

# Automatic Grading of Spreadsheet and Database Skills

**Zlatko J. Kovačić and John Steven Green**  
**Open Polytechnic, Wellington, New Zealand**

[Zlatko.Kovacic@openpolytechnic.ac.nz](mailto:Zlatko.Kovacic@openpolytechnic.ac.nz)  
[John.Green@openpolytechnic.ac.nz](mailto:John.Green@openpolytechnic.ac.nz)

## Executive Summary

Growing enrollment in distance education has increased student-to-lecturer ratios and, therefore, increased the workload of the lecturer. This growing enrollment has resulted in mounting efforts to develop automatic grading systems in an effort to reduce this workload. While research in the design and development of automatic grading systems has a long history in computer education, only a few attempts have been made to automatically assess spreadsheet and database skills.

This paper has three purposes: (1) to describe the design of an assessment in the *Information Systems* course at the Open Polytechnic to assess students' spreadsheet and database skills, (2) to describe the development of an automatic grading system to assess spreadsheet and database skills, and (3) to compare automatic with manual marking to determine if automatic grading system is a feasible method of reducing workload.

The automatic grading system we developed uses Excel's user-defined functions to automatically check whether a feature or a function has been used. Since the outcomes from user-defined functions are scrambled, students verify their own answers by entering the results from these functions into an online quiz. As a result, there is no need for the lecturer to download, open, and check the actual software application. The system recognizes correct answers from these scrambled inputs and allocates marks. This system is integrated into the Moodle learning management platform and linked to the students' academic record database.

The main difference between the automated grading system for the assessment of spreadsheet and database skills described in this paper and existing systems is that the latter systems require the actual software application to be submitted for marking. The system described in this paper does not require markers to handle the application. Instead, it automatically checks the application while students are working on it, but grading is not performed until students answer specific quiz questions.

Practical experience with the automatic grading system has shown that the system significantly decreases turnaround time for the grading of assignments, while providing instant feedback to students on the correctness of their answers. At the same time, the system reduces the workload of the lecturer, freeing lecturers from administration and the time-consuming tasks of checking individual aspects of the spreadsheet and database applications. This allows them to allocate time to student support and other more crea-

---

Material published as part of this publication, either on-line or in print, is copyrighted by the Informing Science Institute. Permission to make digital or paper copy of part or all of these works for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial advantage AND that copies 1) bear this notice in full and 2) give the full citation on the first page. It is permissible to abstract these works so long as credit is given. To copy in all other cases or to republish or to post on a server or to redistribute to lists requires specific permission and payment of a fee. Contact [Publisher@InformingScience.org](mailto:Publisher@InformingScience.org) to request redistribution permission.

tive activities. In addition, the automatic grading system allows for a much finer probing of individual aspects of the spreadsheet and database applications, with no additional work required by student or lecturer.

The methods of marking were evaluated to address the main research question of whether there were significant differences between a human and an automated grading system. A comparison between the methods of marking (human and the automatic grader) based on data from 11 trimesters indicated no significant difference in the average marks and mark distributions in the case of the spreadsheet application. The comparison also showed that although the difference in the average marks in the case of the database application was significant, it did not mean that the effect of the method of marking was meaningful or important, as illustrated with the effect size. Nonetheless, monitoring of the automatic grader results is recommended.

**Keywords:** automatic grading, e-assessment, database skills, distance education, software skills, spreadsheet skills.

## Introduction

Growing student numbers in the 40-plus age range and increased student-to-lecturer ratios have led many academics to consider alternative forms of assessment in tertiary education, such as automatic assessment systems and, within these systems, automatic grading systems. A shift in the role of the lecturer has also triggered work on the further development of these systems. Malmi, Korhonen, and Saikkonen (2002) point out that advances in automatic assessment systems are the result of the changing roles of lecturers in the new learning paradigm; lecturers become facilitators, creating the learning environment, providing guidelines, commenting on students' work, giving feedback, and so on. Advancements in technology have also enabled the development of automatic assessment, as well as generally improving the student learning experience through, for example, learning management systems such as Moodle and Blackboard.

The broadest definition of e-assessment is given in Malmi et al. (2002) and adopted in this paper. Malmi et al. (2002) defines e-assessment as "the end-to-end electronic assessment processes where ICT (information and communication technology) is used for presentation of assessment activity, and the recording of responses," which "includes the end-to-end assessment process from the perspective of learners, tutors, learning establishments, awarding bodies and regulators, and the general public" (p. 6). They also make a distinction between computer-based assessment and computer-assisted assessment. The former refers to assessments delivered and marked by computers, while the latter relies on computers only in part.

There are differences between assessment and grading. In this article, however, the terms 'computer-based assessment', 'automated grading systems', and 'automatic marking systems' have been used interchangeably, as they have been in most of the studies listed in the references. The main reason for doing this is that our automatic grading system does not simply evaluate the learning and performance of students (the core elements of the definition of grading); it also provides detailed feedback for every question and suggests possible reasons for any incorrect solutions and answers. Our automatic grading system, therefore, also helps to improve student learning, which is the goal of assessment. In the following literature review, we kept the original terms authors used in their studies to describe their systems.

Computer education research has focused on automatic grading systems for almost 50 years (see Whitelock & Brasher (2007) for an overview of three generations of automatic assessment systems). However, there have only been a few attempts to build a grading system for the automatic assessment or the grading of Office skills, that is, word-processing, spreadsheet, and database skills (Hill, 2003, 2004; Koike, Akama, Chiba, Ishikawa, & Miura, 2005; Koike, Akama, Morita, & Mura, 2006; Zhenming, Liang, & Guohua, 2003). The major expectation in this research was

that the introduction of automatic grading systems would significantly decrease turnaround time for assessments and, at the same time, reduce the workload of the lecturer by removing the need to check the software application students developed. However, building an automatic marking system is not a quick nor a cheap option, as Koike et al. (2006) emphasized.

### ***Research Goals, Questions, and Hypothesis***

The goals of our research were to build an automatic grading system to assess advanced spreadsheet and database skills and to compare automatic with manual marking. More specifically, our first goal was to convert the practical parts of the *Information Systems* course assignments (building spreadsheet and database applications) into two computer-marked assignments quizzes in Moodle.

The aims of this conversion were twofold:

1. to design and develop tasks for the assessment of spreadsheet and database skills by distance, without opening and checking the software applications,
2. to automate the marking of practical software applications.

The conversion of assignments into the Moodle quiz format was undertaken with the expectation that it would be beneficial for both students and lecturers. Students would get instant feedback, and their work would be marked objectively. The workload of lecturers would be reduced, allowing them to allocate more time to teaching, the updating of course material, and student support.

Our second goal was an evaluation of whether our automatic grading system allocated marks as a human marker would. Given our research interest, the main research question asked in this paper is: do significant differences exist between marks allocated by humans and those allocated by an automatic grading system? It is hypothesized that there is no significant difference.

Before addressing these questions, a brief overview of related work on platforms for the automatic assessment (i.e., grading or marking) of Office skills is presented. In the section ‘E-assessment: Design Issues and Challenges,’ we discuss our approach to the design of a grading system for assessing spreadsheet and database skills. In the section ‘E-assessment: Implementation Issues,’ we give an overview of our automatic grader system and its components, briefly describing the issues we had with its implementation. The next section gives an evaluation of the methods of marking. The last section outlines our conclusions and future work.

## **Literature Review**

### ***Drivers for the use of Computer-based Assessments***

According to Tshibalo (2007), academic workload is increased in higher education and online assessments may help reduce this workload by helping lecturers manage the large volume of marking and assessment-related administration. Swithenby (2006) also listed several drivers for the increasing use of e-assessments. He included both economic (that is, demand for portable qualifications and a cost-effective means of testing) and pedagogical drivers, and the most important pedagogical driver is that of rapid feedback in the form of both marks and comments, since they have the potential for the immediate shaping of learning.

Whitelock and Brasher (2006) also cited several drivers for the adoption of e-assessment, including perceived increases in student retention and the enhanced quality of the feedback the student receives. They also cited the flexibility of e-assessment for distance learning, the strategies it provides to lecturers for coping with large student numbers, the inherent objectivity in marking, and how it makes more effective use of a learning management system.

There are, however, some potential weaknesses and barriers to further increases in the use of computer-based assessments. As Swithenby (2006) pointed out, there is still some “cultural antipathy to computer-based assessment.” According to this view, the use of computer-based assessments is impersonal, suitable only for the assessment of tasks at a low cognitive level or those that require closed responses.

Some academics feel that computer-based assessment cannot test high order skills, such as synthesis and analysis (Musham, 2004). We feel that the acceptance and effectiveness of multiple-choice tests depends heavily on their design. With properly designed computer-based assessments, even the higher ranked objectives of Bloom’s taxonomy (such as application and analysis) can be reached. As the development of computer-based assessment requires academic staff time and a high initial investment, Whitelock and Brasher (2006) see the need for academic staff development time as one of the major barriers for the wider use of computer-based assessment. Similarly, Buzzetto-More and Alade (2006) concluded that “assessment programs are time consuming efforts that require planning and foresight. Effectiveness is dependent on institutional and administrative support as well as a long range plan for sustainability that includes technological preparedness” (p. 266).

### ***Use of Automated Grading Systems for Office Skills***

As previously stated, there have been only a few attempts to develop automated grading systems for assessing Office skills. Waldman and Ulema (2008) briefly described three different methods for implementing a custom grading program for Excel. The first option (which we have partially adopted in the design of our automated grading system) requires writing a grading program in Excel using Microsoft’s Visual Basic for Applications (VBA). The second option requires writing a program that interfaces with a running copy of Excel (Hill’s MEAGER, described later, is an example of such a program). The third method (adopted by Waldman & Ulema, 2008) “is to write a program that simply reads the student’s Excel data and extracts the data for the items that need to be graded” (p. 77).

The third method requires an Excel file to be submitted for marking. As the number of submissions increases both in print and online (as an attachment to an email or uploaded to the learning platform), these submissions slow down the marking process (see Kline & Janicki, 2003). Each student receives a file that contains an embedded identifier which addresses the issue of plagiarism. Once the students return the Excel file, the marker runs the automated grader that compares the work of the students with the grading templates. These grading templates are text files that list the elements of the Excel file to be graded, the acceptable correct answers, mark allocation, and feedback comments for incorrect answers.

The current version of the system developed by Waldman and Ulema (2008) permits only two correct answers to be built in to each spreadsheet cell. This feature decreases the benefits of automated grading when more than two variations are expected in some questions. In such cases, human markers are required. Waldman and Ulema have included the following items in the automatic grading system for spreadsheet applications: name and existence of the worksheet; the value of a particular cell, formula, and range name (cells associated with a range name); the use of range names; and checking the validity of formulas and graphs (existence, type, legend, axis label and data range).

Koike et al. (2005) built an automatic marking system for intermediate Office skills. They wrote programs in VB.NET, and these programs automatically mark MS Word and Excel files. For example, the program for MS Word checks page settings, paragraphs, indents, figures, tables, fonts, colors, texts, and so on. A student may download the program and run it on their files for formative assessments.

One of the most frequently cited automated grading systems is developed by Hill (2003, 2004). Hill developed two automatic grading systems for Microsoft Office applications, namely Excel (Microsoft Excel Automated Grader - MEAGER) and Access (Microsoft Access DataBase Automated Grading System - MADBAGS). These systems were constructed to be used by lecturers.

MEAGER is a generic grading program that compares the Excel workbook submitted by the student against the lecturer-supplied solution workbook and allocates marks for each submitted file. MEAGER also includes reports and tools for the detection of plagiarism. Once students submit their Excel workbooks, MEAGER compares their files with the correct solution, that is, a grading template (an Excel file prepared by the lecturer).

MEAGER is a Microsoft Access application. It extracts attributes from both the student's and lecturer's workbook and stores them in separate tables for each attribute. MEAGER then compares each attribute in the student's and lecturer's tables, identifies differences, and records them in an errors table. The lecturer can use this errors table to identify which particular spreadsheet skills the student is lacking. MEAGER marks and embeds a grade report in the student's workbook. The updated student workbook can then be sent back to the student.

MEAGER grades various worksheet attributes: text labels, numbers, formulas, fonts, cell alignments, number formats, merged ranges, worksheet names, chart types, location and source data, and so on. It does not grade conditional formatting, embedded objects or drawing objects, and controls. MEAGER marks in greater detail than human graders and can detect errors a human grader cannot while reducing the time required to grade assignments. However, there are some weaknesses in MEAGER. For instance, the lecturer still has to download and upload the student workbook when using MEAGER. Also, the automated grader is, in some cases, too rigid to accept an alternative but correct answer if it does not match the specimen answer exactly.

Marking formulas in Excel can be a challenging task (see Hill, 2004). This is because, in some cases, the same result can be achieved in different ways by using different formulas/functions. MEAGER approached this challenge by applying an interpreter for the symbolic manipulation of mathematical expressions (like Maple or Mathematica). An assignment should be written very strictly, with no room for interpretation or the use of an equivalent formula. Otherwise MEAGER would be unable to differentiate between two equivalent formulas, even though they answer the same question correctly. In this case, a lecturer's intervention is required.

We have adopted a different approach to this problem by explaining the tasks and requirements to students clearly and in detail and by specifying in the assignments the particular Excel/Access function or specific formula to be used. So when a function or formula is entered in the online quiz, students know from the assignments which particular function or formula they should use. It might be considered that this tactic is too restrictive, that it reduces the freedom of students to use their own modeling knowledge to solve particular tasks. At the same time, however, this approach ensures that all the software functionalities and features covered in this introductory course are really fully grasped by students and assessed by lecturers. A more advanced course would probably require a different automated grading system, one that would have more flexibility in handling students' solutions and where the focus is more on the optimal solution and modeling approach rather than on a set of required Office skills. Our automatic grading system, however, also accepts an equivalent formula or an alternative solution. This is hard-coded in the quiz. In other words, our system would recognize an alternative solution and mark it as correct.

MADBAGS is an Access Grader that works similarly to MEAGER. It embeds an "Errors" table in the student's database. MADBAGS compares a correct version of an Access database with a student version and records the errors in an Access table. It grades the following database attributes: table data, table structure, field attributes, indexes, primary keys, relations, SELECT queries, total queries, DELETE and UPDATE queries. MADBAGS does not grade data access pages,

macros or modules, and can detect whether a form or report is absent although it does not grade the form or report in detail.

## **E-assessment: Design Issues and Challenges**

The main motivation for the creation of a new system for automatic grading was that all the existing systems reviewed still require the actual software application (that is, Excel workbook and Access database) to be submitted for marking. The automated grading system is then executed in order to compare the student's solution with the "correct" solution and to provide relevant feedback. We wanted to avoid the actual handling of software applications, because marking large cohorts with existing automated grading systems would still be a time-consuming process; it requires downloading a submitted application from the learning management platform, passing it through the automated grading system, and then either returning the report back to student or returning the actual application. Our system was designed with the idea of eliminating all these steps by automated marking while the application is still with students.

The Open Polytechnic uses Moodle integrated with the student database. This enables electronic submissions for essay-like assignments (both as an attachment and as questions in quizzes), online marking and, in the case of Moodle quizzes, storing marks directly (without human intervention) in the student record. This was an additional motivation for us to design and develop online assignments that could be automatically marked.

### ***Course and Assignments Descriptions***

The *Information Systems* course is an introductory Level 5 course for the Information Systems and Technology major at the Open Polytechnic of New Zealand. The purpose of Level 5 on the ten-level New Zealand Qualification Framework is to qualify individuals with theoretical and/or technical knowledge and skills within a specific field of work or study. The course has an average of 100 students per trimester with a multicultural student body of approximately two-thirds women and one-third men. Most students are between the ages of 30 and 50, and the majority of students are employees or self-employed. Students study the course in Moodle through distance learning. Students need to successfully complete a pre-entry test before enrolling on this course. This test ensures they have basic skills in Excel and Access. However, there is still a significant variation in their spreadsheet and database skills.

As part of student assessment, students are required to build spreadsheet and database applications. They also need to create a presentation and write reports to a client in a case study that describes a small, local New Zealand business. The Excel and Access applications are developed around this case study. In this course, only spreadsheet and database skills are graded.

The *Information Systems* course has two assignments:

- Project 1:       Task 1: Build the "Tiki" sales tracking system
- Task 2: Business information system research
- Project 2:       Task 1: Build the "Tiki" database system
- Task 2: PowerPoint presentation
- Task 3: SDLC and telecommunications technologies

Previously, students were asked to submit the actual software applications, the actual PowerPoint presentation, and written reports using Word templates. We designed these practical tasks to cover the "application" and "analysis" objectives of Bloom's taxonomy. Students were also asked to use their applications to answer a few questions related to different scenarios that were relevant to the business from the case study.

## ***The Design of Assignment Tasks***

All the automatic grading systems for spreadsheet and database skills that we reviewed (Hill, 2003, 2004; Kline & Janicki, 2003; Koike et al. 2005; Russell & Cumming, 2005; Waldman & Ulema, 2008) require students to submit the actual application. We designed an assignment that ensured particular features of the Office software were used, but without asking students to submit the actual application. This was the most challenging part of the conversion process. Since the dataset in the case study assigned to students was small in size, students might try to find solutions manually. Therefore, the assignment tasks had to be designed in such a way as to prevent students from finding solutions manually.

We will use the spreadsheet application and pivot table the students were asked to create to illustrate our approach. After they created a pivot table from the raw data, we asked them to change the field settings in the pivot table to summarize the field by standard deviation. While most of them would be able to manually create a two-way pivot table, even those who completed a statistical analysis course would not be easily able to manually calculate a standard deviation and change the display of data in the pivot table.

There is always a possibility for students to short-circuit an assignment task by calculating the final result manually. In such cases, a human may need to review the quiz. So we asked the students to include a few screenshots of the most important steps into their essay questions, to ensure that the tasks were completed as required. However, these essay questions were marked by a human and not by an automated essay grading system.

## **E-assessment: Implementation Issues**

We partially adopted the first implementation option described in Waldman and Ulema (2008) – writing a grading program in Excel using Microsoft’s Visual Basic for Applications (VBA). We used user-defined functions and Excel objects to check if the attributes in a student workbook were according to the requirements and if a particular function/feature was used.

Students collected information from their workbook on discrepancies between the requirements of the task and their solution. The outcome was stored in a separate **Answers** worksheet. The results of the checking procedure were “scrambled” using the Excel random function. This function generated a random number from one of the subset of numbers (“correct” and “incorrect” set of numbers), depending on whether the answer was correct or not, or if the attribute was according to our requirements.

Students were asked to enter these numbers from the **Answers** worksheet into the online quiz. The online quiz recognizes whether the number belonged to the “correct” or “incorrect” set of numbers and allocated marks accordingly. Similarly, Russell and Cumming (2005) built in the so-called “hidden database check” to ensure students did not complete the SQL query manually. They also penalized students for the inefficient use of SQL statements (that is, the statement could be correct, but had a query twice as long as the sample solution query) or to give a quality measure to the SQL statement.

Specific spreadsheet skills, such as conditional formatting and goal seek, were assessed in different parts of the workbook. Utilizing the same group of cells to assess multiple skills in using Excel complicates automatic grading because a subsequent activity overwrites the result of a former activity – as explained in Waldman and Ulema (2008). Therefore, each of such skills should be assessed in a separate worksheet.

## **Description of the Automatic Grader**

Our grading system has two parts: a Moodle quiz based mostly on Cloze-type questions where students get immediate feedback on their submission, and a set of essay-type questions marked by a human marker.

Similarly Amelung, Piotrowski, and Rösner (2006) developed a three-step approach: (1) an electronic multiple-choice test, (2) electronic submission for essay-like assignments and (3) automatic checking and marking of programming assignments with immediate feedback. Implementation of their system was carried out in Plone, another open-source content management system.

For our practical quizzes in both the *Information Systems* course assignments, these are marked automatically without intervention from lecturers. As earlier mentioned, students do not submit the applications. Instead, using the application, they answer questions and insert the requested information into the quiz. The rest of both assignments are essay questions (marked by a human marker).

## **Preparing a spreadsheet for use**

For Project 1 Task 1 (a spreadsheet task), students download some Word documents with data for this task. They also download an Excel file, which contains four worksheets. The students then edit the Excel file as required by the assignment. The Excel template helps us in the clarification of the assignment and in controlling the format of worksheets. This simplifies the marking process and focuses the students' attention on the Excel features rather than on formatting issues.

Detailed instructions are provided on how to prepare the initial workbook. These activities ensure that all the user-defined functions that are required for checking the student workbook attributes are functioning. We ask the students to complete an example quiz to ensure that they will not have technical problems with the quiz. The example quiz forms part of an incremental assessment strategy. Russell and Cumming (2005) also used an incremental assessment strategy by splitting their assessment into four assignments. When the student had completed 75% of a tutorial group, the related assignment was made available.

The example quiz has an **Answers** sheet similar to the **Answers** sheet in our Project 1 quiz. This forces students to prepare the spreadsheet for use before attempting the quiz for Project 1. When they score 100%, the password for Project 1 Quiz is delivered to them. The example quiz does not contribute towards the student's grade, thus students have the opportunity to take the example quiz until they are comfortable navigating the Moodle quiz. Lecturers addressed any technical problems and observed the students' reactions to the quiz.

The **Answers** sheet is password-protected. It contains answers to the following questions:

- How many worksheets does the Excel workbook contain?
- Are the worksheets named and ordered as required?
- Are the column labels replaced with descriptive names?
- Is the currency format used where appropriate?
- Is the requested information inserted in the left/right footers?
- Is the page setup as required?
- Has the AutoFilter been used?

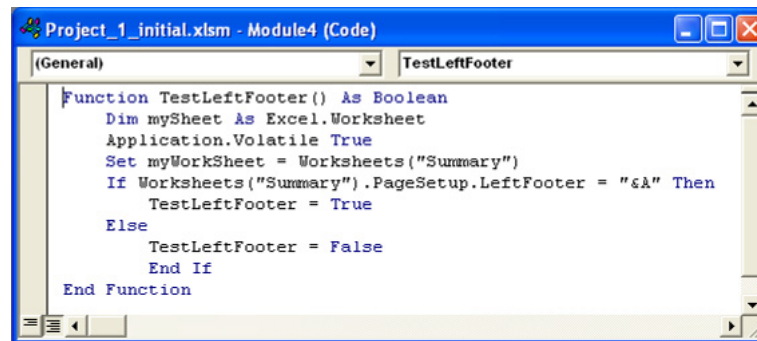
These are requirements built into the assignment. Although students know about these requirements, the numbers the system inserts into this **Answers** worksheet have no meaning to them. They will later enter these numbers in the Moodle quiz.



## User-defined functions

The user-defined functions check to verify that the attributes in a student workbook meet the requirements and if a particular function/feature has been used. The outcomes of these user-defined functions are then used in the **Answers** worksheet to generate random numbers. There are two possible ranges of random numbers: numbers from the first range indicate the correct answer, while numbers from the second range indicate an incorrect answer. Students are asked to enter these numbers in the Moodle quiz.

For each of the questions listed in the previous section, an Excel function (a user-defined function) was created. For example, the function `TestLeftFooter()` in Figure 1 checks if the **Summary** worksheet has the name of the worksheet in the left footer.



```

Project_1_initial.xlsm - Module4 (Code)
(General) TestLeftFooter
Function TestLeftFooter() As Boolean
    Dim mySheet As Excel.Worksheet
    Application.Volatile True
    Set myWorkSheet = Worksheets("Summary")
    If Worksheets("Summary").PageSetup.LeftFooter = "&A" Then
        TestLeftFooter = True
    Else
        TestLeftFooter = False
    End If
End Function
  
```



**Figure 1: User-defined function – TestLeftFooter()**

For further protection, we scrambled the answers from these functions by using the `RANDBETWEEN()` Excel function. This function assigns a random number from a specified interval to the correct answer. Students are asked to enter this number in the online quiz. Any number from the specified interval will be accepted as a correct answer. These numbers change each time a student opens or modifies the Excel workbook.


## How the quiz works


For Excel functions such as `IF` and `VLOOKUP`, each argument was entered separately. This allowed us to allocate marks for a single argument if it was correct and reduced the chances of students incorrectly entering a formula or function. At the same time, it ensured students did not miss some of the arguments.

The following tasks in the Project 1 quiz (Step 6; see Figure 2) illustrate which answers from the **Answers** worksheet students need to insert. These four answers are located here because in the Step 6 task we ask them to “make the **Summary** worksheet fit on a single page and be centered horizontally and vertically when printed in landscape orientation.”

**6**  **Step 6: The IF function** 

Marks: 3

Write down the formula in the first cell containing the IF function 

=IF(    ,  ,  )

Enter these answers from the **Answers** sheet:

Answer (17):  .


Answer (18):  .

Answer (19):  .

Answer (20):  .







**Figure 2: Quiz question**

The system will mark the quiz automatically on submission (as shown in Figure 3). The actual answers have been removed from this screenshot.


**6**  **Step 6: The IF function**


Marks: 3


Write down the formula in the first cell containing the IF function:


=IF(      ,   ,   .   )

Enter these answers from the **Answers** sheet:

Answer (17):   .

Answer (18):   .

Answer (19):   .

Answer (20):   .

[Make comment or override grade](#)

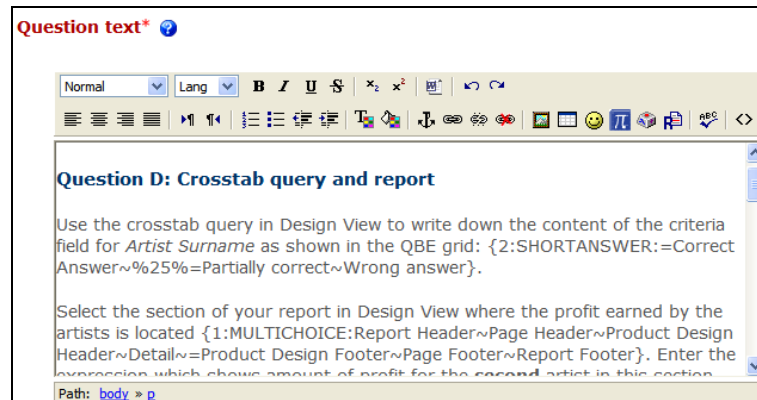
**Partially correct**

Marks for this submission: 2.1/3.

**Figure 3: Quiz answers**

If the student scrolls the mouse over their answer, an indication of whether their answer is correct or not will pop up on the screen. However, students will not be able to see the correct answer. Most of the questions are in the form of a Cloze-type question. The Cloze-type question combines multi-choice, numerical, and short answer types of questions into a single question. It allows the allocation of a fraction of mark for partially correct answers. For example, =Correct Answer is a correct answer, but if the student enters =Partially correct only, they will be awarded 25% of the full mark for this particular question (as shown in Figure 4). The Cloze question also accommodates the multiple ways in which each question can be answered and scored correctly. For instance, if the student is asked to enter the date they can enter 12-May-09, 12/05/09 or other date formats.

Each question in the Cloze type of question can have a different weight, dependant on how important the question is compared to others. For example, the first question (SHORTANSWER) has a weight of 2, while the next question (MULTICHOICE) has a weight of 1. This means the first question will produce twice as many marks as the second one if both are answered correctly.



**Figure 4: The structure of the Cloze type of question**

After the final assignment submission date, scores can be automatically transferred to the student management system overnight, if the link from Moodle has been established.

## Evaluating Methods of Marking

### **Methodology**

An evaluation of the methods of marking was done as a quantitative study. Our research hypothesis was that there is no significant difference between the marks of a human and an automated grading system. A finding of no significant difference would strongly support our claim that the automated grading system developed in this paper worked as well as a human marker.

An appropriate approach to evaluating the methods of marking would be if a human and an automated grading system marked the same papers. We could then determine the inter-rater reliability using statistics such as Cohen's kappa, inter-rater correlation, concordance correlation coefficient and intra-class correlation. This was not possible for the reasons that follow. Firstly, before the introduction of the automatic grader, students submitted the actual spreadsheet and database applications, which were then marked manually. Copies of these applications were not stored, that is, they are not available now to be marked by the automatic grader. Even if we had them, they are not prepared in the way requested in the “Preparing spreadsheet for use” section. Secondly, after the introduction of the automatic grader, students do not submit the actual spreadsheet and database applications. So the actual applications are not available for manual marking.

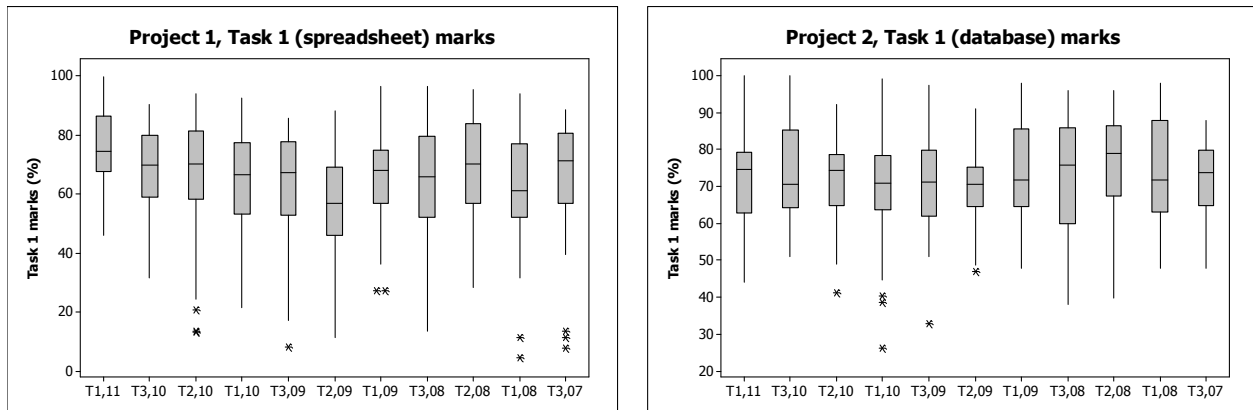
As a result of these two reasons, an evaluation was undertaken by comparing the distributions of marks allocated by the automatic grader and those allocated by a human marker. It was assumed that the student population before and after the introduction of the automatic grader was the same, and that the students on this course were no more and no less smart than past students before the introduction of the automatic grader. If this was the case, a shift in the distribution and the mean value to the right or to the left would be evident. With these assumptions, the comparisons of the mark distributions and mean values are justified and provide enough information to decide if there are significant differences between human and automated grading system marks, at least for the overall marks.

We compared different aspects of distributions of marks allocated by human and automated grading system using histograms, box-plots, and descriptive statistics. The hypothesis that there was no difference between average marks allocated by a human and an automated grading system was tested using the two sample *t*-test and Wilcoxon rank-sum test.

## Results

The online quizzes were first implemented in Trimester 2, 2009. From the regular course evaluation administrated by the Open Polytechnic Academic Office at the end of each trimester, we can conclude that students felt they got immediate feedback and deserved the marks they scored. A few students disliked that the automatic grading system required extreme attention to such details as spelling and spacing in their input. One could easily argue, however, that this is a requirement in the real world – results that are ‘almost’ correct would not be accepted elsewhere. It is critical for students to persevere with the assignment until it is accurate and complete.

For this evaluation we gathered data, that is, assignment marks from 11 trimesters. In the first five trimesters, a human marked both projects. The software applications were marked automatically in the following six trimesters. The distributions of marks for both software application parts of the projects are presented in Figure 5 (the left pane shows the spreadsheet and the right pane shows the database application). The label on the horizontal axis indicates trimester and the year (that is, ‘T1,11’ should be read as Trimester 1, 2011). The first six box-plots on the left on these two diagrams show the distribution of marks allocated by the automated grader (the online quizzes in Moodle). The remaining box-plots show the distribution of marks allocated by human markers in each trimester.



**Figure 5: The distribution of marks by trimesters**

There are obvious variations in mark distributions between trimesters. The outliers were detected in a few distributions; we show these as asterisks. But only two distributions, both in Trimester 2, 2009, are a bit different from the others. In the case of the spreadsheet application, the whole mark distribution is shifted toward lower marks (also indicated with the lowest median – the central line within the box), while in the case of the database application the size of the box (that is, the interquartile range) is the smallest of all the boxes in this pane indicating the smallest variations in the marks. However, after a small adjustment of the automatic grader in the following trimester, the mark distributions became similar to the distributions we had before introducing the automatic grader. Overall, after an initial adjustment of the automated grading system, we can say that Figure 5 shows no significant differences in the distributions of marks by trimesters. In other words, the introduction and use of the automated grading system had little impact on how the marks were distributed.

To check if there are differences in the mark distributions and the average marks between the marks allocated by the method of marking, we have combined all marks from these trimesters for each software application separately. The first two panes at the top in Figure 6 show histograms of marks, while the two panes below these show box-plots for each method of marking and each application. Two histograms for Project 1, Task 1 marks are fairly similar, both being slightly

skewed to the left. The histograms for Project 2, Task 1 marks indicate that the automatic grader gives a symmetric distribution, with the majority of marks concentrated in a rather narrow interval, while manual marking gives a distribution slightly skewed to the left, with greater variations.

These features of the distributions are even more prominent on the box-plots. The mean mark for Task 1 in both projects for each method of marking is indicated and connected. In each case, the distribution of the observations that are not outliers is fairly symmetric. However, the outlying values for both automatic and manual marking suggest non-normal populations. This should not be a major problem when applying the two sample *t*-test because of the large sample sizes.

Overall, the distribution of marks for the automated and manual methods of marking are shaped similarly and the average marks are quite close to each other, particularly in the case of Project 1, Task 1 (spreadsheet application) marks. A small difference between the distributions and average marks in the case of Project 2, Task 1 (database application) requires further investigation. However, both histograms and box-plots provide strong evidence in support of no major differences between human and automated grading.

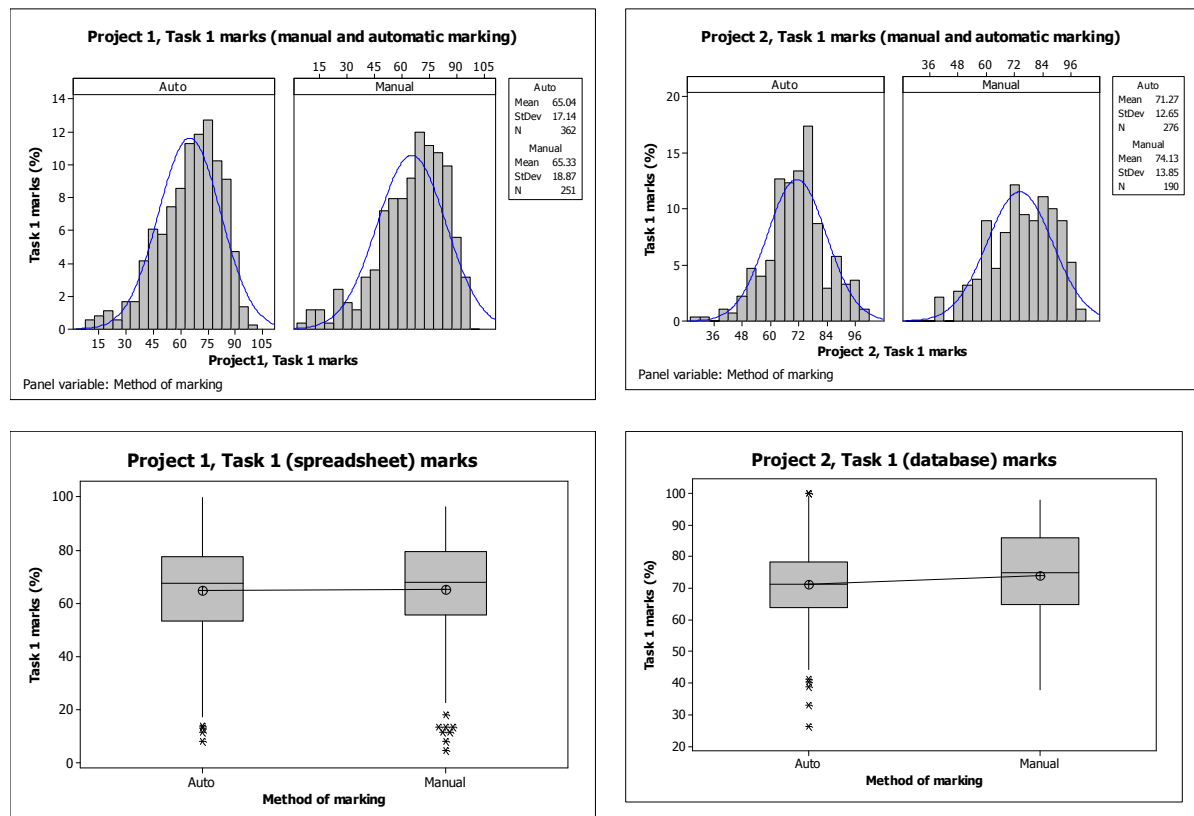


Figure 6: The distribution of marks by method of marking

The main summary statistics are presented in Table 1. Both measures of centre, that is mean and median marks for Project 1, Task 1 marks, are quite similar. The measures of variation, that is, standard deviation, coefficient of variation, and interquartile range, are also quite similar. Comparing the statistics in the last two columns in Table 1 confirms what we have concluded about the distribution of the Project 2, Task 1 marks – that the mean, median, and interquartile range for the automatic grader marks are slightly smaller than the manual marks.

**Table 1: Selected descriptive statistics**

Statistic	Project 1, Task 1 mark		Project 2, Task 1 mark	
	Automatic	Manual	Automatic	Manual
Number of marks	362	251	276	190
Mean	65.04	65.33	71.27	74.13
Standard deviation	17.14	18.87	12.65	13.85
Coefficient of variation	26.36	28.89	17.75	18.69
Minimum	8.18	4.55	26.32	38.00
First quartile	53.48	55.68	63.94	65.00
Median	67.61	68.18	71.18	75.00
Third quartile	77.83	79.55	78.24	86.00
Maximum	100.00	96.59	100.00	98.00
Range	91.82	92.04	73.68	60.00
Interquartile range	24.35	23.86	14.29	21.00
Skewness	-0.11	-0.87	-0.18	0.42
Kurtosis	0.29	0.59	-0.40	-0.52

To test the hypothesis that there are no differences between automatic grader and manual marks we conducted a test of significance. We tested a hypothesis that there are no statistically significant differences between the mean marks obtained by manual marking and those of automatic marking. The result of the two sample  $t$ -test is presented in Table 2.

**Table 2: Test of significance**

	Project 1, Task 1 mark	Project 2, Task 1 mark
<b>Two sample <math>t</math>-test</b>		
Null hypothesis	The means of the marks are equal	
Alternative hypothesis	The means of the marks are not equal	
$t$ statistic	-0.20	-2.27
Degrees of freedom	503	381
$P$ -value	0.844	0.024
Effect size ( $r$ )	0.0089	0.1155
<b>Wilcoxon rank-sum test</b>		
Null hypothesis	The medians of the marks are equal	
Alternative hypothesis	The medians of the marks are not equal	
$W$ statistic	109842.5	60737.0
$P$ -value (adjusted for ties)	0.5493	0.0094

Applying the two sample  $t$ -test to Project 1, Task 1 marks we got  $t=-0.20$ ,  $df=503$ ,  $P$ -value=0.844. Because the  $P$ -value of 0.844 is greater than the standard 5% level of significance, we can say that there is not enough evidence to reject the null hypothesis. In other words, we can attribute the difference in the Project 1, Task 1 means to sampling variance.

When testing the same hypothesis for Project 2, Task 1 marks we got  $t=-2.27$ ,  $df=381$ ,  $P$ -value=0.024. Because the  $P$ -value of 0.024 is less than the standard 5% level of significance, we can reject the null hypothesis at that level of significance. In other words, we cannot attribute the difference in the Project 2, Task 1 means to sampling variance but to the method of marking.

A word of caution is needed when interpreting this result. While the test statistic is significant, it does not mean that the effect it measures (the method of marking) is meaningful or important. In other words, because of the large sample size, even a very small or unimportant effect can turn out to be statistically significant. This is what happened with Project 2, Task 1 marks. To quantify

the effect size we used the correlation coefficient as an objective measure of the importance of an effect. In the case of Project 2, Task 1 marks, the effect size is 0.1155 and that is, according to Cohen (1992), a small effect. That is, the method of marking explains only 1% of the total variance in Project 2, Task 1 marks. The effect size is even more negligible – as in the case of the Project 1, Task 1 marks. Since we kept the outliers that caused the departure of the marks distributions from normality in the dataset, we also used a non-parametric equivalent of the two sample *t*-test, that is, the Wilcoxon rank-sum test, to compare the median marks of these two populations. This test does not depend upon an assumption of normality.

The values of the Wilcoxon rank-sum statistic and *P*-values are presented in the lower part of Table 2. These tell us that we reached the same conclusion at a 5% level of significance, using both the parametric *t*-test and nonparametric equivalent, as the Wilcoxon rank-sum test: there is no statistically significant difference between the Project 1, Task 1 medians when using different methods of marking (automatic vs. manual). However, in the case of the Project 2, Task 1 median marks, the Wilcoxon rank-sum test suggests we reject the null hypothesis of no difference between median marks. The discussion of the effect size above suggests that we should not be too concerned with the test result in the case of the Project 2, Task 1 marks. However, we may consider a fine adjustment of the Project 2, Task 1 test so that the marks distribution more closely matches the distribution of marks obtained when a human marked students' assignments.

In summarising these results, we can say that in spite of a small variation in marks between a human marker and an automated grading system, experience in the past eleven trimesters confirms our main research hypothesis – that there are no significant difference between the marks allocated by a human marker and the marks of a spreadsheet automated grading system. In the case of a database automated grading system the difference is significant, but the effect size is very small.

## Conclusion and Future Work

This paper describes the design of an assessment and the development of an automatic grading system to assess advanced students' spreadsheet and database skills in the *Information Systems* course at the Open Polytechnic.

The automated grading system discussed in this paper differs from other automated grading systems developed for Office skills (Hill, 2003, 2004; Kline & Janicki, 2003; Koike et al. 2005; Russell & Cumming, 2005; Waldman & Ulema, 2008) in one major aspect. That is, students are not required to send the developed software application and lecturers are not required to receive and check the actual application. This is the major advantage of this automated grading system over other systems because, as Kline and Janicki (2003) emphasized, the physical handling of submitted assignments can become difficult and time consuming.

Using the user-defined functions in Excel, we have been able to check the attributes of software applications and if the requested feature or function was used appropriately. Lecturers were not required to open the actual application because students verified their own application by entering scrambled user-defined functions outputs in the online quiz. Alternative correct answers and partially correct answers were coded in the online quizzes. This allowed for flexibility in the acceptance of different solutions.

In a similar way to one of the methods proposed by Waldman and Ulema (2008) for the implementing of a custom grading program, our automated grading system is based on the use of Microsoft's Visual Basic for Applications (the same approach adopted by Koike et al., 2005). Other grading systems – such as Hill's MEAGER system – require a program that interfaces with a running copy of Excel. A major disadvantage of our automated grading system is that the answers are hard-coded into the system. While this is also a limitation of the other grading systems (such as the system developed by Kline & Janicki, 2003, which allows only two correct answers), we

have built in a few alternative acceptable answers to cater for possible variations in the solutions provided by students.

Marking formulas in Excel can be quite difficult because the same result can be achieved using different Excel functions and formulas. Hill (2004) addressed this problem by using an interpreter for the symbolic manipulation of mathematical expressions. We adopted a different approach – that of limiting the students' options so that they did not use a function other than the one expected. This can be interpreted as restricting the freedom of students to solve particular tasks using their own modeling knowledge, but it can be argued that this approach ensures that students fully grasp the core software functions and spreadsheet and database skills required in the course.

An evaluation of the methods of marking (manual vs. automatic grader) confirms that the automatic grader allocates marks in a similar way to a human marker. In the case of the spreadsheet application, the marks distribution for the automatic grader matches the distribution of the human marker. The test of significance confirms that the both mean and median marks for these two methods of marking are not statistically different. While the test of significance suggests rejecting the null hypothesis of no differences between the means and medians for these two methods of marking in the case of the database application, the measure of the effect size shows that the methods of marking have a small effect, explaining only 1% of the total variance.

We hope that this application of automated grading system integrated with the Moodle learning platform is transferable to other subject areas. We found that our automated grader system had the following advantages for students and staff:

For students:

- Instant feedback on formative, real-time assessment quizzes assists learning and motivation.
- Instant feedback on summative, real-time assessment quizzes provides instant feedback on their success.
- Real-time assessment quizzes provide a structure for the less organized, less clear-thinking students.

For staff:

- Although the initial setup involves a substantial amount of thought and effort, the result of this has great ongoing workload benefits. It enables staff to concentrate on providing ad-hoc feedback to student questions not already covered by the other course resources. This is, therefore, also a student benefit.
- As staff workload is reduced, they are able to do research which underpins the degree course and keeps it current. This is, therefore, also a student benefit.
- Moodle quizzes enable lecturers to receive statistics on the validity of their questions.

For future work, we still need to address issues of plagiarism and the authentication of candidates, as in any other form of assessment in distance education. An appropriate evaluation of the methods of marking should also be undertaken. This means that the same assignment should be marked by a human and an automatic grading system and that the use of the different measures of inter-rater reliability will show any systematic discrepancies between a human marker and an automatic grader.

So what can be said about the future of automated assessment / grading / marking or e-assessment in general? We agree with the vision outlined in Whitelock and Brasher (2007) that



“traditional paper-based summative assessments will continue to migrate to computer delivery .... Increasingly, aspects of courses that lend themselves to objective question types, or that use assessments based on visualisations of concepts or procedures, will be completed online. These strategies are likely to be combined with short-answer questions, marked by computer and checked by humans, to probe the learner’s ability to form links between areas of knowledge”. (p. 36)

## References

- Amelung, M., Piotrowski, M., & Rösner, D. (2006). EduComponents: Experiences in e-assessment in computer science education. *Proceedings of the ITiCSE'06*, Bologna, Italy, 88-92.
- Buzzetto-More, N. A., & Alade, A. J. (2006). Best practices in e-assessment. *Journal of Information Technology Education*, 5, 251-269. Retrieved September 27, 2011 from <http://www.jite.org/documents/Vol5/v5p251-269Buzzetto152.pdf>
- Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, 1(3), 98-101.
- Hill, T. G. (2003). MEAGER: Microsoft Excel automated grader. *The Journal of Computing Sciences in Colleges*, 18(6), 151-164.
- Hill, T. G. (2004). Excel grader and Access grader. *ACM SIGCSE Bulletin*, 36(2), 101-105.
- Kline, D. M., & Janicki, T. N. (2003). Enhancing economic and finance learning through automated grading of spreadsheet exercises. *Journal of Economics and Financial Education*, 2(2), 23-29.
- Koike, H., Akama, K., Chiba, M., Ishikawa, T., & Miura, K. (2005). Developing an e-learning system which enhances students’ academic motivation. *Proceedings of the 33rd Annual ACM SIGUCCS Conference on University and College Computing Services*, Monterey, CA, USA, 147-150.
- Koike, H., Akama, K., Morita, H., & Mura, K. (2006). Using an automatic marking system for programming courses. *Proceedings of the 34th Annual ACM SIGUCCS Conference on University and College Computing Services*, Edmonton, Alberta, Canada, 173-178.
- Malmi, L., Korhonen, A., & Saikkonen, R. (2002). Experiences in automatic assessment on mass courses and issues for designing virtual courses. *Proceedings of the 7th Annual Conference on Innovation and Technology in Computer Science Education*, Aarhus, Denmark, 55-59.
- Musham, G. (2004). *Using computer technology to support student assessment*. OET Assignment, University of London.
- Russell, G., & Cumming, A. (2005). Automatic checking of SQL: Computerised grading. *International Journal of Learning*, 12(3), 127-134.
- Swithenby, S. J. (2006). *E-assessment for open learning*. UK: The Open University.
- Tshibalo, A. E. (2007). The potential impact of computer-aided assessment technology in higher education. *South African Journal of Higher Education*, 21(6), 684-693.
- Waldman, M., & Ulema, M. (2008). Automated measurement and analysis of effectiveness of teaching selected Excel topics in an introductory IS class. *The Journal of Computing Sciences in Colleges*, 23(5), 73-82.
- Whitlock, D. & Brasher, A. (2006, June). *Roadmap for e-assessment*. Joint Information Systems Committee Report.
- Whitlock, D. & Brasher, A. (2007). *Effective practice with e-assessment. An overview of technologies, policies and practice in further and higher education*. Joint Information Systems Committee (JISC).
- Zhenming, Y., Liang, Z., & Guohua, Z. (2003). A novel web-based online examination system for computer science education. *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference*, Boulder, CO, USA, 7-10.

## Acknowledgement

This article is a substantial revision of an earlier conference paper entitled “Automated assessment of advanced office skills”, presented at the DEANZ 2010 Conference, April 25-28, in Wellington, New Zealand. ([http://www.deanz.org.nz/home/conferenceDocs/Kovacic\\_Green.pdf](http://www.deanz.org.nz/home/conferenceDocs/Kovacic_Green.pdf))

We take this opportunity to thank editor and anonymous journal reviewers for their valuable comments on the paper. It goes without saying that the remaining shortcomings and errors are our responsibility.

## Biographies



**Dr. Zlatko J Kovačić** is an Associate Professor in the School of Information and Social Sciences at the Open Polytechnic. Zlatko has a varied academic background and diverse research interests, ranging from core interests relating to IT learning and teaching, eCommerce, time series analysis, and multivariate analysis. His current research is focused on the cognitive processes in distance education using computers and communications technologies, on the social and cultural aspects of information technology and the time series analysis of emerging financial markets.



**John Steven Green** is a Senior Lecturer in the School of Information and Social Sciences at the Open Polytechnic. John has published papers in Information Science journals and conference proceedings across the world. His area of research includes planning and social networking, as well as automating teaching and grading in online education.