ReCALL **25**(3): 321–339. 2013 © European Association for Computer Assisted Language Learning
doi:10.1017/S0958344013000141
First published online 26 June 2013

Computer-based oral exams in Italian language studies

C. PAUL NEWHOUSE
Edith Cowan University, Western Australia
(email: p.newhouse@ecu.edu.au)

MARTIN COOPER
Edith Cowan University, Western Australia
(email: m.cooper@ecu.edu.au)

Abstract

In this paper we report on one component of a three-year study into the use of digital technologies for summative performance assessment in senior secondary courses in Western Australia. One of the courses was Italian Studies, which had an oral communication outcome externally assessed with an oral performance for which students travelled to a central location and undertook an interview with two assessors. Apart from the logistical difficulties for both students and the organising body, this method did not leave an enduring record of the process, and raised questions about the reliability of the assessment. Over the three years of this study, we tried several approaches to using digital technology to assess oral performance, including a portfolio of sub-tasks leading up to a video-recorded oral presentation, a computer-based exam, a video recorded interview, and an online exam that included oral audio-recordings. For each of the years online marking tools supported two methods of drawing inferences about student performance from the representations: the more traditional analytical method and the comparative pairs method. Rasch analysis of the results of the two methods showed that both were at an acceptable level of reliability. Overall, students and teachers reported that they liked using audiovisual recordings and online performance tasks for revision but not for summative assessment. The study also demonstrated that the scores from externally marked computer-based oral tasks carried out in class time correlated highly with the scores from traditional face-to-face recorded interviews. Therefore, online assessment of oral performance appears to be an equally effective way to facilitate assessment when compared with traditional methods and offers other affordances, such as convenience and access from a variety of locations, as well as providing an enduring record of student performance.

Keywords: performance assessment, oral language, computer-based exam, portfolio, comparative pairs marking, adaptive comparative judgements

1 Introduction

In this paper we present the results of one component of a three-year study that commenced in 2008 and was conducted by the Centre for Schooling and Learning Technologies (CSaLT) at Edith Cowan University (ECU) in collaboration with the Curriculum Council of Western Australia. The study investigated the potential

for digital technologies to support more authentic forms of assessment in four senior secondary courses in Western Australia: Applied Information Technology; Engineering Studies; Italian Studies and Physical Education Studies. This paper focuses only on the Italian Studies course component of the study. In this component our brief was to develop and trial computer-supported assessment tasks that would validly and reliably measure oral language proficiency. The assessment tasks needed to be manageable in terms of cost and logistics within a typical school environment, and capable of being scaled-up for state-wide implementation. Although the context of the research was the Italian language it was considered that the findings would be equally applicable to the assessment of other language courses.

The need for better forms of assessment is increasingly seen as integral to improving schooling (Kozma, 2009). In 2011 the President of the United States of America spoke at length on the need to measure performance in ways other than traditional exams (eSchool News, 2011: 15). However, these alternative forms of assessment need to generate defensible measures: that is, valid and reliable measures of the intended performance. A further reason to investigate alternative forms of assessment in the Italian Studies course was that the form of oral performance assessment used was expensive and potentially unreliable. It required students to travel to a central location and undertake an interview with two assessors who judged performance in real time. This presented logistical difficulties for both students and the organising body, and the lack of an enduring record, combined with on-the-spot marking, challenged the reliability of this form of assessment. As a result digital forms of assessment were designed, trialed and modified over the three-year period of the study. Principally these were forms of digital audio or audio-visual recording of oral performance as well as forms of computer-based exams. Oral assessment data were captured using digital video cameras, computers, and through 'live' online systems. Hence, we both developed the digital assessment tasks and evaluated their implementation in school settings.

We will firstly give a brief overview of the literature around the assessment of practical performance, methods of marking, and computer-supported language assessment before moving on to the method, results and conclusions of the study with particular emphasis on the final phase in 2010.

2 Literature review

The study connected with three main fields of research subsumed within the general field of assessment: performance assessment, methods of marking, and computer-supported assessment.

2.1 Assessment of practical performance

Many educational researchers argue that traditional paper-based, limited-response assessment fails to either measure learning processes and higher-order thinking skills (Lane, 2004; Lin & Dwyer, 2006), or achieve validity (McGaw, 2006). Lin and Dwyer (2006) argue that the focus should be on capturing "more complex performances" (*op. cit.*: 29) but suggest that this is seldom done due to "technical complexity and logistical problems" (*op. cit.*: 28). Ridgway, McCusker & Pead (2006: 39) see a danger that, "considerations of cost and ease of assessment will lead to the

https://doi.org/10.1017/S0958344013000141 Published online by Cambridge University Press

introduction of 'cheap' assessment systems which prove to be very expensive in terms of the damage they do to students' educational experiences." Hence the strong rationale to consider the efficacy of performance assessment using alternative methods.

In order to be judged a performance needs to either be observed by the assessor or represented in some form. This may involve the assessor observing a student performing or judging the results of a recorded performance. Where the forms of performance can be represented digitally (e.g., audiovisual recording), the work can be made available to assessors easily and cheaply using digital repositories and networked computers. Historically, various methods have been used to assess oral language performance, including interviews, role plays, and group discussions. Often performances have been assessed by raters who are present and, hence, considerable training is required to enable the raters to give reliable scores. The most common approach is a short interaction with a native speaker judged against a set of descriptions of achievement standards (McNamara, 2000). Oral samples may be included in a language portfolio or language dossier, along with other samples of language performance (Cummins & Davesne, 2009; Myers, 2002).

2.2 *Methods of marking*

Task assessment is what is commonly referred to as 'marking'. The performance on an assessment task needs to be judged to determine a score, grade or ranking. Three methods of marking were considered: (1) 'traditional' true score marking, (2) comparative pairs judgements, and (3) analytical marking using standards-based frameworks. An elaboration of each follows. For the current study we used both analytical and comparative pairs marking to assess student output.

The traditional approach is, as Pollitt (2004: 5) puts it, to sum 'true' scores on "micro-judgements". Politt explains that this approach is likely to generate scores with low reliability for the measurement of "performance or ability". Typically the primary requirement is to provide a ranking of students and therefore, he argues, comparisons between performances using more holistic judgements and Rasch dichotomous modelling will provide this and also result in a reliable interval scale. The results of implementing a comparative pairs approach to marking that he helped implement for the e-scape project attested to the saliency of this argument (Kimbell, Wheeler, Miller & Pollitt, 2007).

The comparative pairs approach to marking requires assessors to select a 'winner' between a pair of performances, and repeat this process for many pairs, with the results analysed using a Rasch model for dichotomous data (Pollitt, 2012). Whereas Pollitt (2004) describes the comparative pairs method as "intrinsically more valid", he believes that without Information and Communications Technology (ICT) support it has not been feasible to apply due to time and cost constraints. McGaw (2006) believes that such assessment methods, supported by digital technologies, should be applied in public examinations.

The standards-based analytical method of marking typically uses a framework to construct a rubric for a particular assessment task that describes these standards according to components of the task. The resulting judgements may be represented as a set of levels of achievement or may be combined by converting these to numbers and adding them. However, using Rasch polytomous modelling these judgements may be combined to create an interval scale score. In this paper we refer to this approach as *analytical marking*.

https://doi.org/10.1017/S0958344013000141 Published online by Cambridge University Press

2.3 Computer-supported language assessment

Computer-supported assessment, or computer-assisted assessment, encompasses several applications of computers for assessment processes. The British Psychological Society (2002) produced guidelines that provide a conceptual model: *Assessment generation*; *Assessment delivery*; *Assessment scoring and interpretation*; and *Storage, retrieval and transmission*. All four areas of application were relevant to our study that used a combination of computer-based exams, digital portfolios and digital recordings of performance. There are well-documented examples of computer-based exams—for example, in Canada and Norway (British Broadcasting Corporation, 2009; Carbol, 2007). Our study also used some of the technology of the UK *e-scape* project in which students used handheld computers to respond to questions and capture audiovisual evidence of activity in design and technology, science and geography (Kimbell, *et al*., 2007). Wiegers (2010) reports that in the Netherlands computer-based exams are used to examine skills not able to be assessed on paper, in order to increase alignment with life requirements, to increase flexibility in delivery of assessment, and to reduce workload.

Although computers have been used for over a decade in the assessment of written and listening-based aspects of foreign language learning (Jamieson, 2005; Ockey, 2009), with a recent example of an online tool used for comprehension reported by Vincent-Durroux, Poussard, Lavaur & Aparicio (2011), assessing oral performance using computer technology has not been adopted widely, largely due to limitations in processing language. According to Douglas and Hegelheimer (2007: 125), "no one has yet developed a computer-based simulation that adequately reflects truly interactive oral language". One of the best examples is *Versant for English,* which automatically scores recordings made over the telephone (Pearson Education Australia, 2012). This uses speech recognition and natural language processing technologies; but there is some debate about their validity with concerns that the delivery medium may change the nature of the construct being measured (Downey, Farhady, Present-Thomas, Suzuki, & Can Moere, 2008). Douglas and Hegelheimer (2007) further question how computer-based tasks engage learners.

> It seems reasonable to assume that the use of, say, a video of two people having a conversation in a listening test would enhance authenticity by providing a context and verbal cues for the test takers. However, questions have been raised about whether test-takers actually look at the video, preferring to concentrate solely on the audio input. (*op. cit*.: 117)

Our research used video to simulate conversations as part of oral assessment. Software has yet to be developed that can simulate a true conversation, although there are some sophisticated chat-bots that have been developed (Pandorabots, 2012).

Garmire and Pearson (2006: 162) explain that while computer-based assessment has the potential to increase "flexibility, authenticity, efficiency, and accuracy", it must be subject to "defensible standards". The use of digital technologies in high-stakes school-level performance assessment is relatively rare, due to feasibility concerns about cost, logistics and technical reliability (Lin & Dwyer, 2006). Dede (2003: 9) suggests that the barriers are now not so much technical or economic as "psychological, organizational, political and cultural". That is, participants, educators, leaders and

community members are not adequately convinced of the efficacy of computer-supported or computer-based assessment.

3 Method

In the design of the study we followed Barrett (2005) who suggests every assessment should have three foundation pillars: (1) A model of how students represent knowledge and develop competence in a content domain; (2) Tasks or situations that allow one to observe students' performance; (3) An interpretation method for drawing inferences from performance evidence. In this study we considered how digital technologies could support the representation of student knowledge and competence with respect to a practical performance, the tasks or situations that would deliver such representation, and how that could be judged through methods of marking. This brought together three key features.

1. The representation in **digital files** of the performance.
2. The presentation of these digital representations in an **online repository** accessible to assessors.
3. The judgement of these digital representations using both **analytical** and **comparative pairs** marking methods.

The project built on the work of Kimbell *et al*. (2007) at the University of London and applied their feasibility framework in the final interpretive analysis of all the data collected (see Table 1).

Table 1 *Feasibility framework used to evaluate digital forms of assessment*

Dimension	Description
Manageability	Concerning making a digital form of assessment do-able in typical classrooms with the normal range of students.
Technical	Concerning the extent to which existing technologies can be adapted for assessment purposes within course requirements.
Functional	Concerning reliability and validity, and the comparability of data with other forms of assessment.
Pedagogic	Concerning the extent to which the use of a digital assessment forms can support and enrich the learning experience of students.

The research design for this study can be described as a participatory responsive evaluation with three evaluative cycles, requiring an analysis of the perspectives of the key groups of participants (teachers, assessors, students), and with qualitative and quantitative data collected from each group. There were three phases to the study, each of a year in length; however, the focus of this paper will be on the third phase, with a brief account of the first two phases. The first phase was a 'proof of concept' to explore the feasibility of particular digital forms for external assessment. The second phase focussed on developing a prototype of the digital forms of assessment. Finally, in the third phase, the implementation was scaled up to involve a larger sample of representative schools. In each phase the intention was to implement the same assessment task for each case (school) and collect a range

of data on the implemention and marking of the assessment task. Each year these data were used to refine the assessment task for the following year. Teachers were recruited on the basis that they would agree to implement the assessment task.

In essence, the study involved the development, implementation and evaluation of digital assessment tasks, with participants involved in all three of these aspects.

3.1 Developing the assessment tasks

At the beginning of each phase of the study a situation analysis was conducted by a team comprising researchers, curriculum officers, assessment officers and teachers. In the second and third years these built upon the results emerging from the data from the previous year. As the Italian Studies course already had a tradition of assessing oral performance through a face-to-face 'interview' undertaken at a central location, the initial focus was on improvements to this approach in terms of validity, reliability and logistics. Ultimately approaches were trialled that either simulated a conversation or recorded the student speaking.

3.2 Data collection for the evaluation of the assessment task

Over the three years there were fifteen schools, fifteen teachers and fifteen classes consisting of a total of 184 Year 11 or 12 students whose data were included in the Italian Studies component of the study. For each class the data collected involved: observation of the class completing the assessment; a survey of students; interviews with students, teachers and assessors; and scores generated by the three methods of marking (included marking by the teachers using their own methods). These data were analysed both for each teacher-class case study and for the combined sample.

Initially a researcher liaised with each teacher to plan to implement the tasks. Then students were observed working on the tasks and technical support was provided on-site if required. After the completion of the assessment, a researcher ensured the representations of the students' performance were collated using either an online system or an external storage device. The students were surveyed and a small group interviewed in a forum. The teacher completed a short questionnaire. After the external analytical marking was completed the teacher was provided with a short report on the data collected from the class and was given the opportunity to respond.

3.3 Survey of students

The student survey used a questionnaire of 46 closed-response items and two open-response items. This instrument was created from one used in the *e-scape* project and one used in a previous project conducted by the research centre. Descriptive measures were calculated for each closed-response item. Responses to the open-response items were tabulated to assist in drawing out themes. The following six scales were derived from combining sets of closed items:

eAssess: A score between 1 and 4 to measure the efficacy of the digital assessment tasks.

Apply: A score between 1 and 3 to measure the types of tasks to which computers are applied.

Attitude: A score between 1 and 3 to measure attitude toward using computers.
Confidence: A score between 1 and 3 to measure confidence with using computers.
Skills: A score between 1 and 4 to measure specific ICT skills.
SCuse: An estimation of the average time per day (in minutes) spent using computers at school.

Some descriptive statistics for these scales are shown in Table 2 and distributions for two are represented in the graphs in Figure 1.

Table 2 *Descriptive statistics for the scales based on items from the student questionnaire*

	Year 1 (N = 32)				Year 2 (N = 41)				Year 3 (N = 86)			
	range	$\bar{x}$	SD	α	range	$\bar{x}$	SD	α	range	$\bar{x}$	SD	α
eAssess	1.9–4.0	2.7	0.5	0.9	1.7–3.6	2.6	0.4	0.8	1.8–4.3	2.9	0.6	0.8
Apply	1.6–3.0	2.4	0.4	0.4	1.4–3.0	2.3	0.4	0.5	1.7–3.0	2.3	0.3	0.0
Attitude	1.6–3.0	2.4	0.3	0.1	1.6–3.0	2.6	0.3	0.5	1.6–3.0	2.5	0.3	0.5
Confid	1.0–3.0	2.5	0.5	0.9	1.0–3.0	2.6	0.5	0.8	1.2–3.0	2.5	0.5	0.8
Skills	1.9–3.7	2.9	0.5	0.9	1.6–4.0	3.0	0.6	0.9	1.6–4.0	2.8	0.6	0.9
SCUse	0–120	36	32		0–132	47	41		0–252	52	50	

α = *Cronbach's Alpha reliability coefficient*

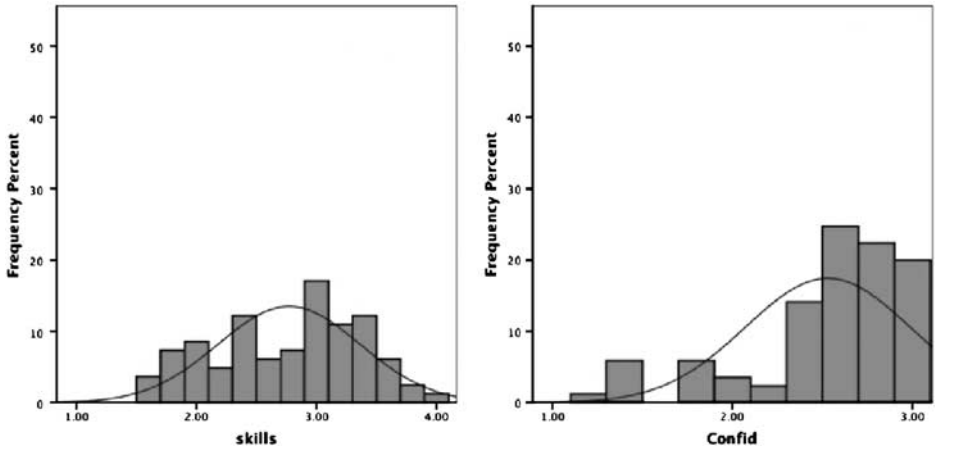

Fig. 1. Graphs for the distribution of scores for Skills and Confidence scales on the student questionnaires in the third phase

In all three phases of the study the survey revealed that students had little experience in doing assessments on computers (72%, 63% and 80% respectively indicated *little* or *no* experience) and felt they would need *some* or *lots* of time to get used to the process (around 80% of students). Almost all had a range of technologies at home, with 65% owning a video camera through to 98% owning an mp3 player, and about 90% owning a laptop computer, with 95% having broadband Internet access. By the third phase 94% indicated using a computer at home on most days. About 72% felt confident with

https://doi.org/10.1017/S0958344013000141 Published online by Cambridge University Press

computers and liked using them. Of the ICT skill areas listed, students felt least confident about web authoring, databases, spreadsheets, and video editing (around 50% indicating no skills). Overall, the students indicated a moderate level of skills with a mean on the *Skills* scale of between 2.8 and 3.0.

3.4 Marking criteria, tools and assessors

Analytical marking criteria were developed by the research team from the assessment tasks and the syllabus outcomes. These were incorporated into the analytical marking tool as a rubric. A single holistic criterion was distilled from these criteria to be used for the comparative-pairs marking. Performances were judged by two external assessors using the analytic method and by a group of assessors using the comparative pairs method. External assessors were recruited from curriculum officers and teachers. Each year a training workshop was held for the assessors to learn how to use the system and to ensure a consistent interpretation of the assessment criteria. During this workshop the first round of judgements was completed. Assessors were then able to continue the process from home or work. They used online marking tools to access the representations of performance and to record judgements.

For the first two phases of the study, marking tools for both analytical and comparative pairs marking were custom built using the *Filemaker Pro* relational database software. For the third phase, the rubrics were sent to the *Willock Information Systems* company and made available through their online assessment module for teachers and assessors for analytical marking. For comparative pairs marking the online marking tool called the *Adaptive Comparative Judgement System* (ACJS) (Pollitt, 2012), developed by TAG Learning for the *e-scape* research using MAPS, was used. Student responses (typed and oral recording) in digital form were downloaded from the *Willock Information Systems* website and uploaded to MAPS. Assessors logged on to the ACJS, were presented with pairs of performances, and indicated their selection of the better of the two. An administrator's logon allowed access to control and reporting tools. At the close, the system calculated a score in logits using Rasch dichotomous modelling and generated a report that included graphs, reliability measures and data on each assessor. A substantial difference with this system compared to the custom built system used in the first two phases was that for the latter all assessors had to complete a pre-determined set of assigned judgements, the results were then downloaded into a spreadsheet and analysed using the RUMMcc software using a Rasch dichotomous model.

For each phase of the study, scores from external assessors and teachers were analysed for the entire sample, as well as for each school case study. This included descriptive statistics for each source of scores and correlations between each. The correlation coefficient between the scores provided one test of the reliability (between external analytical assessors and between methods of marking) and validity (between external assessors and teachers) of the measures.

4 Results of the study

This section summarises how the tasks were implemented, the technologies used, and the main results from the analysis of data over the three phases.

https://doi.org/10.1017/S0958344013000141 Published online by Cambridge University Press

4.1 Phase one – proof of concept

The first phase involved four teachers with a total of 35 Year 11 students developing a folio over about twelve hours and making an oral presentation. The folio was a series of tasks to show development of ideas and preparation for the oral presentation and included a map activity, a retrieval chart and question answers, brainstorm, fact sheet, a word-processed reflection and a one-minute voice recording. The two-minute oral presentation was video recorded by a researcher using a radio microphone attached to the student. The folio tasks were only fully completed by students from one school. There were no technical issues with the video recording of the oral presentations. The digital photographs and word-processed documents received were converted to a PDF file for each student. The native video of the digital camera (.mpg) was converted to WMV to ensure compatibility. These data (PDF file, audio file, video file) were placed into student folders on a server.

4.1.1 Results of marking. Two assessors marked the folios and video recordings online using the analytical method. These assessors were joined by three others for the comparative pairs marking to judge the recordings of presentations, each making 120 comparisons. All five were Italian language educators.

Scores from the analytical assessors were very low (mean = 35%) because many students did not complete all the folio tasks. The teacher, however, marked according to his/her own methods and this resulted in higher scores (mean = 61.4%). There was a very strong significant correlation between the scores given by the two external analytical assessors ($r = 0.93$, $p < 0.01$), indicating high reliability.

The RUMM software used to analyse the data from comparative pairs marking provided a Separation Index (SI) of 0.832 as one indicator of reliability. There was a moderate to strong correlation ($r = 0.70$, $p < 0.01$) with the scores from analytical marking, but only a weak correlation with the teachers' scores ($r = 0.48$, $p < 0.05$).

4.1.2 Perceptions of participants. The *external assessors* were generally positive about the processes of online marking except for having to wait for a video to download before viewing. The *teachers* identified the need for more explicit links between the portfolio sub-tasks and the final presentation and that the task conditions needed to be more comparable between schools. They also felt that the task was too rehearsed and did not reflect the students' ability for oral communication.

The *students* generally agreed (62%) with the assertion that "It was easy to use digital technologies for the assessment tasks and oral exam". About two thirds agreed that, overall, digital technologies were good tools for parts of the assessment. However, 70% disagreed that it was better doing the oral exam with digital technologies than face-to-face with an examiner. The mean on the eAssess scale was 2.7 (Table 1), only slightly above the mid-point (Figure 2). In the open-response items they indicated enjoying the task but felt nervous with the video camera and would prefer just to be audio-recorded.

https://doi.org/10.1017/S0958344013000141 Published online by Cambridge University Press

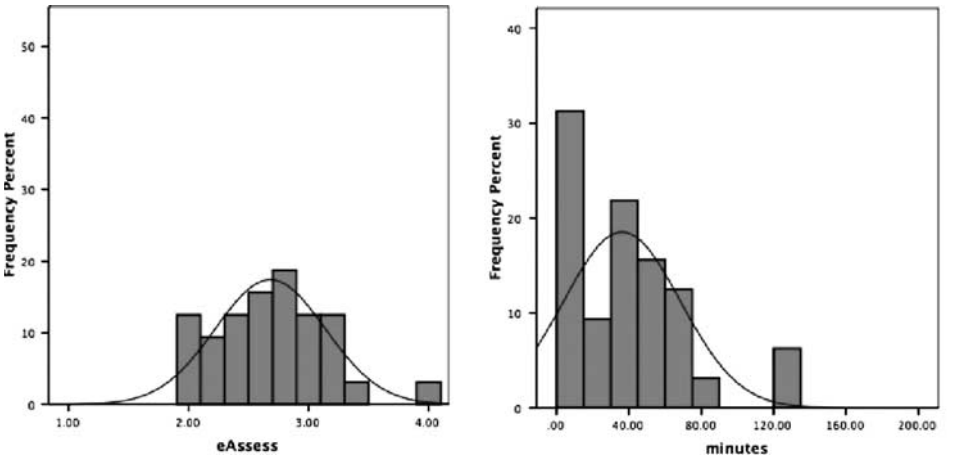

Fig. 2. Graphs for the distribution of scores for the eAssess (LHS) and School Computer Use (RHS) scales on the student questionnaire in the first phase of the study

4.2 Phase two—prototype

The second phase of the study involved four teachers with a total of 52 Year 12 students. A major goal for the first phase was to prepare students for their final oral examination. However, the teachers did not believe that this had satisfactorily occurred and thus it was decided to redesign the assessment tasks. Part of the final examination involved the selection of an image to form the basis of a 'conversation'. To mirror this, the oral performance became a video-recorded conversational interview and the folio became a number of computer-based oral recordings to prepare for the interview. For the folio, the students responded to either stimulus pictures with a question or a set of written questions (see Figure 3), with their responses digitally recorded using a microphone connected to the computer with the *e-scape* exam management system (Kimbell, *et al.*, 2007).

The recorded interview comprised two parts closely matched to the final official examinations. The first consisted of selecting a set of stimulus materials with a focus question, and following a 15-minute preparation time, engaging in a 4-minute conversation with the external assessor. The second was participating in an 8-minute in-depth conversation initiated by the student. An external assessor and the classroom teacher judged the performances on the spot. Students from all schools involved were video-recorded with recordings being edited into short clips (.wmv) for each student and placed on a server.

The computer-based exam tasks were completed by 52 students from four schools with no technical malfunctions. The *e-scape* client software, including the content, were preloaded onto USB flash drives, an onerous task involving duplicating and then launching the software (with a live internet connection) and logging in as the student. Later the researcher had to connect the USB drive and login to upload the audio recordings into MAPS. Uploading at the time of the task was not possible due to firewall restrictions at the schools involved.

https://doi.org/10.1017/S0958344013000141 Published online by Cambridge University Press

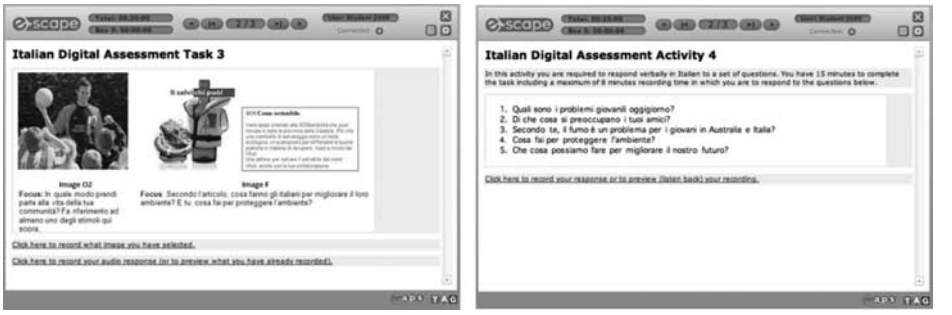

Fig. 3. The two types of e-scape activities – left = image & stimulus question and right = stimulus questions

4.2.1 Results of marking. External marking was conducted in the same manner as in the first phase, except that the external assessment of the in-class online tasks was done through MAPS. As a result of this, the students were provided with a score and feedback.

The external assessors and teachers on average gave a similar set of scores by analytical marking (mean of 61% and 58% respectively). The external assessors took an average of ten minutes per student portfolio. There was a strong correlation between the two analytical assessors ($r = 0.77$, $p < 0.01$). There was also a strong correlation between the external assessors and the teachers' oral exam marks ($r = 0.80$, $p < 0.01$). Additionally, there was a strong correlation ($r = 0.74$, $p < 0.01$) between the online tasks scores and the external markers' average for the oral examination, which suggests that the scores from these types of online tasks may be used as a reasonable predictor for the oral exam scores.

4.2.2 Perceptions of participants. Teachers, assessors and students felt that the tasks were of an appropriate standard and that the stimuli were well selected. Students appreciated feedback on their performance; however, the teachers believed that the tasks did not accurately reflect a conversation and would be better if modified so that the students listened to, rather than read, the stimulus questions.

The *assessors* had some concerns about the use of only three grade descriptors for each criteria. They also found it difficult having to juggle between two browser windows and manually locate a portfolio in the MAPS system. For comparative pairs marking, there were complaints about the time taken to view videos. However, overall assessors were positive about the process and found it easier to use than the system used in the previous year.

The *teachers* felt that the feedback process took too long to be received in some cases and this had a tendency to divorce the feedback from the actual task. They felt the oral examination was run well and that the video camera had little impact on performances. They believed that it was necessary still to have two markers present to help with on-the-spot moderation of results. However, they were supportive of the concept of a video recording that could be assessed after the examination.

The *students* generally agreed (66%) that is was easy to use digital technologies for the oral exam. However, most (73%) disagreed with the assertion that "I would prefer

https://doi.org/10.1017/S0958344013000141 Published online by Cambridge University Press

to do the oral presentation examination using digital technologies than face-to-face with an examiner". The mean on the *eAssess* scale was 2.6 (Table 2), around the mid-point, with scores following close to a normal distribution (Figure 4). From the open-response items, students indicated appreciating being able to analyse and critique their own performances on the online tasks through reviewing the audio recordings and feedback. However, they also indicated feeling distracted by others when carrying out the in-class online tasks. They reported feeling nervous being video recorded, and over 70% felt that it would be better to do the oral face-to-face rather than to use digital technologies.

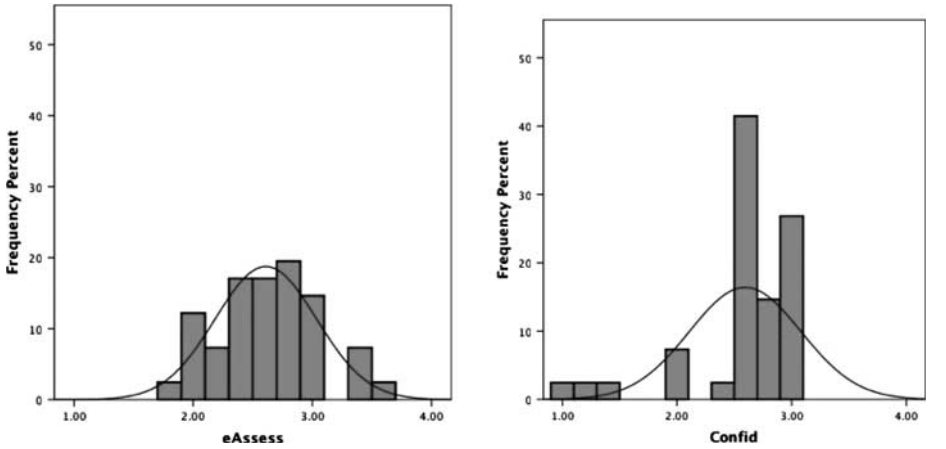

Fig. 4. Graphs for the distribution of scores for eAssess (LHS) and Confid (RHS) scales on the student questionnaire in the second phase of the study

4.3 Phase three – scale up

For the third phase the assessment was implemented at seven schools with a total of 97 Year 11 students. The assessment task was expanded to include further outcomes from the course—*Viewing, Reading and Responding*, and *Listening and Responding*. Because the teachers in the previous year did not believe the online tasks had adequately simulated a conversation, the computer-based exam was modified to include video clips of conversations, the third of three components: Parts A, B and C. Part A was three pre-assessment online tasks designed to allow familiarity with the system in about 80 minutes. Parts B and C (20 and 12 minutes respectively) were conducted online on the same day and essentially formed one assessment task. In Part B students listened to a radio interview in Italian and then responded to questions by typing in English and Italian (see Figure 5). In Part C students listened to three short video clips of a dialogue between a student and her teacher (see Figure 6). At the end of each clip one of the actors turned to the camera and asked a question. Students then completed an oral response in Italian that was recorded on the computer. The intention was to simulate a conversation.

All tasks were completed through a web browser, using an online system provided by *Willock Information Systems*. All typed responses and audio recordings were

https://doi.org/10.1017/S0958344013000141 Published online by Cambridge University Press

Fig. 5. Part B of the computer-based exam as viewed in a browser

Fig. 6. Part C of the computer-based exam as viewed in a browser

uploaded automatically to the system's server. Computers needed to have *Java* installed, an Internet connection, web-browser, and headsets with a microphone. Prior to implementation, school computer systems were tested. At the two schools with *Apple MacBooks*, problems were experienced with the audio-recording elements of the assessment task. Therefore at these schools Part C was not completed.

4.3.1 Results of marking. Student performances were marked analytically by two assessors using tools provided by the *Willock* system. However, for comparative pairs marking the responses had to be downloaded and then uploaded to the MAPS portfolio system to be attached to the ACJS. For each student there was a PDF file containing text-based responses to Part B and up to three audio files for Part C. The exam output for only 50 of the 97 students (those that completed *PartBC* and had at least one audio file) were marked in this way by eight assessors. The holistic criterion used was, "*Which of the pair of students being judged was the best communicator in the Italian language?*" – where 'communicator' represented both the student's ability to speak fluently in Italian (reflected in audio responses) as well as to understand the spoken language (reflected in responses to listening task). The time taken for each judgment was from about two minutes to 7.5 minutes per student.

https://doi.org/10.1017/S0958344013000141 Published online by Cambridge University Press

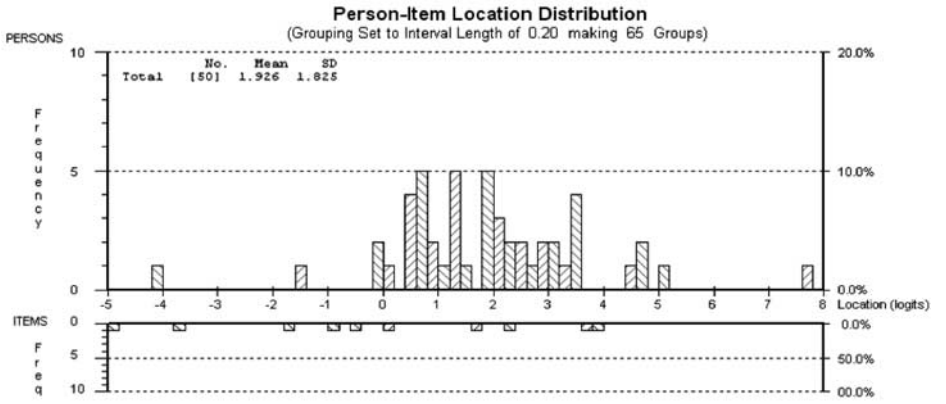

Fig. 7. Person-item location distribution for analytical marking of Part C (oral) of the assessment task

The teachers scored the performances well above the analytical assessors (means of 70% and 59% respectively). There was a strong correlation between the two assessors ($r = 0.88$, $p < 0.01$), indicating good reliability. There was a low correlation between the average of their scores and those awarded by the teachers ($r = 0.48$, $p < 0.01$).

The results of analytical marking were analysed using a Rasch polytomous model for each of Parts B and C. Scores for each criterion for the two external assessors were included for each student as separate items. Both analyses supported the contention that the two components realised reliable scores.The Part B listening component realised a high Cronbach's Alpha index of 0.89, and no item showed significant misfit. For the Part C oral component, the analysis yielded a high Cronbach's Alpha index of 0.91. Figure 7 shows the range of the person locations and item locations were similar. The total chi-square probability was 0.64, no item showing significant misfit, and all thresholds were in order.

The ACJS provided a reliability coefficient of 0.939 after the 13th round. Only ten scores had a standard error of measurement above 1.1. The system provided statistics on the consistency of the eight judges, with only one lying outside one standard deviation. There was a moderate to strong correlation ($r = 0.77$, $p < 0.01$) with the scores from analytical marking. However, for some students there were substantial differences in ranking from the two different methods of marking (up to 25 positions).

The absolute differences between the analytical rank and pairs rank were compared with absolute differences between the rankings from the two analytical assessors. There was a strong correlation between differences in ranking based on method of marking and the results from each of the two analtyical assessors ($r = 0.89$ and 0.93, $p < 0.01$), but there was no correlation between these and the differences in ranking between the two assessors. This lack of relationship between differences in ranking based on method of marking and differences based on different assessors suggests different sources of difference. The seven students with a difference in ranking between the two methods of marking of more than 20 were reviewed by

https://doi.org/10.1017/S0958344013000141 Published online by Cambridge University Press

looking at the comments of assessors and the nature of the performance. It was concluded that these students may have undervalued Part C because this was not marked by their teacher and that the comparative pairs assessors gave more weight to Part C than Part B.

4.3.2 Perceptions of participants. In general the assessment task was acknowledged by the teachers to be faithful to the course. However, the oral component was considered by both teachers and students to be inferior to the traditional face-to-face oral communication assessments. This was due to both technical problems and the nature of simulating a conversation. There were some concerns identified about distractions with audio-recording and inconsistencies in the administration of the tasks, with some students being permitted to write their responses before reading them and others having to answer 'on the fly'. The *teachers* generally considered the assessment task to be interesting, consistent with the course themes, requiring a variety of skills, set at the correct standard, and with potential for assessment in Italian. However, they felt that the oral communication component was best done traditionally with a face-to-face examiner, primarily because it did not represent a natural conversation.

The *assessors* agreed that the tasks were suitable for students and were similar to the types of tasks used for language learning. Additionally, the breadth of the tasks allowed students of all levels to demonstrate performance. They identified various levels of engagement with the tasks, largely dependent on whether the students' results were going to contribute to the actual school grades. The students who performed well appeared to be those who had completed the pre-assessment tasks. Other factors possibly influencing the standard of performance included that students felt strange talking to a computer, that they were unable to ask for clarification, that some spent too long preparing an answer, reducing the authenticity of conversation, and a lack of exposure to a variety of ways of assessing oral communication. One assessor commented that the production type tasks would be likely to carry more weight (this was borne out statistically with the oral component appearing to carry more weight in the pairs judging).

The *students* generally (70%) indicated that doing the assessment tasks on the computer was easy, although only 40% preferred this to the traditional face-to-face exam. The mean score of 2.9 on the *eAssess* scale indicated a generally positive perception of the digital assessment task (Figure 8). The only contributing item with over 50% disagreeing was "Overall, digital technologies are good tools for Italian assessment tasks and examinations". Given the responses to the other items, it is likely that this response represents a lack of experience in tackling this type of activity and limited use of computers, with most less than 50 minutes a day (Figure 8). In the open-response items they indicated that this form of assessment gave them more flexibility and less pressure because they could go at their own pace; furthermore, it seemed less formal and intimidating, they were able to use the dictionary, and they could pause the recording. Most students also considered this form of assessment to be easier, faster, neater, and for some it was enjoyable. However, many students experienced some technical problems, felt self-conscious when recording, and were distracted by others recording.

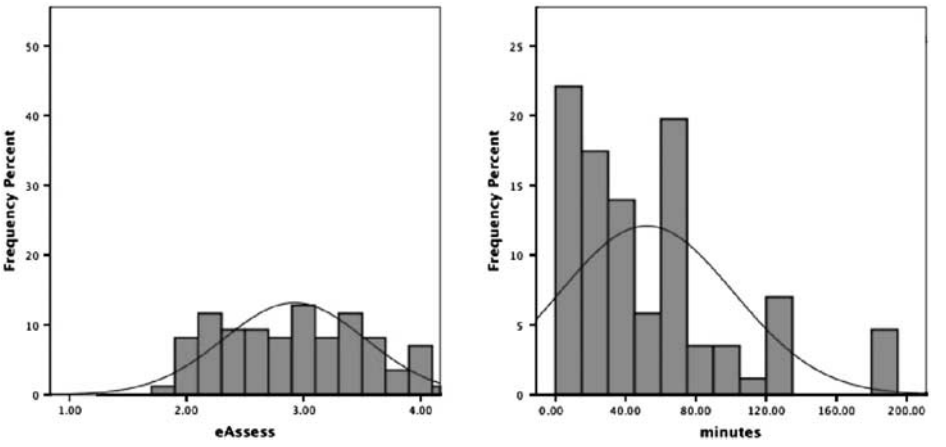

Fig. 8. Graphs for the distribution of scores for eAssess (LHS) and School Computer Use (RHS) scales on the student questionnaires in the third phase of the study

5 Discussion of feasibility

The results of data analysis were interpreted using the feasibility framework with the functional dimension divided into validity and reliability.

In terms of the *Manageability Dimension,* all forms of assessment tried were manageable in all schools provided the teacher could book enough computers. In the first two phases the process of collating students' work would not be manageable on a large scale. The video recording of oral interviews was easily manageable and could be scaled up. The online tasks using the *Willock* system were smooth and efficient. However, development of listening and oral tasks (scripting, audio recording, and video recording) was time consuming. Some technical expertise was necessary on the part of the invigilator.

In terms of the *Technical Dimension,* the main issue concerned firewalls, authentication and network speed where online activities are involved, particularly audio recording requiring Java or Flash. In addition, using external microphones could be difficult, with sound cards, drivers and settings needing to be checked. Online systems need to ensure audio recordings are uploaded; about 30% of files did not reach the *Willock Information Systems* repository. This may require a locally saved backup copy and confirmation with the user that the process has been successful. Video recording of student oral performance provided few challenges using digital video cameras and radio or zoom microphones. Once performance representations were uploaded to servers there were few technical issues with marking.

In terms of *validity* on the *Functional Dimension,* both teachers and students generally believed that face-to-face oral language assessment had greater validity than computer-based oral assessment. However, in the second phase of the study a strong correlation was found between scores on the online oral tasks and the scores on the face-to-face oral exam. Many students did not believe that the technologies enabled them to perform at their best; however, some felt under less pressure than in face-to-face interview-based assessment. Many students had little or no experience

https://doi.org/10.1017/S0958344013000141 Published online by Cambridge University Press

with completing online assessment tasks. Teachers generally felt that video recording of face-to-face exams could lead to fairer assessment.

In terms of *reliability of measurement* on the *Functional Dimension,* by the third phase of the study both the scores from external analytical marking and from comparative pairs marking were found to have high levels of reliability. For example, Rasch analysis of the results of analytical marking yielded initial Cronbach's Alpha indices of 0.887 and 0.914 for the two assessed components, and inter-rater reliability correlations were strong. Comparative pairs judging results achieved a Cronbach's Alpha reliability coefficient of 0.94. However, further analysis showed that the oral component carried more weight and therefore it would probably be preferable to assess the listening and oral components separately.

In terms of the *Pedagogic Dimension,* it was clear that using computers to support oral language learning and assessment was foreign to most teachers and students. Therefore, it was not surprising that initially many students were apprehensive about being video recorded or recording their own oral responses. However, eventually many students found online oral tasks useful in preparing for oral exams.

6 Conclusion

Ultimately the outcomes of the study need to be viewed as a balance between the affordances of the digital forms of assessment used and the constraints identified through the feasibility framework. From the beginning the potential of video-recording of oral performances and the post-performance online judging was clearly manageable, technically feasible, valid and reliable and readily aligned with current practices. In contrast, any form of assessment that required students to use computers, particularly for oral recording, was not perceived by teachers and students to be adequately valid for summative assessment and posed some technical constraints in some schools. Computer-based oral tests were found to be manageable and could be judged reliably leading to results equivalent to face-to-face assessments. Students found that the digital technologies enabled them to critically reflect on their performance, they could proceed at their own pace, and have a distraction-free listening experience using headphones. However, the video recording of oral performances created a level of apprehension among students that may have led to reduced performance. Teachers maintained a strong belief in the primacy of a real 'conversation' as the most effective form of assessing oral performance, despite the logistical difficulties and the threats to the validity of scores from real-time judgements.

In general the schools had the required technologies to undertake digital assessments that deliver media (text, audio, and video) and captured student performance, whether that was in the form of text, audio, or video. There are commercially available systems that can accommodate these kinds of assessments. However, in some schools there were serious technical constraints in capturing oral recordings online. Teachers and assessors were positive about the experience of online marking, finding the systems convenient and fast. Online marking of digital representations of oral performance appeared to be at least as reliable as traditional face-to-face methods and offered greater convenience and a method of storage of student performance. This suggests that it is possible to assess oral performance reliably and validly using

means other than the traditional face-to-face conversational method. This may have logistical benefits for both the students and the awarding body.

The results of the study would suggest that awarding bodies should consider the potential of digital forms of assessment for oral communication outcomes. However, care would be required if the collection of audio and video assessment materials entailed online systems in real time. There are clear cost-benefit advantages to assessing oral language performance using digital methods of simulated conversation. It is likely that computer access and the limitations of online technologies will be quickly overcome to support this approach to assessment. However, these approaches would need to be gradually introduced to allow teachers to understand the changes in emphases needed in their teaching.

Acknowledgements

The study discussed in this paper was the work of a research team organized by the Centre for Schooling and Learning Technologies at Edith Cowan University (http://csalt.education.ecu.edu.au/). The work of everyone in this team, along with curriculum officers from the Curriculum Council of WA, and the teachers and students involved, has contributed to the research outcomes presented in this paper. Professor Richard Kimbell was a consultant to the study.

References

Barrett, H. C. (2005) *Researching Electronic Portfolios and Learner Engagement: The REFLECT Initiative*. New York: Taskstream.

British Broadcasting Corporation (2009) Norway tests laptop exam scheme. *BBC News*. http://news.bbc.co.uk/2/hi/technology/8027300.stm

Carbol, B. (2007) *Transition to Online Testing: An ROI Analysis*. Kelowna BC: Society for the Advancement of Excellence in Education.

Cummins, P. W. and Davesne, C. (2009) Using electronic portfolios for second language assessment. *The Modern Language Journal*, **93**(Focus Issue): 848–867.

Dede, C. (2003) No cliché left behind: why education policy is not like the movies. *Educational Technology*, **43**(2): 5–10.

Douglas, D. and Hegelheimer, V. (2007) Assessing language using computer technology. *Annual Review of Applied Linguistics*, **16**: 115–132.

Downey, R., Farhady, H., Present-Thomas, H., Suzuki, M. and Can Moere, A. (2008) Evaluation of the usefulness of the Versant for English Test: a response. *Language Assessment Quarterly*, **5**: 160–167.

eSchool News. (2011) Obama: Too much testing makes education boring. *eSchool News*, **14**(5): 15.

Garmire, E. and Pearson, G. (eds.) (2006) *Tech Tally: Approaches to Assessing Technological Literacy*. Washington: National Academy Press.

Jamieson, J. (2005) Trends in computer-based second language assessment. *Annual Review of Applied Linguistics*, **25**: 228–242.

Kimbell, R., Wheeler, T., Miller, A. and Pollitt, A. (2007) *e-scape: e-solutions for creative assessment in portfolio environments*. London: Technology Education Research Unit, Goldsmiths College.

Kozma, R. B. (2009) Transforming Education: Assessing and Teaching 21st Century Skills. In: Scheuermann, F. and Bojornsson, J. (eds.), *The Transition to Computer-Based Assessment*. Ispra, Italy: European Commission, Joint Research Centre, 13–23.

Lane, S. (2004) Validity of High-Stakes Assessment: Are Students Engaged in Complex Thinking? *Educational Measurement, Issues and Practice*, **23**(3): 6–14.

Lin, H. and Dwyer, F. (2006) The fingertip effects of computer-based assessment in education. *TechTrends*, **50**(6): 27–31.

McGaw, B. (2006) *Assessment to fit for purpose*. Paper presented at the 32nd Annual Conference of the International Association for Educational Assessment, Singapore.

McNamara, T. (2000) *Language testing*. New York: Oxford University Press.

Myers, M. J. (2002) Computer assisted second-language assessment: to the top of the pyramid. *ReCALL*, **14**(1): 167–181.

Ockey, G. J. (2009) Developments and challenges in the use of computer-based testing for assessing second language ability. *The Modern Language Journal*, **93**(Focus Issue): 836–847.

Pandorabots (2012) http://www.pandorabots.com/

Pearson Education Australia (2012) Versant. http://www.versanttest.com

Pollitt, A. (2004) Let's stop marking exams. Paper presented at the International Association for Educational Assessment Conference, Philadelphia, USA.

Pollitt, A. (2012) The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, **19**(3): 281–300.

Ridgway, J., McCusker, S. and Pead, D. (2006) Report 10: Literature review of e-assessment. In: Facer, K. (ed.), *Futurelab Series*. Bristol, UK: Futurelab.

The British Psychological Society (2002) *Guidelines for the Development and use of Computer-Based Assessments*. Leicester, UK: The British Psychological Society.

Vincent-Durroux, L., Poussard, C., Lavaur, J. and Aparicio, X. (2011) Using CALL in a formal learning context to develop oral language awareness in ESL: an assessment. *ReCALL*, **23**(2): 86–97.

Wiegers, J. (2010) E-assessment in The Netherlands, innovations for the 21st Century. Paper presented at the 36th International Association for Educational Assessment Bangkok, Thailand. http://www.iaea.info/documents/paper_4d22770a.pdf