Assessment of Performance Quality

Metrics for Process rather than Result

Peter Summons

*Information Systems Group, School of Management, The University of Newcastle, Callaghan
NSW 2308, Australia*

mgpfs@cc.newcastle.edu.au

Keywords: assessment, basic skills, competencies, computing

Abstract Assessment grades provide a final result typically regarded as an indicator of how well a student has mastered the concepts underlying assessment tasks. Student knowledge is normally derived from a dimension whose metric is the amount of correctly completed tasks. This result is outcome-based and indicative of *what* was done, without consideration of *how* it was done. This paper proposes that a complete assessment requires more ; that information regarding how students attempt assessment tasks is also necessary to provide a feedback mechanism for future assessment and teaching content. This paper presents a proposed research direction for an ongoing Australian project.

1. INTRODUCTION: HISTORICAL FOUNDATIONS

Studies by D'Ambra (1991) and Ablong (1991) support that knowledge of results can enhance learning and performance outcomes. Well-designed assessment can also be regarded as a learning a learning tool for students 'Quality assessment tasks also provide clues to students about what is important for them to learn' (McIntosh 1996). A practical implementation of these principles is provided by the utilisation of practice assessment tasks in conjunction with detailed feedback of assessment results. The assessment tasks can show an aspect of a course objective and thus can provide incentive and direction (Summons et al. 1997a, 1997b).

Not only can assessment and feedback benefit students but it can also be used as a feedback mechanism for educators, providing insight as to the

quality of teaching, the merit of the assessment material, and also providing direction for future teaching content and delivery. This is the proposal for future direction of the Automated Assessment Management Project (AAMP), being conducted at The University of Newcastle and Deakin University, Australia.

The use of computer software applications is taught in similar introductory subjects at Newcastle and Deakin Universities. At Newcastle, INFO101 is taught across two campuses and averages more than 1300 students per year. Deakin University offers a similar subject to more than 750 students per year. It is also taught across multiple campuses as well as being offered in an external, off-campus mode. Each subject incorporates a practical course content where students learn Microsoft Word, Excel and Access on IBM PCs. In line with the view that the '...best way to learn is by doing' (Ablong 1991) assessment of the application software component in both subjects uses skill-based, hands-on, practical tests conducted during students normal tutorials.

The Automated Assessment Management Project was started to provide a mechanism to solve the logistic and human difficulties of assessment in large classes (1000 plus), involving both on-campus and off-campus delivery. An instrument and a methodology to assess computer literacy skills (specifically Spreadsheet) were developed and have been used extensively in the last few years (Summons et al. 1997a, 1997b, 1998). To date approximately 6000 students have been tested with this instrument at Newcastle alone. We have utilised on-line, interactive real-time assessment and marking of students as well as asynchronous testing and marking.

The instrument and methodology we use reflects the initial aims of the project, many of which have been achieved (Summons et al. 1997a, 1997b). These involve both educational and logistic goals that utilise the assessment instrument to:

- Standardise marking and eliminate bias in marking individuals within groups, involving both inter- marker (different concepts) and intra-marker (moods over time) bias
- Rationalise Course Logistics & Cost enabling
- Enable cost reductions through rationalisation of staff numbers and staff time spent in Marking
- Facilitate easy Articulation and Post-testing of students
- Implement fast Turnaround of results to students
- Act as a Learning Tool through the use of Practice Tasks where students learn by doing with practice assessment tasks
- Provide comprehensive, individual feedback is provided to students
- Change the role of the tutor to the guiding expert for remediation
- Show students realistic course expectations

- Reduce student advantage gained from 'cheating'.

The project has been supported by internal teaching grants from The University of Newcastle and an Australian National Teaching Grant (Summons et al. 1997c) between the University of Newcastle and Deakin University, with Newcastle taking the overall coordinating role.

The computer program used in the AAMP is not meant to cater for every possible variation of student answer. The intention is that, although it will handle the vast majority of student responses, there will always be a need for human intervention with the ability to override or modify the software-generated score in specific circumstances. Experience has shown that this may account for one or two cases per thousand in each assessment. An implicit goal in this approach is that the overriding power resides with one person only – the course co-ordinator. The approach has been successful in eliminating past scenarios, where many independent and uncoordinated spur-of-the-minute marking decisions were made by individual tutors (Summons et al. 1998b).

Often in subjects with very large student numbers the tutor has more intimate knowledge of the potential ability or capability of the student than the central course co-ordinator. This results from the closer relationship involving more personal contact between the tutor and student. The tutor is aware of the student as a person, as opposed to the central course co-ordinator who may be only aware of the student by name, or by number, as one of the individuals making up a class of more than six to seven hundred students. The tutor is therefore in a better position to act as a local expert and mentor for the student than the course co-ordinator.

The assessment results and the individual feedback to the student outline what the student did incorrectly. The tutor may then work in conjunction with the student, using the assessment result to try and establish what the student should do to correct future work. This may be simple correction advice or it may be a set of remedial activities.

Note that at this point there is no presumption that the tutor has to be involved in this sort of remedial activity. Certainly with the pressure of significant course content it may have to be initiated by the student. Even if a student does initiate a request for clarification of an incorrect result, there is more chance that a tutor will show the student the correct method rather than try to establish *why* the individual student was incorrect. There is certainly no formal expectation that knowledge gained by the tutor in student remediation will be passed on to the course co-ordinator, apart from exceptional circumstances. This would, of course, be expected to occur if the tutor noticed a majority of students with a common misconception or complaint, indicating the possibility of an incomplete learning outcome in the former or a mistake in the marking guide/program for the later. If a small

number of individuals in each tutorial group suffered the effects of either of these problems then it might possibly remain unreported, as individual tutors would probably not notice the accumulation of case evidence unless they were teaching in multiple tutorials.

Initial surveys of the AAMP indicated a problem with the tutors acting in the role of mentor in conjunction with the automated assessment (Summons et al 1998). They showed that tutors were ineffective as remediators when they took a totally passive part in the marking process. The simple remedy of ensuring tutors had a detailed copy of the marking scheme quickly overcame this problem. However, this break in the assumed continuity of lecturer/tutor/student knowledge highlighted the potential problem described earlier; that the knowledge of how students were interpreting a question and why they were answering things incorrectly, might possibly stop at the level of the tutor. This would be important feedback to provide to the lecturer to enable modification of the content or the style of presentation of course material. Unfortunately, the assessment task results available to the central course co-ordinator merely indicates student outcomes, even if they are expressed in reasonable detail. How could we capture this additional knowledge? That answer forms the basis for the proposed and ongoing research presented in this paper.

2. THE PROPOSED RESEARCH: HOW VERSUS WHAT

Quite often we only see the results of a question and we interpret these to suit our purposes. We see *what* the student has produced, that is the outcome, but not *how* it was produced.

AAMP application testing is accomplished under the control of a computer program interacting with the student in real-time. The assessment program initiates the application and accesses both the test questions and the application data required for the assessment task, and also directs the storage of the completed student answer. It controls a student's ability to redo a question anew and governs the storage of the replacement answer. It also controls a student's ability to modify completed sections within the duration of the assessment time.

The interactive assessment program incorporates a marker function. This either marks at the end of the assessment period, or during it as the student moves from question to question. The marking function stores individual student's results and marks for each answer and for the overall test. In addition to the present creation and storage of marks as outcomes of the

student's competency, the automated assessment instrument can easily incorporate technical details regarding a student's performance.

It is intended to extend the present functionality of the assessment program and marking function to include the collection and storage of such information. This new information is specific to a student's performance for each question and also to their performance in the overall test. It consists of details such as the:

- The test
- Time spent per question
- Time spent on subsequent returns to a question
- Number of returns to specific questions
- Number of new attempts at specific questions, where a previously attempted answer is abandoned
- Number of 'modified' answers, where a previous attempt is retained but modified
- Sequence of correct and incorrect attempts for specific questions.

Other indicators could be aggregated but would require more implementation complexity. Examples of these include recording:

- Events, such as the use of 'undo', 'delete' and 'backspace' that may be applicable to specific applications at the task level
- The mouse clicks per question. These can also be associated with screen position but this would not be useful on a first-cut pilot study.

The indicators above can be matched with traditional outcome-based performance measures reflecting success, such as the student's final score on completion of a question. The combined data could indicate parameters, specific to the assessment task population that refer to the degree of difficulty experienced by students in completing particular questions.

Measures of the temporal difficulty experienced by students in answering questions coupled with indicators of student achievement, for example the percentage of correct responses to the question, can give an insight to the real complexity of the question. The temporal distribution of students who answer a question correctly (or incorrectly) adds additional information to our interpretation of the relevance and complexity of the question. As a simple example, if a question is answered correctly by the majority of the class and is finished quickly then it may represent an easy task or a well-understood concept. The reasons for poor temporal performance could be investigated and this may help future teaching. It should be noted that there could be many reasons for poor performance specific to the topic, to the course facilities, or to student demographics. Examples include a difficult concept being assessed, a poorly presented concept, or a poor or ambiguous assessment question.

Comparison of actual student answer-time versus the assessor's expected time for answering a question provides another dimension to the interpretation of how well a concept has been learned by students. Additional information is also provided when examining the results of groups of students who have answered a question in a similar time. This allows a finer granularity of student classification by creating an extra dimension of comparison. If individual student identification was supported then commonalities of students forming temporally similar groups could be investigated in a similar manner to the present capability for finding commonalities of students with similar achievement levels. This information could then be used demographically to identify advantages or disadvantages of specific tutorial groups.

Performance data can therefore be beneficial by providing extra clues for tutors regarding the most appropriate remedial activities for individual students. Traditionally, students only tend to request remedial help for questions that were answered incorrectly. If a tutor instigates remedial feedback, the fact that a student answered a question correctly may mean that the concept in the question is not included as a remedial activity as it is assumed that the student knows the concept. However, if the student took a long time answering a question, even though it was eventually answered correctly, this may indicate potential difficulty with that question concept. If students return to the same question many times, even if they do not change their (correct) answer, this may also indicate that they are unsure of the question concept. Likewise, if a student changes their answer before settling on the final correct answer, then it may indicate uncertainty of the question concept. When questions are answered incorrectly, temporal information is also useful, but this is probably not as important from a remedial viewpoint as the question concept is likely to be included in the set of remedial concepts for the student.

For the moment however, the proposed performance data to be gathered at The University of Newcastle is only intended to serve as an aggregated indicator of trends and unusual performance for the central co-ordinator. Ethical considerations constrain the gathering of performance data in that the ability of the central co-ordinator to obtain individual identification of performance indicators other than question marks will be suppressed. At the moment the assessment program identifies individual students by their system log-in identifiers, which reflects their unique student number.

To satisfy ethics requirements for gathering data at The University of Newcastle, students will have the option of choosing not to have performance data gathered about them. This will be in the form of a check box on the assessment that they can select if they wish to stop data gathering. The students doing the test will be notified that the purpose of the data

gathering is to refine future assessment tasks and evaluate the presentation of course content. They will also be advised of the anonymity of the performance data in respect to student identity. This will be done both within the assessment instrument and as prior notification within tutorials.

Figure 1. Data Set Collection

The assessment program will provide two sets of collection instruments, as depicted in Figure 1. One will collect details of Student ID and Grade, as is done at present. The other will collect details of grade and performance parameters, but will not have student identification incorporated. Whilst theoretically, individual identification from a statistical collection such as the later data set is still possible, the data set will only be available to authorised personnel interested in aggregated data indicators rather than specifics.

3. CONCLUSION

The incorporation of performance parameters provides additional temporal and sequence information regarding how a student performs in a test, rather than the traditional outcome-based parameter which simply indicates what they have done. This additional information allows us to form a more complete picture of the assessment and the assessed course content. It can help to indicate individual student weakness areas and also identify student groups with a finer granularity than is possible with outcome-based indicators alone.

Even though restrictions to the use of the performance parameters will apply at The University of Newcastle, the performance metrics will help provide a mechanism to add to our knowledge of how well concepts are taught and learned by students, and how appropriate our assessment is. It should also be noted that it is possible to use the technique in other, simpler forms of assessment carried out on computers, such as the multiple-choice style of testing.

This paper has described work-in-progress to provide additional performance metrics for the traditional outcome-based, practical assessment of computer applications. It has identified performance parameters that can be gathered and it has indicated how these measures can provide a richer understanding of the assessment outcomes, both for remedial activities for students and as feedback for future course direction.

REFERENCES

- Ablong, A. P. (1991) Theories of Learning and Their Application to Prototyping. In *Proceedings of the International Working Conference on Shaping Organisations Shaping Technology*, Adelaide, South Australia, Oct., pp. 1-11.
- D'Ambra, J. (1991) Computer Literacy – No Longer an Option. In *Proceedings of the Australian Computer Conference ACC'91 MOSAIC*, Adelaide, South Australia, Oct., pp. 111-122.
- McIntosh, W. J. (1996) Assessment in Higher Education. *Journal of College Science Teaching*, Vol. 26, Sept./Oct., p. 52-53.
- Summons, P. F., Coldwell, J., Henskens, F. and Bruff, C. (1997a) Automating Assessment and Marking of Spreadsheet Concepts. In *ACSE'97: the 2nd Australasian Computer Science Education Conference*, Melbourne, July, p. 178-184.
- Summons, P. F., Coldwell, J., Henskens, F. and Bruff, C. (1997b) Automated Evaluation of Spreadsheet Concepts – Excel in Action. In *SITIS'97 Proceedings of the Second NSW Symposium on Information Technology and Information Systems*, Sydney, Feb., pp. 87-93.
- Summons, P. F., Coldwell, J. and Bruff, C. (1997c) 1998 Australian National Teaching Development Grant – Individual. *Automated Assessment and Marking Management Package*, the Committee for University Teaching and Staff Development [<http://services.canberra.edu.au/CUTSD/announce/1998Indv.html#Newcastle>].
- Summons, P. F., Bruff, C., Simon and Coldwell, J. (1998) Automated Assessment and Management Project. In *CATE'98 IASTED International Conference on Computers and Advanced Technology in Education*, Cancun, Mexico, May, pp. 181-184.
- Summons, P. F. and Simon (1999) Authentication Strategies for Online Assessments. In *ACSE'98 the 3rd Australasian Conference on Computer Science Education*, Brisbane, July, pp. 101-105.

BIOGRAPHY

Peter Summons is a lecturer in Information Systems at The University of Newcastle, Australia. He is currently doing a PhD in Knowledge Acquisition for Medical Decision Support Systems. He has worked as a Systems Engineer and a consultant in industry. He is currently teaching information management and supervising MIT information systems projects.