Research Paper Submission

Developing a Research Design for Comparative Evaluation of Marking and Feedback Support Systems

John R. Venable, Ashley Aitken, Vanessa Chang, Heinz Dreher, Tomayess Issa, Brian von Konsky, and Lincoln Wood
Curtin University
{J.Venable, A.Aitken, V.S.Chang, H.Dreher, T.Issa, B.vonKonsky, L.Wood}@curtin.edu.au

Marking and provision of formative feedback on student assessment items are essential but onerous and potentially error prone activities in teaching and learning. Marking and Feedback Support Systems (MFSS) aim to improve the efficiency and effectiveness of human (not automated) marking and provision of feedback, resulting in reduced marking time, improved accuracy of marks, improved student satisfaction with feedback, and improved student learning.

This paper highlights issues in rigorous evaluation of MFSS, including potential confounding variables as well as ethical issues relating to fairness of actual student assessments during evaluation. To address these issues the paper proposes an evaluation research approach, which combines artificial evaluation in the form of a controlled field experiment with naturalistic evaluation in the form of a field study, with the evaluation to be conducted through the live application of the MFSS being evaluated on a variety of units, assessment items, and marking schemes. The controlled field experiment approach requires the assessment item for each student to be marked once each using each MFSS together with a manual (non-MFSS) marking control method. It also requires markers to use all the MFSS as well as the manual method.

Through such a design, the results of the comparative evaluation will facilitate design-based education research to further develop MFSS with the overall goal of more efficient and effective assessment and feedback systems and practices to enhance teaching and learning.

Keywords: Marking and Feedback Support System, Teaching Technology Evaluation, Research Design

Introduction

Marking and provision of formative feedback on student assessment items are essential activities in teaching and learning. While essential, they are also time consuming, difficult to do well, challenging, and potentially error prone, combining both intellectual and communication activities with mundane clerical work. Marking and Feedback Support Systems (MFSS) are (usually computerised) information systems (IS) developed to improve the efficiency and effectiveness of human (not purely automated) marking and feedback provision, e.g. for essays or other open-ended written assignments. Anticipated benefits of developing and using MFSS include reduced marking time, improved accuracy of marks, improved student satisfaction with feedback, and improved student learning. Recently, there has been much progress in the development and availability of MFSS (Berger, 2011; Burrows & Shortis, 2011; Campbell, 2008; Colbran, 2011b; Gibson, et al., 2011; Thompson, et al., 2011; Weinberger, 2011; Weinberger, et al., 2011; Wren, Campbell, Heyworth, & Bartlett, 2011; Wren, Campbell, Heyworth, & Lovering, 2011; Young, 2011). MFSS features vary substantially from system to system, but generally support easier and more accurate application of marking guides and rubrics and/or easier provision of feedback. This paper discusses issues relating to evaluation of MFSS and the design of a comparative evaluation research approach that addresses those issues. The results of the comparative evaluation will facilitate design-based (or design science) education research and evidence-based education through further research and development of MFSS, with the overall goal of more efficient and effective assessment and feedback to enhance teaching and learning. While the research design described in this paper was developed for a particular research project, the approach could be used for other projects evaluating the same or other MFSS.

This paper is organised as follows. The next section introduces Marking and Feedback Support Systems – their goals or intended benefits, their features, and some examples of MFSS. A description of the literature on research methods relevant to education and the development and evaluation of MFSS follows. The next two sections consider MFSS themselves and issues relating to their evaluation in terms of rigour as well as ethics. The following section describes a comparative evaluation research design that addresses those issues. Finally, remaining issues and further research are discussed in the conclusion section.

Marking and Feedback Support Systems

As noted above, marking and provision of formative feedback on student assessment items are essential but difficult and potentially error prone activities in teaching and learning. Marking and providing students with feedback can be frustrating (Bailey & Garner, 2010) and stressful (Hogan, et al, 2002) and are among the most time-consuming (Ferns, 2011), teaching activities. Marking is particularly difficult when it should also produce effective feedback to students, which aids their learning from formative assessment. The provision of quality feedback significantly increases the time that marking takes and is very difficult for units with large numbers of students (Gibson, et al., 2011). Furthermore, markers are under pressure to mark quickly so they can return feedback to the students as soon as possible 1) before the students become dissatisfied and 2) to provide a learning opportunity while the material is still fresh in the students' minds, thereby improving student learning.

In order to address the time-consuming nature of marking and to increase the quality and timeliness of feedback (along with other goals, see below), researchers as well as educational technology providers have developed and continue to develop information systems that support teachers in their marking and feedback work. Burrows and Shortis (2011) use the terms Online Marking and Feedback Systems (OMFS) and Marking and Feedback Tools (MFT) more generally. However, unlike OMFS, this research is not concerned whether the information system is online rather than being standalone (e.g. running on a PC). Furthermore, this research *is* interested in supporting, rather than automating marking (e.g. with multiple choice quizzes or automated essay grading). We therefore use the term Marking and Feedback *Support* Systems (MFSS), to emphasise the support of human users in marking, rather than automation of marking. Having set out our choice of terminology, the research design described in the paper could be used for evaluating any of the classes of tool or system named above.

MFSS Goals and Benefits

The purpose of Marking and Feedback Support Systems (MFSS) is to improve the efficiency and effectiveness of marking and feedback activities in teaching (Berger, 2011; Burrows & Shortis, 2011; Campbell, 2008; Colbran, 2011b; Gibson, et al., 2011; Thompson, et al., 2011; Weinberger, 2011; Weinberger, et al., 2011; Wren, Campbell, Heyworth, & Bartlett, 2011; Wren, Campbell, Heyworth, & Lovering, 2011; Young, 2011). This section considers these goals and desired benefits in more detail.

Improved efficiency of marking and feedback

- Reduced marking time – Reduction of the effort in marking and feedback is important to teachers so that they can reduce their workload or devote effort to other teaching activities. Managers of teaching are also concerned with teachers having reduced workload in order to reduce their staff stress levels and improve their general well-being, to allow time to do other activities and improve quality, and/or to allow staff to effectively teach greater numbers of students.

Improved effectiveness of marking and feedback

- Improved accuracy and consistency of marks – Correct and fair marks are important to teaching staff, students, and administrators. More ‘objective’ marks and marks that more correctly assess work are important to all concerned.
- Higher quality feedback – For formative assessment, a key goal is to provide feedback that is clear, meaningful, and facilitates the students’ learning. Some characteristics of feedback quality that come to mind include …
  - Legible – handwritten feedback is sometimes illegible
  - Meaningful and comprehensible
  - Specific (rather than generic) to what work the student actually submitted
  - Establishes the relationship to the material taught and general lessons
  - Helpful and constructive rather than overly critical and negative
  - Related to the educational goals
- Timelier feedback – The sooner feedback is returned, the timelier it is. Feedback received while the assessed work is fresh in the mind of the students should be easier for the student to comprehend and relate to their efforts. Students will also then have more time to incorporate what they learn from the feedback into their continued learning. Thus timelier feedback should have more utility and lead to increased learning.
- Improved student satisfaction with feedback – Happier students are better for everyone!
- Improved student learning through feedback received – This is the real goal of the feedback in the first place.

Effective marking and feedback would address the seven principles of good feedback practice suggested by Nicole and colleagues and applied to technology-supported assessment (Nicol & Macfarlane-Dick, 2006; Nicol & Milligan, 2006). According to the seven principles, good feedback:

1. “helps clarify what good performance is (goals, criteria, standards);
2. facilitates the development of reflection and self-assessment in learning;
3. delivers high quality feedback that helps learners self-correct;
4. encourages interaction and dialogue around learning (peer and teacher-student);
5. encourages positive motivational beliefs and self-esteem;
6. provides opportunities to close the gap between current and desired performance;
7. provides information to teachers that can be used to help shape their teaching.” (Nicol, 2011, p. n.p.)

MFSS Features

MFSS features that address the above goals vary substantially from one MFSS to another. Some of the available features include the possibilities below.

- More easily apply marking guides and rubrics
  - More accurately apply marking guides and rubrics – Through specific statements of what is desired in marking rubrics
  - More easily select marks according to shown criteria – Through the ability to tick check boxes or move sliders
- More easily written comments
  - Pre-written comments – The ability to select pre-written comments can save writing or typing
  - Adding new comments – The ability to record new comments that are added to a database and can be selected when evaluating subsequent assessment items
  - Editing comments – The ability to modify pre-written comments to make them more specific to the student’s actual work
- More easily place comments at appropriate places on assessed student work submissions – When comments can clearly point to a particular point in a student’s work being assessed, it is more relevant and specific to that student’s actual work

- More easily calculate total marks – Eliminating error-prone and time-consuming clerical and administrative tasks, such as totaling weighted marks and applying percentage penalties
- More easily record marks – Automatically transferring marks into a mark recording system (e.g. a learning management system) reduces errors and administrative workload
- More easily produce feedback and summaries – Generating and formatting a feedback document also reduces errors and workload
- More easily transmit feedback to students – Automatically sending feedback, e.g. via email, also reduces workload and possibly errors (e.g. sending feedback to the wrong student)

Available MFSS

Many different MFSS have been developed as part of educational research or are becoming available on the commercial market and their number is growing quickly. Burrows and Shortis (2011) identified 29 different technologies related to marking and feedback, although only 15 of those could be considered MFSS. The project for which the research method in this paper was developed (which began before Burrows and Shortis was published) included two other research prototypes. This section first considers research prototype MFSS, then commercial products.

Among the MFSS research products are ABBA, EPSS, tsAAM, Semi-Automatic Essay Assessment based on a flexible Rubric, and SAFS, each of which is briefly described here. ABBA (Aitken Black Board Approach) was developed by Dr Ashley Aitken at Curtin University (no published information is available) and uses a combination of Blackboard's online test feature and online submission of assignments to support marking assignments, including inclusion of pre-written comments. EPSS (Electronic Performance Support System) was developed by Dr Alistair Campbell at Edith Cowan University (Campbell, 2008). Current incarnations run on Apple Macintosh and iPad computers and support touchscreen marking according to rubrics, ability to add and share comments between markers, automatic calculation and totaling of marks according to a rubric, embedding of video of performances being marked in an automatically generated assessment report, and automatic (one-button) transmission of the assessment reports to each student (Wren, Campbell, Heyworth, & Bartlett, 2011; Wren, Campbell, Heyworth, & Lovering, 2011). tsAAM (technology supported Assignment Assessment and Moderation) (Berger, 2011) was developed by Daniel Berger under the supervision of Dr Christian Gütl at Technical University Graz (Austria) and Prof Heinz Dreher at Curtin University. The research prototype has been further modified and extended by researchers at Curtin University. Features include marking rubrics with easy selection of mark levels, ability to add positive (marks added), negative (marks deducted), and neutral (no mark implications) comments, either from pre-written comments or added on an ad hoc basis and retained for further use, as well as automatic calculation of marks and automatic generation and formatting of reports to be sent to the student. Semi-Automatic Essay Assessment based on a flexible Rubric (Weinberger, 2011; Weinberger, et al., 2011) was developed by Andreas Weinberger with the same supervision as the work by Berger. Short Answer Feedback System (SAFS) (Gibson, et al., 2011), developed at Curtin University, automatically marks short answer quizzes without human intervention, sending the result and feedback to the student via email.

Commercially available MFSS products include Blackboard's Gradebook rubric and marking and feedback features, Waypoint, Moodle Workshops, ReMarksPDF and ReMarksXML, Turnitin's Grademark, iAnnotate and Audacity, each of which is briefly described here. Blackboard version 9.1 contains a rubric feature where a marker can easily indicate a level of student attainment in a rubric and provide feedback, with the marks are automatically totaled and recorded within the Blackboard grade centre. Waypoint (Skeele et al, 2007) can operate as a standalone product or be integrated into Blackboard. Moodle Workshops (Cole, 2005) is a configurable tool in the open source Moodle learning management system supporting open comments and staff, peer, and self assessment. ReMarksPDF (Colbran, 2011a, 2011b) is available from ReMarksPDF.com and allows annotation of assignment submissions in pdf format, with pen, highlighting and drawing tools, colour coding, special 'stamps' or symbols, text based (typed) comments, and sound (voice) comments. Rubric/criterion based marking with automatic totals, reporting, and a marks database that allows importing or exporting. Grademark is a product from Turnitin (https://www.turnitin.com/static/products/grademark.php) that supports custom rubrics and allows a marker to drag and drop standard or custom marks and comments directly onto electronic copies of student papers. iAnnotate (http://www.ajidev.com/iannotate/) is an iPad app that allows the user to annotate PDF files (e.g. provide marks and feedback on student assignments). ReView (Thompson, et al., 2011) is available from acidgreen education at http://www.review-edu.com/ and provides marking rubrics with slider controls and the ability to add comments. Students self-assess against each marking criterion using the ReView system. Student self-assessment is not visible to the marker until after marking is completed. The discrepancy between the student self-assessed mark and actual mark for each criterion assists markers to target feedback where the gap is greatest. iFeedback (Young, 2011) allows markers to provide marks using a weighted criteria model and to provide personal feedback. Audacity (Audacity, 2011) is a general tool for recording and editing audio, such as feedback comments, which can then be appended to documents.

The above represent many, but only some of the available MFSS, with a variety of features. The first question then is, “Which is the best MFSS?”, or at least “Which of the extant MFSS are better at achieving the stated goals for MFSS and what are their individual advantages and disadvantages?” The second question is how should one go about answering the first question, i.e. what is an appropriate method for comparatively evaluating these different MFSS. Answering the second question is the topic of this paper.

Evaluation of MFSS

Some prior work has been done on evaluating MFSS. Burrows and Shortis (Burrows & Shortis, 2011; Shortis & Burrows, 2009) performed two kinds of evaluations. First, they evaluated MFSS against criteria of the features that they included (or omitted). Second, they conducted experiments using university staff to conduct mock evaluations of student assignments. Data was gathered in the form of open comments on different topics and Likert-scale ratings of various features. Following the experiment, focus groups were used to discuss and confirm the aggregated findings. Important limitations of this work are that it did not gather data about real costs of the setup or assessment effort/time required and especially that it did not examine the outcomes for students in terms of student satisfaction or improved student learning. The research design described in this paper addresses these limitations (except for assessment of the improvement of student learning, as shall be seen later).

Design Experiments, Design-Based Research and Design Science Research

In this section, we consider research approaches relevant to evaluation of MFSS, which can inform our choice of how to answer the second question above – “How should go about the comparative evaluation of Marking and Feedback Support Systems to determine which are better for different purposes and what are their advantages and disadvantages?” Appropriate methods from education include Design Experiments and Design-Based Research. A similar perspective in the Information Systems field is Design Science Research (DSR). We briefly consider each of these in turn.

Design Experiments (in Education)

Design Experiments (Brown, 1992; Collins, 1992) are an approach where an educational innovation (a new educational approach or technique) is designed and then conducting an experimental study of that innovation. The study is made by making an intervention in an actual educational setting that puts the innovation to real use to see whether the anticipated benefit is achieved (or not) and what other problems may occur. Both Brown and Collins use the term “design science”, which will be discussed further in later sections, to characterise the research paradigm.

Brown (1992) emphasises the holistic nature of the intervention in that the overall situation needs to be adapted to the innovation in order to achieve the most benefit, but with a resulting difficulty in isolating the effect on the dependent variable (i.e. “training worked” or “training didn’t work”, p. 145) of the new innovation to a particular aspect within the larger holistic evaluation. On the other hand, Brown asserts that such in situ investigation has large advantages over unrealistic lab studies because the innovation may not work in the complex environment, with its numerous, mutually dependent aspects, of real educational settings. Brown favours more in-depth, idiographic investigation of such complexities over simplification and reductionism.

Collins (1992) on the other hand emphasises a program of multiple design experiments, in which multiple innovations are evaluated and compared, evaluation is as objective as possible, but with multiple evaluations in different contexts. For Collins, the aim of research should be to develop a design theory for educational innovations.

Design-Based Research (in Education)

Design-Based Research (Dede, 2004; Dede, et al., 2004; Design-Based Research Collective, 2003; Reeves, et al., 2005) builds on design experiments and “blends empirical educational research with the theory-driven design of learning environments” (Design-Based Research Collective, 2003, p. 5). Importantly, the DBR approach has particularly been adopted in accommodating research in computer and information system innovations for teaching and learning, e.g. in Dede, et al. (2004) and Schmidt, et al. (2010).

The Design-Based Research Collective (2003) proposes five characteristics of good Design-Based Research (DBR) including that theory should be an outcome, development and evaluation of innovations should be in cycles, outcomes should be accessible to practitioners, research needs to account for and fit in with the needs of authentic settings and develop refined understandings of why an innovation works or doesn’t, and finally methods in the evaluation stages must adequately document linkages between the intervention and the outcomes.

Dede (2004), like Brown (1992) focuses on the need not to simplify in order to obtain a clear control group, but to deal with complexity. Similar to Collins (1992), Dede proposes that multiple studies be conducted instead of simplistic single studies that sacrifice realistic implementation in favour of control of the confounding variables present in the complexity of real situations.

Design Science Research (in Information Systems)

This section establishes the relevance of the Design Science Research (DSR) paradigm, as discussed in the field of Information Systems, to Education generally and to MFSS evaluation specifically. It also introduces some aspects of the IS dialog on DSR that are relevant to the evaluation of MFSS (and other educational innovations).

The DSR paradigm (Hevner, et al., 2004; March & Smith, 1995) has recently received extensive attention in the Information Systems (IS) discipline. DSR is relevant to education and MFSS for three reasons. First, methodologically, DSR is closely related to design experiments and design-based research. Second, DSR emphasises the invention, design, and development of new technologies, techniques, and methods (all relevant to education), yet still achieving research rigour with respect to evidence of their utility in meeting their goals. Venable (2010) has suggested that all applied disciplines (e.g. education) could benefit from considering the DSR paradigm and the discussions about it in the IS discipline, particularly developments concerning DSR methods (including evaluation) and design theory. Third, MFSS are a kind of Information System (IS), i.e. a system that collects, processes, and stores data in order to produce information and communicate it to people.

Work on DSR methods starts from the basis of two activities: build and evaluate (Hevner, et al., 2004; March & Smith, 1995). However, more detailed models have also been developed. Nunamaker, et al. (1991) proposed a five step process (focused largely on computer-based systems as the technology): construct a conceptual framework, develop a system architecture, analyze and design the system, build the (prototype) system, and observe & evaluate the system. Venable (2006b) identified four major activities around DSR: theory building, problem diagnosis, solution technology invention, and evaluation. A particular project may move iteratively through all of these activities in any order, depending on circumstances. Of these four activities, solution technology invention is very similar to build and evaluation is the same as evaluate. Peffers, et al. (2008) developed a six stage DSR process model, including identify problem & motivate, define objectives of a solution, design & development (analogous to build), demonstration, evaluation, and communication, with iteration if needed back to earlier stages.

Evaluation in Design Science Research

Various different means for evaluation have been identified in the DSR literature. Nunamaker, et al. (1991) divided evaluation methods into two groups: observation (case studies, survey studies, and field studies) and experimentation (computer simulations, field experiments, and lab experiments). Hevner, et al. (2004) identified five classes of evaluation methods, including analytical, case study, experimental, field study, and simulation.

Building on Nunamaker, et al. (1991), Venable (2006a) distinguished between artificial evaluation and naturalistic evaluation. Naturalistic evaluation is conducted *in situ* and ideally has three characteristics (Sun & Kantor, 2006): a real system being used by real users for a real problem. To the extent that an evaluation does not meet all these characteristics, the evaluation is unrealistic and artificial. According to Venable (2006a), naturalistic evaluation is “the real ‘proof of the pudding’”. Importantly, only through naturalistic evaluation can one be certain of the workability and effectiveness of the solution in its intended organisational context(s). Approaches for naturalistic evaluation include case studies, survey studies, field studies, and action research. Approaches for artificial evaluation include computer simulations, role-playing simulations, field experiments, and lab experiments. Importantly, artificial evaluation may offer the only possibility to control for confounding variables and be certain that an improvement or achievement of the goals is due to the innovation (and only the innovation). However, artificial evaluation methods can establish efficacy, not effectiveness. Naturalistic evaluation, or what Fritz and Cleland (2003), writing in the medical field, called an effectiveness or pragmatic approach, “seek to examine the outcomes of interventions under circumstances that more closely approximate the real world” (Fritz & Cleland, 2003, p. 164).

This relative advantages of artificial and naturalistic evaluation highlights a key issue – how to make the evaluation rigorous in terms of controlling for confounding variables in the utility of the innovation (using artificial evaluation), yet also ensuring that the solution works in real, complicated situations (using naturalistic evaluation). One possibility as suggested by Brown (1992) and Collins (1992) is to have multiple evaluations (generally these would be naturalistic in the terms of Venable (2006a)). Another possibility is to use both artificial and naturalistic evaluation methods, most likely in multiple evaluations. In the next section, we consider how these issues play out for evaluation of MFSS.

Issues in Research Evaluating Marking and Feedback Support Systems

Ideally, the comparative evaluation of MFSS would achieve rigour both in the sense of controlling for confounding variables (ensuring that benefit achieved was due to the particular MFSS and only the particular MFSS being evaluated), as well as in the sense of realistic application that surfaces potential problems in using the MFSS in real contexts, with real users, and using real MFSS (not toy prototypes or simulations).

Confounding Variables in Evaluating MFSS

Evaluating MFSS has a number of potential confounding variables, which might bias or otherwise systematically affect the evaluation results. Potential confounding variables include the following:

- The student – differences in perceptions and expectations
- The unit – differences in level, topic, and size
- The assignment or assessment item – differences in type, length, characteristics to be assessed
- The marking rubric used – differences in length, detail, quality, and suggested or required feedback text
- The marker – differences in motivation, experience and knowledge in the domain and experience marking in the domain and the type of assessment
- Learning effects – both learning how to use the tool and learning about marking the assessment item gained during performance of the marking task

Ethical Issues

An important ethical issue is that it is inappropriate to assess some students solely with one marking and feedback method and other students solely with a different marking and feedback method, because some students might thereby be disadvantaged in their learning and/or their results (i.e. some students might receive poorer quality feedback or lower marks simply because of having their work assessed using a different MFSS).

Combining the ethical and confounding variables issues, a way needs to be found that students are all marked the same way, yet different MFSS are employed and evaluated, and still controlling for the confounding variables. The next section describes a research design to achieve this.

Research Design for Evaluating Marking and Feedback Support Systems

We are currently engaged in a research project to comparatively evaluate four different MFSS (ABBA, tsAAM, EPSS, and Blackboard Gradebook marking, feedback and rubrics features). Of the many available MFSS, these were selected primarily for our interest and convenience, as ABBA, tsAAM, and EPSS were developed and are under continued development in our research group and our university is converting to Blackboard 9.1 with its new marking rubric feature. Other MFSS could (and should) be evaluated and the same method described here can be used to do so.

Goals of the Evaluation

There are seven characteristics to be evaluated in the research: efficiency in terms of (1) tool installation, (2) set up times and (3) resources, efficiency in terms of (4) marking effort, effectiveness in terms of (5) marking reliability and (6) accuracy, and effectiveness in terms of (7) student satisfaction with feedback provided through use of the tool. Evaluations should be in both absolute terms and in comparison to other tools and traditional, manual approaches. Additionally, information will be sought about the overall marker and student experience and suggestions for improvements.

The independent variable in the experimental design is the MFSS used (or traditional manual, non-MFSS supported, marking and feedback used as a control) and the dependent variables will be the seven efficiency and effectiveness characteristics described above.

Research Design

The research design we have come up with to meet the above needs is to conduct a Design Experiment as part of a program of Design-Based Education Research – developing new and improved MFSS. The research design draws on the Design Science Research evaluation approaches in the IS literature. The evaluation design is a combination of a field experiment (artificial evaluation) with a field study (naturalistic evaluation) (Venable, 2006a). The MFSS will be employed for marking in actual units by actual markers, using actual MFSS (some commercial, some research prototypes) in line with the three aspects of naturalistic evaluations proposed by Sun and Kantor (2006).

In order to account for differences in unit (level, topic, and size), assessment item (type, length, characteristics to be assessed), and the marking rubrics used (length, detail, quality, and suggested or required feedback text), there need to be multiple empirical trials of the MFSS to be evaluated across a variety of units, assessment items, and marking rubrics. Units to be covered in the experiments will include introductory and advanced units (undergraduate and postgraduate), small and large units with smaller or larger numbers of markers, as well as ‘softer’ and ‘harder’ topics. For practical reasons, in this particular project we plan to cover seven different units over two semesters, but further evaluations will likely be conducted later in other projects as the research proceeds. Also for practical reasons, all units are in the field of information systems, but other subject areas can be addressed in the future. Furthermore, control for differences in characteristics of units/topics, assessment items, markers and rubrics will be obtained by using multiple MFSS tools within a single unit, assessment item and its rubric, and by each marker.

To control for differences in students, assessment items will be marked using all MFSS being evaluated in a particular trial. Each student will receive marks and feedback generated by each of the MFSS being evaluated on that unit/assessment item/marking rubric. This will also address the ethical issue that no student should be disadvantaged by being marked by different tools (on that particular unit and assessment item). It would be inappropriate to simply divide each unit into a control group (manual marking, no MFSS), and one or more treatment groups (e.g. a tsAAM-supported group and/or an EPSS-supported group). Instead, all students’ assessment items will be assessed using all the tools being evaluated in a particular trial.

To control for differences in markers, each marker will use all of the different MFSS being used in a particular trial (as well as a manual control marking method) on that unit/assessment item/marking rubric. However, in order to account for learning effects, each different marker on a particular field experiment will use the MFSS (and manual marking method) in a different order.

To summarise, instead of simple division into a control group and one or more treatment groups, each student’s assessment item will be assessed by more than one marker, each of which will use a different MFSS tool (or no tool) for that student. Doing so incurs a cost of multiple marking of each assessment item, but this cannot be avoided for the ethical reason noted above. For example, student 1 might be assessed by marker A using MFSS X, by marker B using MFSS Y, and by marker C using no MFSS. In addition to avoiding ethical issues, another benefit of using this research design is that students will receive feedback using each of the marking approaches used and can evaluate both absolutely and comparatively all of the approaches. Furthermore, such a research design provides a control over student characteristics.

However, one cannot have a single marker using only one tool or the marker becomes a confounding variable. Therefore, each marker will assess some students using no tool and (an) equal (or nearly equal) number(s) of students using the tool(s) being assessed. For example, marker A might mark students 1-10 using no MFSS (traditional manual), then students 11-20 using MFSS X, and finally students 21-30 using MFSS Y, while marker B would first mark students 1-10 using MFSS X, then students 11-20 using MFSS Y, and finally students 21-30 using no MFSS (manual), and marker C would mark students 1-10 using MFSS Y, students 11-20 using no MFSS, and students 21-30 using MFSS X. Thus, each student has their assessment item marked three times, one using each assessment approach/ MFSS. Feedback can be provided using all three approaches and no student is disadvantaged by the marking method (or by the marker for that matter). This provides a further control over the marker as a potential confounding variable. Table 1 below illustrates the above example of the research design, with marking progressing from left to right. To evaluate three MFSS plus a manual control would require 4 markers and 4 groups of students similarly distributed onto different MFSS and marker combinations.

Table 1: Example distribution of MFSS across student and marker combinations

	Students 1-10	Students 11-20	Students 21-30
Marker A	No MFSS (manual)	MFSS X	MFSS Y
Marker B	MFSS X	MFSS Y	No MFSS (manual)
Marker C	MFSS Y	No MFSS (manual)	MFSS X

Following completion of the marking by all markers, the markers together with one or more of the researchers should meet to review, discuss, and moderate the marks to determine the correct mark for each student's assessment item. Differences between the correct (moderated) mark and each marker's prior individual mark will be used to determine the accuracy of the mark, which will then be related to the tool used (but averaged against marker differences, on average across tools. After marking and moderation, all of the feedback and marks developed using all the MFSS and a manual approach used will be provided to the students together with the moderated mark, which will serve as their final mark for the assessment. This will further address ethical issues and the fairness of the marking.

Following return of the marks and feedback, the students will be surveyed to obtain their satisfaction with the feedback, their preferences and opinions about the different feedback provided using the different methods, and their suggestions for improvement.

Prior to marking, data will be gathered on MFSS tools used and their features, MFSS installation and set-up times, and the traditional (manual) methods used (which may vary considerably).

During marking, data on marking times and marks awarded (numeric) will be gathered from markers. Toward this end, we have designed a spreadsheet to record starting and stopping times for marking each assessment item and calculating the total time. During and following marking, markers are asked to record any comments and suggestions for improvement and to rate their satisfaction, ease of use, ease of learning, etc. of the MFSS (or manual method). These are recorded on the same spreadsheet.

The research design mixes a field experiment and a field study. The field experiment part controls the confounding variables and collects quantitative data on specific measures to rigorously test whether the benefits are achieved. The field study part collects qualitative data on opinions, experiences, and suggestions together with data about specific features, units characteristics, assessment item characteristics, and marking schemes, which may also be correlated with outcomes and assess the efficacy of the method in a real, complex environment. Furthermore, by keeping copies of the assessments, we also will have rich data on the feedback provided, which can be further analysed at a later date.

Conclusion

In this paper, we have described issues in the evaluation of MFSS and a research design combining an artificial (field experiment) and naturalistic (field study) evaluation approaches. The design provides both controlled and comparative evaluation of the different MFSS being evaluated as well as an evaluation of the overall efficacy of real MFSS, used by real users, for a real task (actual assessment of an actual assignment in an actual unit). Overall, the results of a study following this design should provide systematic, empirical evidence of the relative and overall utility of the MFSSs evaluated and their features.

Naturally, the research design does have limitations and disadvantages. One disadvantage is the cost of having multiple marking of assessments. Another is that it may be difficult to obtain sufficient markers who are both willing and who have sufficient knowledge of the domain being assessed to adequately carry out a similar assessment study. A key limitation is that the method does not allow comparison of the MFSSs' effects on learning achieved using the different feedback from the different methods, since all students receive feedback using all methods. Providing students with different feedback using only one of the MFSS being evaluated in order to attempt to measure differences in learning using different feedback methods would require a decision giving higher priority to the resulting knowledge as to which feedback methods are better over the disadvantage given to some students in their learning due to inferior feedback and feedback methods. The chosen research method opted not to make such a decision. Nonetheless, the method we have proposed does allow for measuring the relative *perceived* value of the different feedback approaches, which we believe is an acceptable surrogate for actual learning.

Future research could refine the approach taken and also apply it more systematically to the many other MFSS which are not being evaluated in our study, as well as for other course/unit topics and other forms of assessment. Development of a coherent, cumulative body of knowledge evaluating different MFSS and their features would be useful both for educators who are considering what MFSS to acquire and deploy, as well as for MFSS designers and researchers in continuing to develop and improve MFSS in a program of design-based education or design science research.

Acknowledgements

This research is supported by a grant from the Australian Council of Deans of Information Communication Technology (ACDICT) Learning and Teaching Academy (ALTA) and the Curtin University School of Information Systems.

References

Audacity. (2011). Audacity: the free cross-platform sound editor. 2011, from http://audacity.sourceforge.net/

Bailey, R., & Garner, M. (2010). Is the feedback in higher education assessment worth the paper it is written on? Teachers' reflections on their practices. *Teaching in Higher Education, 15*(2), 187-198.

Berger, D. (2011). *Supporting Tool for Moderation in the Grading Process of Summative Assessments: Design and Prototype of a Software Tool for Moderation and Assessment with Variable Rubrics.* Graz University of Technology, Graz, Austria.

Brown, A. L. (1992). Design Experiments: Theoretical and Methodological Challenges in Creating Complex Interventions in Classroom Settings. *The Journal of the Learning Sciences, 2*(2), 141-178.

Burrows, S., & Shortis, M. (2011). An Evaluation of Semi-automated, Collaborative Marking and Feedback Systems: Academic staff perspectives. *Australasian Journal of Educational Technology, 27*(7), 1135-1154.

Campbell, A. (2008). *Performance Enhancement of the Task Assessment Process through the Application of an Electronic Performance Support System.* Edith Cowan University, Joondalup, WA, Australia.

Colbran, S. (2011a). *Evaluation of the usefulness of self-assessment, peer assessment and academic feedback mechanisms.* Paper presented at the Australian Technology Network Assessment Conference 2011: Meeting the Challenges, Perth, Western Australia.

Colbran, S. (2011b). *ReMarksPDF: A new approach to moderation involving multiple assessors.* Paper presented at the Australian Technology Network Assessment Conference 2011: Meeting the Challenges, Perth, Western Australia.

Cole, J. (2005). Workshops *Using Moodle: Teaching with the popular open source course management system*. Sebastopol, CA, USA: O-Reilly Media.

Collins, A. (1992). *Toward a design science of education*. Paper presented at the New directions in educational technology.

Dede, C. (2004). If Design-Based Research is the Answer, What is the Question? A Commentary on Collins, Joseph, and Bielaczyc; diSessa and Cobb; and Fishman, Marx, Blumenthal, Krajcik, and Soloway. *Journal of the Learning Sciences, 13* (1 Special Issue on Design-Based Research), 105-114.

Dede, C., Nelson, B., Ketelhut, D. J., Clarke, J., & Bowman, C. (2004). *Design-Based Research Strategies for Studying Situated Learning in a Multi-user Virtual Environment.* Paper presented at the International Conference on Learning Sciences.

Design-Based Research Collective, T. (2003). Design-Based Research: An Emerging Paradigm for Educational Inquiry. *Educational Researcher, 32*(1), 5–8.

Ferns, S. (2011). *Allocating academic workload for student consultation assessment and feedback.* Paper presented at the Australian Technology Network Assessment Conference 2011: Meeting the Challenges, Perth, Western Australia.

Fritz, J. M., & Cleland, J. (2003). Effectiveness Versus Efficacy: More Than a Debate Over Language. *Journal of Orthopaedic & Sports Physical Therapy, 33*(4), 163-165.

Gibson, W., Robinson, L., & Yorke, J. (2011). *Work in progress: The Short Answer Feedback System (SAFS).* Paper presented at the Australian Technology Network Assessment Conference 2011: Meeting the Challenges, Perth, Western Australia.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly, 28*(1), 75-105.

Hogan, J., Carlson, J., & Dua, J. (2002). Stressors and stress reactions among university personnel. *International Journal of Stress Management, 9*(4), 289-309.

March, S., & Smith, G. (1995). Design and natural science research on information technology. *Decision Support Systems, 15*(4), 251-266.

Nicol, D. J. (2011). Technology-supported assessment. from http://wiki.alt.ac.uk/index.php/Technology-supported_assessment

Nicol, D. J., & Macfarlane-Dick, J. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199-218.

Nicol, D. J., & Milligan, C. (2006). Rethinking technology-supported assessment in terms of the seven principles of good feedback practice. In C. Bryan & K. Clegg (Eds.), *Innovative Assessment in Higher Education*. London: Routledge, Taylor and Francis Group.

Nunamaker, J. F. J., Chen, M., & Purdin, T. D. M. (1991). Systems Development in Information Systems Research. *Journal of Management Information Systems, 7*(3), 89-106.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2008). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45-77.

Reeves, T. C., Herrington, J., & Oliver, R. (2005). Design Research: A Socially Responsible Approach to Instructional Technology Research in Higher Education. *Journal of Computing in Higher Education*, *16*(2), 97-116.

Schmidt, M., Galyen, K., Laffey, J., Ding, N., & Wang, X. (2010). Leveraging Open Source Software and Design Based Research Principles for Development of a 3D Virtual Learning Environment. *SIGCAS Computers and Society, 40*(4), 45-53.

Shortis, M., & Burrows, S. (2009). *A review of the status of online, semi-automated marking and feedback systems.* Paper presented at the ATN Assessment Conference, Melbourne, VIC, Australia.

Skeele, R. W., Carr, V. B., Martinelli, J., & Sardone, N. B. (2007). *Innovation in e-assessment: Exploring a multidimensional tool.* Paper presented at the 2007 World Conference on E-Learning i Corporate, Government, Healthcare, and Higher Education, Montreal, Quebec, Canada.

Sun, Y., & Kantor, P. B. (2006). Cross-Evaluation: A new model for information system evaluation. *Journal of the American Society for Information Science and Technology, 57*(5), 614-628.

Thompson, D., Lawson, R., & Boud, D. (2011). *(2011) Efficient pre-assessment intervention to enhance student judgements using ReView*. Paper presented at the Australian Technology Network Assessment Conference 2011: Meeting the Challenges, Perth, Western Australia.

Venable, J. R. (2006a). A Framework for Design Science Research Activities. *Proceedings of the 2006 Information Resource Management Association Conference (CD ROM), Washington, DC, USA, 21-24 May 2006.*

Venable, J. R. (2006b). *The Role of Theory and Theorising in Design Science Research*. Paper presented at the DESRIST 2006.

Venable, J. R. (2010). *Information Systems Design Science Research as a Reference Discipline for Other Business Disciplines*. Paper presented at the International Academy of Business and Public Administration Disciplines Conference

Weinberger, A. (2011). *Semi-Automatic Essay Assessment based on a flexible Rubric.* Graz University of Technology, Graz, Austria.

Weinberger, A., Dreher, H., Al-Smadi, M., & Guetl, C. (2011). *Analytical Assessment Rubrics to Facilitate Semi-Automated Essay Grading and Feedback Provision.* Paper presented at the Australian Technology Network Assessment Conference 2011: Meeting the Challenges, Perth, Western Australia.

Wren, J., Campbell, A., Heyworth, J., & Bartlett, R. (2011). *Improving marking of live performances involving multiple markers assessing different aspects.* Paper presented at the Developing student skills for the next decade: Proceedings of the 20th Teaching and Learning Forum, Perth, Western Australia.

Wren, J., Campbell, A., Heyworth, J., & Lovering, C. (2011). *Improving assessment outcomes through the application of innovative digital technologies.* Paper presented at the Australian Technology Network Assessment Conference 2011: Meeting the Challenges, Perth, Western Australia.

Young, S. (2011). *iFeedback - a new tool for grading and providing timely, detailed individual feedback to students.* Paper presented at the Australian Technology Network Assessment Conference 2011: Meeting the Challenges, Perth, Western Australia.