CET-SET (College English Test—Spoken English Test) during almost the same period of time. The result showed that the scores were highly correlated, the content validity was consistent, and the test-takers’ language output was almost the same.

With the advances of information technology, computers have been extensively used in testing language abilities. Computer-aided SOPI, as a result, is becoming increasingly popular in testing speaking ability due to its conspicuous advantages such as standardized format, high rating reliability as well as the convenience and low cost of administration. Cai’s (2005: 66-75) research further proved that, in addition to high correlation and consistency, computer-aided SOPI had more advantages over OPI in terms of rating reliability, validity and practicality. The review of the relevant studies leads to the conclusion that computer-aided SOPI could be a desirable substitute for OPI or SOPI, especially in the Chinese context, where there are normally a huge, unwieldy number of test-takers.
In most Chinese universities, there are usually a large number of students choosing oral English courses. As a result, an increasing number of colleges and universities choose computer-aided SOPI for oral English achievement tests in order to ensure higher consistency and efficiency. In current school-based SOPIs, a common practice is that dozens of teachers, sometimes more than one hundred, independently rate and report their own students’ taped performances. However, as students’ language ability varies from class to class and teachers also differ considerably in rating, low reliability has invariably been a serious concern for school-based assessment of oral English.

Current research in assessing speaking mostly centers around large-scale standardized proficiency tests. With the increase of learning-related and informal assessments, more studies need to be conducted into the tasks and testing processes relevant to that context (Luoma, 2004: 190). Rating in school-based achievement tests, for example, deserve particular attention since large-scale standardized speaking tests can afford to hire and train a pool of raters and provide well-written rating criteria while school-based course exams cannot. By comparison, achievement tests are more important in terms of evaluating students’ speaking abilities because they are more closely linked to teaching syllabus and because they can provide more useful feedback about teaching and learning. Obviously, if the rating reliability is not high enough, the assessment results might be misleading. Therefore, it is highly necessary to conduct research to ensure high rating reliability for school-based oral English achievement tests. In recent years, a growing number of oral English testing packages have been developed. However, so far, little research has been conducted to improve rating reliability through development of a computer-aided rating system, especially for school-based oral English achievement test.

Since the issue of CECR in 2004, teaching speaking for non-English majors has been increasingly becoming an independent course in universities in China. Nevertheless, due to the relatively large number of test-takers and lack of professional grading training, there arises a serious problem, namely the low rating reliability of school-based oral English achievement tests. Therefore, how to improve the rating reliability will be of great significance for oral English teaching and research for non-English majors.

3. Research Methodology

Rating is an interaction between raters, criteria and performances (McNamara, 1995: 173). Needless to say, the factors affecting reliability, for example, raters’ subjectivity, cannot be well filtered out if the rating process is not properly controlled and supervised. This is particularly true of school-based achievement tests, since most raters are college English teachers and do not usually receive special training before they start rating.

3.1 Design of a Computer-aided Rating System

According to Luoma (2004: 179), two types of rating reliability are particularly relevant to speaking assessment. The first type is intra-rater reliability or internal consistency, referring to the degree to which raters agree with themselves over a period of time on
the ratings that they give. The second type is inter-rater reliability, meaning the degree of agreement among raters. In order to improve the two types of reliability in school-based oral-English achievement tests, a five-part computer-aided rating system (CARS) was developed (see Figure 1).

**Part I. Tape Treatment.** This part has two functions: task division and random distribution. The treatment is to divide each test-taker’s performance according to task types and then distribute them randomly so that raters can have equal access to performance of different levels and will only have to rate one task.

**Part II. Rater Training.** The training part includes three steps: 1) instantiation of rating criteria, usually done by a small group of master raters. Benchmark tapes for each level will be produced; 2) collective training, involving masters’ explanation and demonstration of instantiated criteria, and raters’ trial rating; 3) on-line training and qualification verification. Raters must pass qualification verification before they start to rate the oral tests.

**Part III. On-line Scoring.** Cool Edit Pro 2.0 is employed for sound playing so that sound waves are visible and also available for editing. Raters are encouraged to “catch” typical errors in test-takers’ utterance for feedback. Skipping silent parts in test-takers’ tape is also allowed. Score recording is done by clicking corresponding buttons.

![Figure 1. Flow chart for CARS](image-url)
Part IV. Reliability Verification. This part checks inter- and intra-rater reliability by respectively comparing raters’ scoring with benchmarks and with their own rating repetitions over different periods of time. If the reliability is too low, further on-line training is required until they pass qualification verification again.

Part V. Data Uploading & Output. This part helps raters upload relevant data and automatically carry out such tasks as score calculation, statistic analysis and result reporting.

3.2 Research Questions
This study was designed to verify whether the design of CARS can significantly improve the rating reliability in school-based oral English achievement tests. Therefore, it was intended to address the following research questions:

1. Can CARS significantly improve inter-rater reliability? If so, how?
2. Can CARS significantly improve intra-rater reliability? If so, how?
3. Why can CARS help improve rating reliability?

3.3 Research Design
Sixty samples were chosen from a school-based oral English achievement test held in December 2008 for criterion instantiation and corresponding benchmark tapes. The rating criteria and corresponding instantiation were created by three master raters. Six English teachers carried out two rounds of rating with an interval of one week. In order to minimize possible influence or bias caused by the raters’ familiarity with test-takers, all the names of sound files were coded in numbers. In both rounds, in order to avoid the possible influence caused by different tapes on rating, the same 30 samples, together with 10 repetitions, were used for rating experiment to check inter- and intra-rater reliability. The repeated files were given different filenames so that the raters would not know which tapes were rated twice. The repetitions used in the 2nd round were different from the ones in the 1st round so as to reduce the influence caused by the raters’ possible impression of the previous rating. A one-week interval was intended to minimize the influence caused by any possible impression of the first round of rating.

In the 1st round, the raters received a one-hour collective training according to the above design, and then scored 40 taped performances (including 10 repetitions) independently by using RealPlayer. Each taped performance consisted of three tasks so that there were altogether 120 tasks for each rater. The three tasks in each file were scored successively and their scores were recorded manually in corresponding tables. This is referred to as traditional rating in the present study.

One week later, the 2nd round of rating was conducted with the same rating criteria. The same six raters were randomly divided into three pairs and each pair rated one task through on-line rating with CARS. This time, in addition to the 30 samples used in the 1st round and 10 repetitions, an additional 80 samples were used. Different from the first round of rating, each pair of raters rated only one task but had to deal with a total of 120 tasks independently. In this way, the rating workloads for two rounds were equivalent. The scores were saved in a corresponding database. At the end of each round, a brief conference was held among the six raters and notes were taken for qualitative data-collection.
3.4 Data Collection and Analysis

The data used for this study consisted of two parts, the quantitative and the qualitative parts. The quantitative part included the scores for the three tasks of the samples shared in two rounds of rating. For both rounds, the scores given by each pair of the raters were compared respectively for inter-rater reliability checking and the repetition rating scores were compared for intra-rater reliability checking. That is, for each task in both rounds, 40 pairs of data were collected for inter-rater reliability analysis and 20 pairs for intra-rater reliability analysis. SPSS 13.0 was employed to conduct correlation and cross-tab analyses for the aforementioned comparisons.

The most common way of expressing reliability is through correlation, which is a statistical indicator for the strength of relationship between variables (Luoma, 2004: 182). The variables in this study were pairs of scores provided by different raters and also the scores by the same raters over two different periods of time. Cohen’s kappa coefficient was used to measure the agreement between the evaluations of two raters when both were rating the same object. Cohen’s kappa coefficient (k), a statistical measure of inter-rater agreement for qualitative (categorical) items, is generally thought to be a more robust measure than simple percentage agreement calculation since it takes into account the agreement occurring by chance (Bachman, 2004: 201-02). The values of k vary from 0 to 1, indicating different degrees of agreement, and range from “no agreement” to “almost perfect agreement”, as shown in Table 1 (Also see http://en.wikipedia.org/wiki/Cohen’s_kappa).

<table>
<thead>
<tr>
<th>K</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0</td>
<td>No agreement</td>
</tr>
<tr>
<td>0.0 — 0.20</td>
<td>Slight agreement</td>
</tr>
<tr>
<td>0.21 — 0.40</td>
<td>Fair agreement</td>
</tr>
<tr>
<td>0.41 — 0.60</td>
<td>Moderate agreement</td>
</tr>
<tr>
<td>0.61 — 0.80</td>
<td>Substantial agreement</td>
</tr>
<tr>
<td>0.81 — 1.00</td>
<td>Almost perfect agreement</td>
</tr>
</tbody>
</table>

The qualitative part of data included the raters’ reflections on the two rounds of rating and their suggestions on further upgrading the design of CARS for higher rating reliability. The corresponding results were reported and discussed as follows.

4. Results and Discussion

4.1 Quantitative Analysis

4.1.1 Inter-rater Reliability Checking

Correlation coefficient (r) and Cohen’s kappa coefficient (k) were calculated between the pair of scores for the same tasks for inter-rater consistency. In the present study, a 0-5
point scale was employed to grade students’ oral performance for each task, so the scores for each task assumed the value of one of the integers from 0 to 5 and had much of the property of categorical items. Therefore, Cohen’s *kappa* coefficients (*k*) for each pair of raters were also calculated for inter-rater reliability. The results were reported in Table 2.

**Table 2. Inter-rater Reliability Report**

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st Round</td>
<td>.800**</td>
<td>.703**</td>
<td>.814**</td>
</tr>
<tr>
<td></td>
<td>.260**</td>
<td>.186*</td>
<td>.119</td>
</tr>
<tr>
<td>2nd Round</td>
<td>.863**</td>
<td>.858**</td>
<td>.881**</td>
</tr>
<tr>
<td></td>
<td>.463**</td>
<td>.638**</td>
<td>.516**</td>
</tr>
</tbody>
</table>

** means *p < .01; * means *p < .05.

For correlation coefficient (*r*), “values in the .8 or .9 range are usually considered good while values in the .5 or .6 range are considered worryingly weak” (Butler, 1985; Cronbach, 1990, See Lumoa, 2004: 182). Table 2 shows that the *r* values for the three tasks in the first round of rating are .800, .703, and .814 respectively but .863, .858, and .881 for the second round. The correlation coefficients are all significant at 0.01 level and the *r* values in the second round of rating are all higher than their counterparts in the first round. However, this is still not enough to tell whether the difference between the two rounds is significant in terms of reliability due to two possible reasons: on the one hand, the score difference by different raters on a 0-5 point scale cannot be too marked; on the other hand, most students did relatively well in the achievement test. As a matter of fact, the scores mainly fall within the range of 3-4. Therefore, it is not difficult to get a relatively high *r* value in this study and, although the *r* values in the second round are higher than those in the first round, it does not tell to what degree the inter-rater reliability in the second round of rating had been improved.

By comparison, the *k* values present a clearer picture: in the first round of rating they are .260**, .186* and .119 respectively. That is, according to Table 1, the inter-rater reliability has just reached the degree of slight agreement in the first round of rating. However, the *k* values for the second round are .463**, .638**, and .516** respectively, and this means that, with CARS, the inter-rater reliability has reached **the degree of moderate agreement** in the second round, about two grades higher than the first round.

**4.1.2 Intra-rater Reliability Checking**

As mentioned above, in both rounds of rating, 10 repetitions were used for evaluating intra-rater reliability. In each round, one pair of raters dealt with the same task, so 20 pairs of scores (from first rating and repetition rating respectively) were collected for each task. Since Cohen’s *kappa* coefficient (*k*) can be used for measuring inter-rater agreement, it can also be employed for evaluating intra-rater reliability by comparing the first rating scores with the repetition scores. Therefore, both correlation coefficient (*r*) and Cohen’s *kappa* coefficient (*k*) were calculated for intra-rater reliability. In addition to making analyses for
three tasks respectively, the data were combined for the three tasks and a further analysis was conducted based on a total of 60 samples. The results are shown in Table 3.

Table 3. Intra-rater Reliability Report

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 1-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Round</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( r )</td>
<td>.829**</td>
<td>.768**</td>
<td>.848**</td>
<td>.809**</td>
</tr>
<tr>
<td>( k )</td>
<td>.172</td>
<td>.272*</td>
<td>.249*</td>
<td>.236**</td>
</tr>
<tr>
<td>2nd Round</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( r )</td>
<td>.876**</td>
<td>.710**</td>
<td>.907**</td>
<td>.877**</td>
</tr>
<tr>
<td>( k )</td>
<td>.656**</td>
<td>.449**</td>
<td>.586**</td>
<td>.584**</td>
</tr>
</tbody>
</table>

** means \( p < .01 \); * means \( p < .05 \).

Table 3 shows that the \( r \) values for two rounds are .829, .768, .848 and .876, .710, .907 respectively. They are all around 0.8 and also significant at .01 level. That is, all the pairs of scores between initial rating and repetition rating in both rounds are significantly correlated. It also shows that the \( r \) value for Task 2 in the second round is slightly lower than that in the first round so the \( r \) values alone cannot tell whether intra-rater reliability has been significantly improved or not.

However, the \( k \) values show clear differences between the two rounds. In the first round, the values for three tasks are .172, .272*, and .249* respectively, none of which is significant at 0.01 level. By comparison, in the second round, they have been improved to .656**, .449** and .586** respectively, all significant at .01 level. The analysis based on the total of 60 pairs of data for the three tasks produces similar results. The \( r \) values are .809** and .877** while the \( k \) values are .236** and .587** respectively. Therefore, the \( k \) values indicate that intra-rater reliability has risen from slight agreement in the first round to moderate agreement in the second round of rating.

It can be concluded that the employment of CARS has improved both intra- and inter-rater reliability. At the same time, this research has also indicated slight disagreement between correlation coefficient and Cohen’s \( \kappa \) coefficient. It follows that for reliability checking, Cohen’s \( \kappa \) coefficient (\( k \)) is more precise and reliable than correlation coefficient since it has filtered out the agreement by chance and has better discrimination; it helps make the difference more discernable.

### 4.2 Qualitative Analysis

As mentioned above, at the end of each round of rating, a conference was held among the raters. The conference was designed to solicit views and suggestions from the raters about the rating system. In the meantime, tape-recording was not adopted so as to ensure that the raters aired their views freely and the data collected were more valid. Instead of recording, notes were taken. After the conference, all the data obtained were carefully analyzed and summarized. At the conference, raters mostly agreed on the advantages of rating with CARS over the first round of rating, which could bring about higher reliability.
4.2.1 Advantages of Instantiated Tapes

At the conference, the raters pointed out that as students’ performance always varied in a large range and raters’ understanding of criteria differed from person to person, it was very difficult to accurately assess students’ oral performance, especially with the traditional criteria of mere verbal description. By comparison, instantiated criteria were easier to apply because raters could frequently compare test-takers’ performance with the benchmark tapes of different levels so that the rating results were more accurate. Most raters agreed that, with a range of benchmark tapes in mind, they not only referred to the verbal description of the criteria, but also compared students’ performance with benchmark tapes. Therefore raters could use the criteria more easily and accurately.

4.2.2 Advantages of Single-task Rating

In most school-based achievement tests, teachers successively grade all tasks of their own students’ taped performance. In this case, the bias caused by teachers’ impression of their students’ everyday classroom performance is unavoidable. Common suggestions of reducing subjectivity include rating test performances one task at a time and trying to score performances anonymously (Brown & Hudson, 2002, See Luoma, 2004: 179).

With CARS, test-takers’ performances were divided according to task types and then randomly distributed so that raters only had to deal with one single task and had equal access to performances of different levels. At the conference, the raters unanimously pointed out that, when dealing with more than one task, it was more difficult to memorize all the details of the whole set of criteria for different tasks and a rater’s judgment of a later task was likely to be influenced by his or her impression of the test-taker’s performance in the earlier tasks. As one rater put it, “raters tend to assign a test-taker’s oral performance to a certain grade once he or she begins to talk.” Therefore, it was difficult to avoid the influence of the first impression made by test-takers’ performance in earlier tasks. And the uneven distribution also had negative influence on rating. For example, raters maintained that, when rating a relatively poor group of students, they were likely to give comparatively high scores and vice versa. Therefore single-task rating will not only help improve raters’ understanding of rating criteria but also avoid cross-task influence and reduce cross-group influence.

4.2.3 Advantages of On-line Training and Verification

At the conference, the raters all agreed that rating speaking was highly subjective and that raters were likely to deviate considerably from the criteria. For example, the raters’ internal standards may change due to fatigue or increased familiarity with tasks and performances. As a result, some raters might be too strict while others might be too lenient. Mullen (1980: 91-101) argues that two raters are required in any speaking test, as individual raters tend to have different patterns of rating. Indeed, most investigations of reliability in oral testing recommend the use of at least two raters (de Charrupe, 1984: 63-79) in order to avoid the possible impact that a single rater may have on the test score. However, the employment of two or more raters means much more human power and cost and therefore may not be workable for a school-based achievement test. The present computer-aided rating system
Research on Improving Rating Reliability for School-based Oral English Achievement Tests

helps to solve the above problem without incurring extra cost.

They all agreed that the automatic verification of CARS “forced” them to develop a stronger sense of responsibility, which, in turn, won them higher inter- and intra-rater consistency. To be specific, due to different backgrounds and lack of special training, not all raters for school-based exams were equally absorbed in or devoted to rating. And in traditional rating some raters were likely (sometimes very likely) to grade the taped performances without adhering to the rating criteria but by their own experience. However, as the raters were all teachers, most of them were willing to do their best in rating, especially when they knew their rating performance was assessed. And they would try to improve their rating when they found their weaknesses. Therefore, the verification of consistency could serve as an important reminder for those who otherwise tended to be “absent-minded” in rating.

4.2.4 Advantages of Sound Wave “Reading”

The raters all showed much interest in using a sound wave device, which helped engage them in the rating process, because, with this device, they not only listened to but also “read” test-takers’ performance. For example, when finding some typical errors (identified by human intelligence) in a test-taker’s performance, they could “catch” them for feedback simply by copying the corresponding sound waves, as is shown in the highlighted part between 3:30-4:00 in Figure 2. In achievement tests, such typical errors, which were otherwise not possible to record, were not only a very important evidence for rating but also a significant addition to feedback. Moreover, the availability of sound waves made it possible for raters to save much time by skipping silent parts as shown in the part between 1:00-2:30 in Figure 2. In this way, much time was saved. In both cases, raters could focus more attention and efforts on rating. Some raters put it in this way:

Sound waves engaged us both aurally and visually in the rating process, which added a new dimension to rating. It was easier for us to achieve agreement both with ourselves and other raters.

![Sound Wave Sample of Test-Takers’ Oral Performance](image)

**Figure 2.** Sound Wave Sample of Test-Takers’ Oral Performance

4.2.5 Lowers Rating Workload

The raters all agreed that, compared with traditional rating, the computer-aided rating system saved them much time and effort and they could be fully focused on rating since CARS automatically carried out such tasks as score calculation, statistical analysis and
score reporting. Less workload means less fatigue so that higher reliability will follow.

4.3 Discussion

4.3.1 Influence of CARS on Rating Reliability

The present rating system has such functions or advantages as instantiation of rating criteria, task division and random distribution, on-line training and reliability verification, sound wave “reading” as well as lower rating workload. The above analysis shows that, these functions could help raters employ rating criteria more accurately and focus more attention on rating, which brings about significant improvement in both inter-rater and intra-rater reliability.

On the one hand, CARS can help raters use rating criteria more accurately and easily. Instantiation of rating criteria offers bench-mark tapes of different levels, which can be directly used to “measure” students’ oral performance. Task division and random distribution provides equal access to taped performances of different levels so that cross-group influence on rating, namely, the influence caused by uneven distribution of students’ oral performances, could be greatly reduced. Single-task rating can help raters better master specific criterion or criteria, which helps avoid the cross-task influence. And on-line training can offer a relatively individualized training program. Raters can receive training according to their individual demands. In summary, all these functions of CARS make it easier for raters to achieve higher accuracy in using rating criterion or criteria.

On the other hand, with CARS, raters could be more focused on rating. On-line verification will remind raters of their possible failure to achieve a relatively high consistency. The use of Cohen’s kappa coefficient ($k$) for reliability verification can filter out the agreement by chance and has comparatively good discrimination. This greatly raises raters’ reliability awareness. Since one’s attentiveness will be greatly improved when listening to and looking at something at the same time, wave “reading” can help raters be more focused on rating through processing the waves visually and manually. The automatic processing of statistic data and related calculation work greatly relieves raters’ workload, so that raters can spend more effort in rating. In short, CARS helps raters pay more attention and efforts to the actual rating process, which leads to higher consistency.

4.3.2 Further Improvement of CARS

The analysis of the two types of data also suggests that the design of CARS can be upgraded to ensure higher reliability through the following ways.

First, on-line training and verification mechanism should be upgraded. In the present design, raters are required to receive on-line training when they fail to pass qualification verification. As a result, further on-line training is more often than not seen as a signal of failure, which makes raters feel unduly nervous. Some of the raters put it in this way:

CARS should be designed to improve rating by offering raters more convenience but not punishment. So it should provide raters with access to on-line training and benchmark tapes whenever they deem it necessary, but not just as a means of “punishment”. 
Therefore, the system should be upgraded to give an easy access to on-line training so as to help raters improve their understanding and use of instantiated rating criteria, and then adjust their judgment.

Second, self-discipline development mechanism should be involved in the future upgrading. Although self-discipline or sense of responsibility is not an applied linguistics issue, it is extremely important for solving such linguistic problems as improving rating consistency. At the conference, the raters all agreed that it was extremely difficult for non-professional raters to achieve high consistency in rating oral English performance and a sense of responsibility or self-discipline was vitally important in ensuring a relatively high consistency. Without it, no rating could be reliable. It was suggested CARS should be upgraded in order to motivate raters to further strengthen their self-discipline.

5. Conclusion

The present research has indicated that, in school-based oral English achievement tests, CARS can significantly improve both inter- and intra-rater reliability through the instantiation of rating criteria, task division and random distribution, on-line training and reliability verification, providing sound waves for raters to “read” and “write” as well as lowering the raters’ workload.

However, this research was conducted on a small-scale experiment, and it is difficult to carry out a large-scale study of a similar kind due to practicality reasons. If similar surveys can be conducted online, raters can offer their ideas at any time in rating and a larger quantity of data will be available. Therefore, CARS can be upgraded to include an online survey mechanism to help gather ideas for further improving computer-aided rating. This will bring about higher reliability and validity for school-based oral English tests, which, in turn, will exert more positive washback effect on oral English teaching and learning.

References

Department of Higher Education of MOE. 2007. 大学英语课程教学要求 [College English

Heaton, J. B., 1988, Writing English Language Tests. Pearson Education.


(Copy editing: Anthony ZAK)