International Journal of Learning, Teaching and Educational Research
Vol. 5, No. 1, pp. 26-34, May 2014

Computer-aided Assessment Standardisation for Writing and Oral Presentation Assessments: Design, Development and Implementation

Dr Voyce Li
The English Language Centre
The Hong Kong Polytechnic University
Hong Kong

Abstract. Computer-aided assessment (CAA) has been widely applied to summative assessments in English language teaching and learning. However, its usage is limited to computer-marked exercises, e.g. multiple choice questions or short answers. Assessments on essay writing or oral presentation are still lacking without human intervention. In addition, computer-aided tools for assessment benchmarking have been commonly neglected. This should be of concern to IT specialists when facilitating language assessment through technology. An online English language assessment standardisation platform (ELCAS) was introduced to reduce discrepancies among raters. The platform was further developed, and adopted by the English Language Centre, Hong Kong Polytechnic University for several years, and the project team won a Faculty Award in 2009 for its outstanding performance in developing this online assessment benchmarking tool to assist teachers in achieving consistency and inter-rater reliability in grading assessed assignments. The primary contribution of this paper is to share the ideas and design of the platform, experience of its development, and the problems encountered during implementation which are of interest for CAA standardisation.

Keywords: computer-aided assessment standardisation; ROLE- and CASE-based concepts; writing and oral presentation assessments; holistic and component grading

Introduction

In the past, when the English Language Centre (ELC) taught a very limited variety of subjects the rater training was done in pre-assignment and pre-course meetings. The ELC provided a website with some student scripts corresponding to various grades as references for script markers. The activity of benchmarking was done offline. However, in view of the substantial increase in number of English subjects offered to host departments and the consequential diversification of assessment needs and criteria, it had become logistically

impossible in terms of scheduling a large number of standardisation meetings. A better system was needed, in which teachers could be trained individually and asynchronously to rate scripts to an agreed standardised grade; therefore the ELC undertook a funded project to upgrade the pre-existing online assessment benchmarking tool to cater for the increasing complexity in achieving consistency and reliability in grading assessed assignments by developing a rater training functionality.

The Ideas for a Development of the System

The initial user requirements were to 1) expand the pre-existing system to cater for more subjects and 2) add a rater training functionality to achieve assessment standardisation for holistic and component grading. The project team looked for suitable open source content management software for further development but failed. Possibly, as Web 2.0 was attractive to educators, this was where software developers focused their attention. Therefore, computer-aided tools for assessment benchmarking were neglected. Since it was difficult to find a suitable kit for such development, the team eventually decided to build the product in-house.

Simple-and-flexible (SNF) is the key concept of the design for the platform. The merit ideas are the adoption of ROLE- (Sandhu, Coyne, Feinstein & Youman, 1996) and CASE-based (Aamodt & Plaza, 1994) approaches. The ROLE safeguards different levels of tasks to authorised users; the CASE defines different situations to respond to the need of users. The system reacts with the users based on the status returned by either the ROLE or CASE or both. These two approaches maximise the flexibility for the change of user requirements. In addition, same categories of data based on their own criteria were put into an array with delimiters together with the record in order to 1) simplify the data structure, 2) ease the change of criteria, and 3) reduce the access time to the database.

The Task Flow for Subject Leaders and Markers

There are two main tasks for subject leaders and one for markers (Table 1). The task flow first starts with a subject leader creating an assessment entry. Secondly, markers grade the scripts selected by the subject leader of an assessment. Lastly, the subject leader finalise the grades among all markers.

Six steps for creating assessment entry were identified in the first row of table 1 at the column of 'Tasks of Subject Leaders'. In step 1, options for essential assessment details are provided for subject leaders to select in order to avoid human errors (e.g. typos). For assessment type (writing, individual or group presentation/discussion), a selection menu for number of speakers will then appear when group presentation/discussion is selected. For marking mode (holistic or component), the assessment criteria and a selection menu corresponding to the weightings of each criterion appears when component marking is selected (Fig. 1).

Table 1: Task flow of subject leaders and markers

Tasks of Subject Leaders	Tasks of Markers
To create assessment entry, 1. select i. subject ii. assignment number iii. assessment type iv. number of reused scripts v. marking mode vi. markers 2. confirm details 3. select an existing task sheet / upload a new task sheet to an existing task category / upload new task sheet to a new task category 4. select old scripts (optional) 5. upload new scripts 6. DONE	To do nothing
To add more markers and/or scripts (optional)	To grade and/or comment, 1. click on a script to grade 2. read the essay or listen/watch the audio or video file 3. select a grade/component grades for holistic/component marking 4. accept or override the overall grade computed with component grades 5. give comments if needed 6. save the input and click next script to grade 7. save intermit input if needed 8. click on the scripts whenever for changes 9. click 'Submit' once all the scripts have been graded.
To finalise grade, 1. (i & ii as markers) 2. review the grades submitted by the markers 3. (iii - viii as markers) 4. click 'Submit' once all the scripts have been finalised.	To do nothing
	Check the discrepancies

Step 2 lets subject leaders check what they have selected in step 1 from a tidy web form. Subject leaders are allowed either to click 'Confirm' to go to the next step or make changes by clicking 'Edit' to go back to step 1.

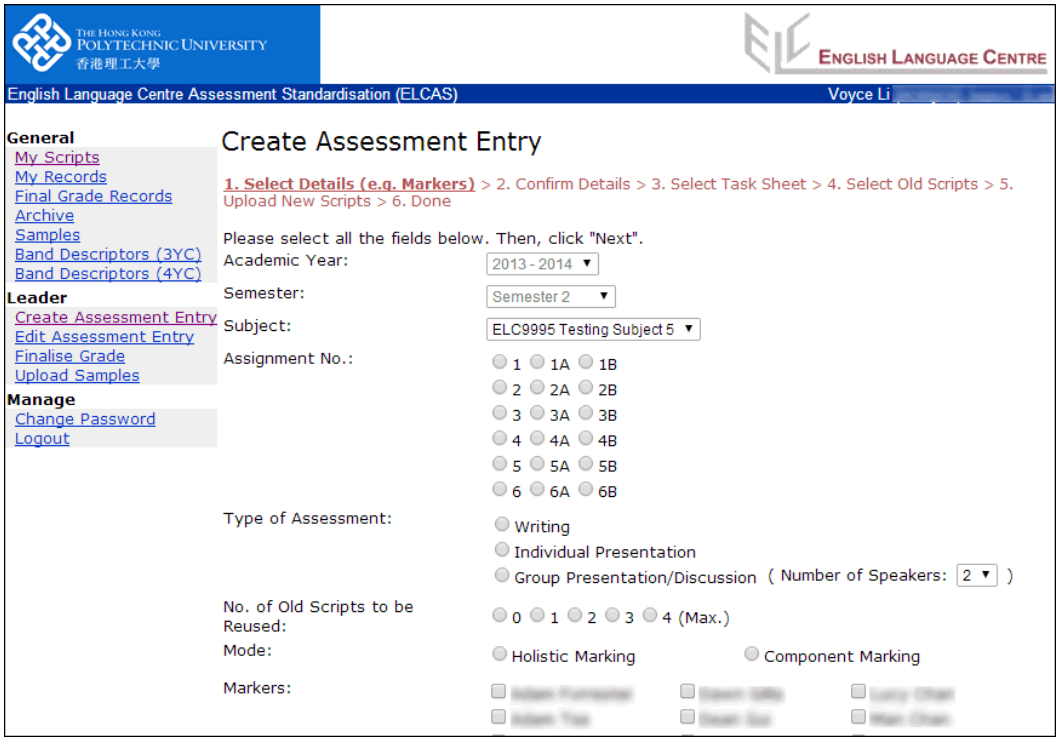

Figure 1: The interface for creating a new assessment entry

In step 3, three cases and the corresponding actions were defined. Case 1: the assessment task paper already exists in the repository – provide selection menu for choosing, and also the selected task paper is able to be viewed to avoid any mistake. Case 2: the task sheet is new to an existing task category – provide an upload function to upload the new task sheet onto a particular task category. Case 3: the task category is brand new – provide a textbox to add a new task category, and then allow a new task sheet to be uploaded onto that task category.

Step 4 can be skipped if the default value (pre-set '0') for number of reused scripts is not changed in step 1. Number of selection menu for the reused scripts will appear corresponding to the number of reused scripts selected in step 1. The selected reused scripts are able to be viewed to avoid any mistake.

Step 5 lets subject leaders upload five students' essays (doc or pdf) or presentations (wma or wmv) at most in one assessment if no reused script(s) is/are selected in step 4. If some scripts have been chosen in step 4, the number of upload will be reduced to 6 scripts at most in an assessment. For group discussion, only one audio or video file is allowed. The maximum file size for each upload is limited to 120MB. Step 6, the last step is to indicate the assessment has been successfully created.

After the assessment has been created, the set of assessment files including the task sheet and student scripts will be pre-loaded on the first page for the markers selected in step 1 when they log into the system to facilitate their grading. Since markers might change their mind after some scripts have been marked, the system allows changes before submission, and also allows an incomplete assessment to be saved whenever the markers need a pause.

The interface for subject leaders to finalise grades is similar to that for markers except subject leaders can see all grades from markers for each script of the assessment on a table to facilitate the process of standardisation. After the grades have been finalised by the subject leaders, markers are then able to see the finalised grades, and the grades given by other team members anonymously. It is important in the rater training to allow raters to learn the discrepancies without pressure. All the finalised scripts will be indexed for further benchmarking after the current semester.

The Three Phases of Development

Performance, Cost, Time and Scope (PCTS) are the constraints of project management that have mutual influence (Lewis, 2005). Since the Cost was fixed and limited, the scope for the phase I development was scaled down to the minimum in order to maintain performance under time pressure.

1. Phase I – Guinea Pig

In the academic year of 2007-08, the system only supported two tasks: grading from markers and finalisation from subject leaders. The system allowed markers to grade and give comments on a set of selected scripts of student writing. After all markers had finished their grading, the subject leader finalised the grades for the same set of student scripts, and then notified markers about the discrepancies if there were any. All the pre-standardisation work was done offline. The preparation work included collecting task papers and student scripts (scanned into pdf format if the original files were not electronic), and then uploaded onto the server manually. Once the corresponding files were ready on the server, an assessment standardisation entry was created at the backend. The system was rough and non-expandable at that moment.

2. Phase II – On the Track

In the second year, the project ran out of money; however, more requests came after a review. We were requested to 1) allow subject leaders to create assessment standardisation entries and upload the task papers and student scripts whenever they needed, 2) support assessment standardisation on oral presentation (recorded as videos or audios) for individuals (1-to-1: one task paper mapped to one student recording with grades and comments) and group discussions (1-to-many: one task paper mapped to more than one student recording with grades and comments), 3) enhance component marking to support various sets of assessment criteria for different subjects, and also subjects can have different assessment criteria in different semesters, 4) support three at most out of the same set of finalised student scripts along with some

new student scripts to be used for new exercises of standardisation, and 5) build an archive with the finalised scripts for reference purpose.

Upon receiving new user requirements, the original system became inadequate - it lacked flexibility for absorbing changes. The team, therefore, gave up the old system and re-built it with new ideas. The ideas of ROLE- and CASE-based were brought in. The design of the system started at the point with a database structure. This time, the constraint of Cost was eliminated – it was absorbed by regular working hours. We were given a more flexible time for expanding the scope. The final platform adopted ASP.NET with C# programming language (object oriented), and built on the top of Windows OS.

3. Phase III – Refining

All old benchmark scripts were indexed and archived as references for markers. The scripts can be accessed on the same platform by searching by semesters and/or subjects. In addition, the top ten markers with the least discrepancies were listed by semesters as an achievement of rater performance. Interestingly, forgetting of passwords was found to be a common phenomenon in each semester. Instead of resetting password ad hoc by requests, a function to retrieve passwords by the users themselves was introduced.

Implementation Issues

The trial run took place in 2007. After a re-construction, the platform was officially launched in 2008. Up to now, over ten thousands assessment records are kept from the database. Implementation is always a stage of the emergence of unexpected issues. The issues were identified either by observation or reported by users (the teachers), and solved immediately (for critical or minor changes) or during semester break (for non-critical or major amendments). However, there are still some issues that cannot be solved without human intelligence.

- Request of changes after submission

The top issue is 'request of changes after submission'. In order to maintain data integrity, changes are not allowed after a process is confirmed to be completed. For example, after an assessment entry is created. Some markers may start their grading based on the set of task paper and scripts selected by the subject leader. In this stage, any changes of the task paper and/or the selected scripts could possibly affect the validity of the grades, which have already been given to some scripts. To improve the system, deletion and amendment of the corresponding files are still restricted before a cascading data checking done by a human. However, subject leaders are allowed to add more markers (no limit) and/or scripts (totally 6, it is 5 initially at most in one assessment) even the assessment has been created. As the same token for grades finalisation, after markers have submitted the grades for an assessment, and if the subject leader has already started to standardise the grades, markers are not allowed making any changes to the grades and/or comments in the assessment.

- File size is an issue

Besides the issue of allow-or-not-allow-changes, preparing video or audio files for oral presentation assessments encountered far more difficulties than that for writing. Li's (2010) findings show that lack of essential equipment and the complexity of recording process are the barriers of students for their submissions of oral presentation assessments, and the ratio of submissions in writing to recording is 7:2. We came across several problems while supporting assessments on individual or group presentation/discussion. First, the system received video files in extreme large size, meaning that all of these files had to be converted into a streaming format in a lower bit rate in order to save storage at the server side, and also shorten the start time of playing at the client side. File size is always an issue - the system was adjusted to limit each upload to a maximum of 120MB and to accept only .wma or .wmv file format.

- Shared video mapped to multiple students' records

When assessing writing or individual presentation, one student script or video file is mapped to the grade(s) or comments to this particular student. When assessing group presentation/discussion, however, multiple upload of the same discussion video file for the number of students in the video becomes unpractical. To solve this problem, the program was amended to map one video file to multiple assessment records if it is an assessment for a group of students. The markers were told to grade the students from left to right in the order of Student 1, Student 2 and so on. However, students were not fixed in one position in some video files. Eventually, students were labelled as Student 1, Student 2, … in the video files.

- Multiple subject leaders mapped to one subject

During the time when the platform was being developed, each subject was coordinated by one subject leader. Due to the new 334 curriculum, two subject leaders were assigned to co-coordinate one subject to share their workload. Keeping abreast with the latest changes in pedagogy, the system was adjusted to support more than one subject leader in one assessment by the merit of the concept of ROLE-based.

- Multiple assessment criteria with different weightings

Under the demand of component grading, the set of assessment criteria with fixed weightings for academic writing, which had been used for some years, no longer fitted the assessment requirements of the subjects developed for the new curriculum. Therefore, the database and the program were revised to accommodate the data that was used to compute the final grades. In addition, an option of NA was added for individual criterion. This means irrelevant to the particular assessment and allows that criterion to be taken away from the set of assessment criteria.

- Dual-mode marking mapped to one single assessment

At the first launch of the platform, assessments were allowed to be marked either holistic or in components. Nevertheless, most of the assessments were marked holistically. It was requested that some scripts which had been given

holistic grades, could also be later marked using component grades. In accordance with the need of the users, the system was revised to allow holistic and component grading for different scripts in the same assessment (i.e. holistic for students A and B but component for student C). There has been a long discussion as to whether holistic or component assessment mode is better to address the issue of fair judgements and staff workload in the ELC. Finally, a policy was introduced to require all assessments to adopt component grading in 2012.

Conclusions and Future Study

The project won a Faculty Award in 2009. Its prototype was modified in May 2010 for another project to support secondary school English teachers assessing students' writing and reading skills during summers in 2010 and 2011 that were based on the criteria identified in the Curriculum and Assessment Guide issued by the Education Bureau. Finally, this project won a Faculty Award in 2011 and a President's Award in 2012. To achieve such sustainable development, simple thinking (with wide vision) and flexible action (with deep consideration) are crucial. Based on the 6 years of experience in developing and implementing the platform of CAA Standardisation for assessing students' writing, reading and oral presentation skills, I argue that the success or failure of the projects critically depend on the direction of design at the earliest stage. The processes are sensitive dependency on initial conditions according to the concept of Lorenz's butterfly effect (Lorenz, 2000). A bad or undesired initial status even could make any rectification impossible.

Apart from the impact of design on the development process, the assessment mode used in assessing students' writing and oral presentation skills is another issue affecting the quality of outcomes during the implementation. There are a number of options for different types of rating scales; however, little research has been conducted on how different rating scales affect rater performance (Barkaoui, 2007). In fact, Barkaoui's findings indicate a higher inter-rater agreement with holistic scale than that with component scale. Also, in Schaefer's (Schaefer, 2008) study, a six-component rating scale was used, where some raters tend to rate higher ability writers more harshly, but lower ability writers more leniently. The results indicate a potential rater bias in EFL writing assessment when component grading was adopted. Further studies on how different rating scales affect to rater performance are worthwhile.

References

Aamodt, A. & Plaza, E. (1994). Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. Artificial Intelligence Communications 7(1), 39-52.

Lewis, J.P. (2005). Project Planning, Scheduling & Control. 4E. McGraw Hill. ISBN 978-0-07-146037-8.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. Assessing Wriitng, 12(2), 86-107.

Li, V. (2010). eSelf-assessment: A case study in English language learning (Hong Kong) for enhancing Writing and Oral Presentation Skills. Education Technology and

Computer (ICETC), 2010 2nd International Conference, 4, 371-375, 22-24 June 2010, Shanghai. ISBN: 978-1-4244-6367-1, DOI: 10.1109/ICETC.2010.5529662.

Lorenz, E. (2000). The Butterfuly Effect, in R. H. Abraham & Y. Ueda (Ed.), The Chaos Avant-garde: Memories of the Early Days of Chaos Theory, World Scientific Publishing Co. Pte. Ltd., 91-94.

Sandhu, R., Coyne, E.J., Feinstein, H.L. & Youman, C.E. (1996). Role-Based Access Control Models. IEEE Computer, 29(2), 38–47.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. Language Testing, 25(4), 465-493.