COP
Provided

The Open University

Open Research Online

The Open University's repository of research publications
and other research outputs

Applying latent semantic analysis to computer assisted assessment in the Computer Science domain: a framework, a tool, and an evaluation

Thesis

How to cite:

Haley, Debra (2009). Applying latent semantic analysis to computer assisted assessment in the Computer Science domain: a framework, a tool, and an evaluation. PhD thesis The Open University.

For guidance on citations see FAQs.

© 2009 D. T. Haley
Version: Version of Record

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data policy on reuse of materials please consult the policies page.

oro.open.ac.uk

Applying Latent Semantic Analysis to Computer Assisted Assessment in the Computer Science Domain: A Framework, a Tool, and an Evaluation

Debra Trusso Haley B.A., M.S.

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Computer Science

Department of Computer Science
Faculty of Mathematics, Computer Science and Technology
The Open University

December 2008

Dedication

To my family, with love.

To Charles, my travelling companion. It's been a good journey so far. May we have many more exciting places to see. Thank you for all the dinners you made during this last gruelling 18 months and for nourishing me in so many other ways.

To my two fine sons, key participants in our annual last family vacations. To David, who is travelling a similar path. I proudly acknowledge that you have surpassed me on the way. To Steven, who has chosen his own path. I salute your inner strength. You two have taught me all about motherly love and pride.

Abstract

This dissertation argues that automated assessment systems can be useful for both students and educators provided that the results correspond well with human markers. Thus, evaluating such a system is crucial. I present an evaluation framework and show how and why it can be useful for both producers and consumers of automated assessment systems. The framework is a refinement of a research taxonomy that came out of the effort to analyse the literature review of systems based on Latent Semantic Analysis (LSA), a statistical natural language processing technique that has been used for automated assessment of essays. The evaluation framework can help developers publish their results in a format that is comprehensive, relatively compact, and useful to other researchers.

The thesis claims that, in order to see a complete picture of an automated assessment system, certain pieces must be emphasised. It presents the framework as a jigsaw puzzle whose pieces join together to form the whole picture.

The dissertation uses the framework to compare the accuracy of human markers and EMMA, the LSA-based assessment system I wrote as part of this dissertation. EMMA marks short, free text answers in the domain of computer science. I conducted a study of five human markers and then used the results as a benchmark against which to evaluate EMMA. An integral part of the evaluation was the success metric. The standard inter-rater reliability statistic was not useful; I located a new statistic and applied it to the domain of computer assisted assessment for the first time, as far as I know.

Although EMMA exceeds human markers on a few questions, overall it does not achieve the same level of agreement with humans as humans do with each other. The last chapter maps out a plan for further research to improve EMMA.

Author's Declaration

All of the work presented in this thesis describes original contributions of the author (Exceptions noted on the Acknowledgments Page). Some of the material in this dissertation was published previously in the following papers:

Haley, Debra Trusso, Thomas, Pete, De Roeck, Anne and Petre, Marian (2007a). Seeing the Whole Picture: Evaluating Automated Assessment Systems. ITALICS Special Issue on Innovative methods for teaching programming.

Haley, Debra Trusso, Thomas, Pete, Petre, Marian and De Roeck, Anne (2007b). EMMA - a Computer Assisted Assessment System based on Latent Semantic Analysis. ELeGI Final Evaluation. Technical Report 2008/14. Milton Keynes, UK. The Open University.

Haley, Debra Trusso, Thomas, Pete, De Roeck, Anne and Petre, Marian (2007c). Tuning an LSA-based Assessment System for Short Answers in the Domain of Computer Science: The Elusive Optimum Dimension. 1st European Workshop on Latent Semantic Analysis in Technology Enhanced Learning, Heerlen, The Netherlands.

Haley, Debra, Thomas, Pete, De Roeck, Anne and Petre, Marian (2007d). Measuring Improvement in Latent Semantic Analysis-Based Marking Systems: Using a Computer to Mark Questions about HTML. Proceedings of the Ninth Australasian Computing Education Conference (ACE2007), Ballarat, Victoria, Australia, Australian Computer Society Inc.

Haley, Debra, Thomas, Pete, Nuseibeh, Bashar, Taylor, Josie and Lefrere, Paul (2005a). The Learning Grid and E-Assessment using Latent Semantic Analysis. Towards the Learning GRID: Advances in Human Learning Services, IOS Press.

Haley, Debra Trusso, Thomas, Pete, De Roeck, Anne and Petre, Marian (2005b). A Research Taxonomy for Latent Semantic Analysis-Based Educational Applications. Technical Report 2005/09. Milton Keynes, UK. The Open University.

Haley, Debra Trusso, Thomas, Pete, De Roeck, Anne and Petre, Marian (2005c). A Research Taxonomy for Latent Semantic Analysis-Based Educational Applications. International Conference on Recent Advances in Natural Language Processing'05, Borovets, Bulgaria.

Thomas, Pete, Haley, Debra, De Roeck, Anne and Petre, Marian (2004). E-Assessment using Latent Semantic Analysis in the Computer Science Domain: A Pilot Study. Proceedings of the eLearning for Computational Linguistics and Computational Linguistics for eLearning Workshop at COLING 2004., Geneva.

Haley, Debra, Thomas, Pete, Nuseibeh, Bashar, Taylor, Josie and Lefrere, Paul (2003). E-Assessment using Latent Semantic Analysis. Proceedings of the 3rd International LeGE-WG Workshop: Towards a European Learning Grid Infrastructure, Berlin, Germany.

Acknowledgements

With great pleasure and gratitude, I would like to thank the following:

Pete Thomas, my primary supervisor, for his always prompt and valuable feedback and for sharing his knowledge of mathematics and statistics. He helped me when I needed him and left me to get on with the work when I didn't. It was his suggestion to use LSA to mark short answers and EMMA is based on his early prototype. And especially, for his kindness and good cheer. He is a treasure.

Marian Petre and Anne De Roeck, my brilliant secondary supervisors. Marian, for helping me learn to write and Anne for always telling me the truth. Their faith in me kept me going.

The PG Forum was my Thursday morning rock of stability. I owe endless thanks to Marian Petre and Trevor Collins for teaching me how to do research and for providing a space for social contact and emotional support.

Bashar Nuseibeh was very good to the Haleys, both in securing funding and as a teacher, mentor and friend.

Finally, I thank the Open University for the generous bursary, my fine office and computer, and the stimulating environment. While I will not miss the pressures of doing a Ph.D., I will surely miss my many OU friends.

I am grateful to acknowledge the work of several people funded by the ELeGI project. Charles Haley wrote code to extract student answers to relevant questions from the assessment course scripts and to use these answers to populate the database. Diane Evans proof read and corrected the student answers for those cases where the code failed to extract the answers correctly. Adam Gawronsky wrote code for a study to compare human markers – it presented 60 randomly selected answers to 18 different questions to the markers, collected their marks, and updated the database.

The work reported in this study was partially supported by the European Community under the Innovation Society Technologies (IST) programme of the 6th Framework Programme for RTD - project ELeGI, contract IST-002205. This document does not necessarily represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.

Table of Contents

Table of Figures

Table of Tables

Table of Equations

List of Acronyms

AC1	first order agreement coefficient for IRR
CAA	computer assisted assessment system
CS	computer science
DFD	data flow diagram
ELeGI	European Learning Grid Infrastructure
EMMA	ExaM Marking Assistant
HOL	higher order learning
IR	information retrieval
IRR	inter-rater reliability
LSA	latent semantic analysis
LSI	latent semantic indexing
MCQ	multiple choice question
OU	The Open University
SD	standard deviation
SM	simple metric
SVD	singular value decomposition
tfidf	a type of weighting function – term frequency/inverse document frequency

Chapter 1. Introduction

This dissertation documents EMMA (ExaM Marking Assistant), a computer assisted assessment system (CAA) based on a statistical Natural Language Processing technique called Latent Semantic Analysis (LSA). LSA has been used to assess essays; EMMA stretches the technique by marking short, free text answers. The dissertation provides a framework for describing and evaluating CAAs and applies that framework to EMMA. The goal for writing EMMA was to develop a marking program that achieved results as consistent as human markers: I wanted EMMA's marks to match human marks as closely as human marks matched each other. The framework provides all the necessary information to demonstrate how close EMMA comes to the goal.

Subsection 1.1 discusses assessment: its importance, the pros and cons of CAA, the cost effectiveness of CAA, and its ability to evaluate higher order learning. Subsection 1.2 gives the motivation for the research documented in this dissertation, subsection 1.3 lists the research questions, and subsection 1.4 provides a roadmap of the rest of the dissertation.

1.1 Assessment

This section sets the stage for the rest of the dissertation by summarising the assessment literature. It starts by discussing the importance of assessment and then moves to the advantages and disadvantages of CAA. Next, it highlights some of the issues of the cost effectiveness of CAA. Finally, it explores the question of whether or not CAA is suitable for assessing higher level learning. EMMA is an attempt to fill the need for a CAA system that addresses higher level learning.

1.1.1 Importance of assessment

McAlpine (2002 p. 4) gives the following description of assessment:

> " ...assessment is a form of communication. This communication can be to a variety of sources, to students (feedback on their learning), to the lecturer (feedback on their teaching), to the curriculum designer (feedback on the curriculum) to administrators (feedback on the use of resources) and to employers (quality of job applicants)."

Assessment is "a critical activity for all universities" (Conole & Bull, 2002 pp. 13-14) and "there is no doubt" about its importance (Brown, Bull & Pendlebury, 1997 p. 7). Assessment is "widely regarded as the most critical element of learning" (Warburton & Conole, 2003b). One researcher claimed "… the most important thing we do for our students is to assess their work" (Race, 1995). One reason for the importance of assessment given by several researchers (Brown, Bull & Pendlebury, 1997; Berglund, 1999 p. 364; Daniels, Berglund, Pears & Fincher, 2004) is that assessment can have a strong effect on student learning. Brown, Bull & Pendlebury (1997 p. 7) claimed students learn best with frequent assessment and rapid feedback and added that one reason assessment is so important is that the right type of assessment can lead to deeper learning (1997 p. 24).

1.1.2 The growth of interest in Computer Assisted Assessment (CAA)

Computer Assisted Assessment (CAA) is assessment delivered and/or marked with the aid of computers (Conole & Bull, 2002). A 2002 study reported an increasing interest in and use of CAA in the preceding five years (Bull, Conole, Davis, White, Danson & Sclater, 2002). The number of papers published at the annual CAA conferences at Loughborough University supports the 2002 study. The number has grown from 20 in 1999 (the third year of the conference and the first year for which figures are available) to 40 in 2007 (www.caaconference.com) with an average of about 37 papers a year.

Brown, Bull & Pendlebury (1997 p. 40) claimed that the increased interest in assessment in the previous ten years "arises from the [British] government's pincer movement of insisting

upon 'quality' while at the same time reducing unit costs" and predict "further cuts in resources" (1997 p. 55). They claim a 63% cut in per student resources since 1973 (1997 p. 255).

Ricketts & Wilks (2002b p. 312) agreed with Brown, Bull & Pendlebury (1997) for the increasing interest in CAA – decreasing resources per student require a cost savings, which can be gained by decreasing tutor marking time. A 2003 survey (Carter, Ala-Mutka, Fuller, Dick, English, Fone & Sheard, 2003) gave a related reason for the interest in CAA: increasing enrolment. They cited the increasing number of ITiCSE (Integrating Technology into Computer Science Education) papers as evidence for the increased interest in CAA.

The next subsections look at some reasons given in the literature for not using CAA followed by reasons given in favour of using CAA.

1.1.3 Arguments against CAA

A national survey (Bull, 1999) found that the "main disadvantages [of CAA] are perceived to be access to and reliability of hardware and software, the amount of time needed (sic) create and organise delivery, and the difficulty of writing good questions". Bull (1999) noted that the cost of CAA was a concern, both in the human time needed to learn to use the software as well as the cost of CAA systems. The cost of human time was again mentioned four years later (Warburton & Conole, 2003a).

A broad range of concerns is expressed by ALT, the Association for Learning Technology, which cautions (Alt, 2003):

> - "The immaturity and volatility of some learning technology mean that there is a lot of work involved in keeping up with available products, especially with a market that is shaking out. Accordingly, much effort is wasted through poor understanding of the technology and its application.

- There are a lot of products and services which are not especially suited to UK FE and HE pedagogic models.
- It is possible to make expensive errors when there is a misalignment between technology, pedagogy and institutional infrastructure or culture. These errors are often repeated in parallel between educational institutions.
- Standards and specifications are evolving, hard to understand, easy to fall foul of, and tend to be embraced with zeal, without the cost and quality implications being properly understood.
- Much effort is also dissipated through a poor understanding of the theory and pedagogy that underpins the use of the technology."

Another worry mentioned in the literature is the suitability of CAA for assessing higher level learning outcomes. It is widely believed that these outcomes can only be measured by essay questions and that CAA is not able to handle essays. Subsection 1.1.6 surveys this issue.

1.1.4 Arguments for CAA

The perceived advantages of CAA greatly outnumber the perceived disadvantages presented in subsection 1.1.3. This subsection highlights the arguments made in favour of CAA.

1.1.4.1 Save time/costs

Saving time, and therefore reducing costs, is by far the most common reason given for using CAA. According to Mason & Grove-Stephenson, (2002), marking takes up 30% of a teacher's time. Unfortunately, they fail to provide evidence or a citation for this claim, but if true, it supports the commonly held view that saving markers' time results in a substantial cost savings. The Ceilidh project (Benford, Burke, Foxley & Higgins, 1996) claimed a "massive" savings in marking time by using CAA. Another paper (Summons, Coldwell, Henskens & Bruff, 1997) reported that using CAA reduced marking time by 67%, saving 100 hours of tutor wages. Joy & Luck stated that CAA reduces marking time (1998). Woit & Mason (2003) used five years of data to justify their claim of a reduction in marking time. A 2003 international survey (Carter,

Ala-Mutka, Fuller, Dick, English, Fone & Sheard) found that CAA saved markers' time, particularly when the class size was large. A study of a CAA system called *submit* demonstrated time savings for both markers and lab staff (Venables & Haywood, 2003).

The survey and studies in the previous paragraph support the claim that CAA is widely perceived to save time. However, the above studies consider only the actual time spent marking without considering various administrative duties including the time spent learning the system and preparing questions, let alone the cost of computer hardware and software. Section 1.1.5 explores the issue of cost-effectiveness.

1.1.4.2 Accommodate large class sizes

Related to cost savings, the increase in class sizes is another argument for using CAA. Preston & Shackelford (1999) provided figures that illuminate the problem: their university has "hundreds" of students enrolling in their first two basic computer science classes, which requires "over a hundred" markers to mark "over 4,000 assignments" a week. Croft, Danson, Dawson & Ward (2001) claimed that using CAA is the only way to ensure regular assessment of large groups of students.

One paper (Ricketts, Filmore, Lowry & Wilks, 2003) made an interesting point: larger class sizes have led to the situation where "the cost of assessment in higher education now surpasses the cost of teaching" because "the cost of assessment is a cost per student, whereas the cost of teaching is more related to the hours per course". This means that larger class sizes directly impact the cost of assessment.

1.1.4.3 Reduce marker bias and improve consistency

Another goal of using CAA is to reduce marker bias and improve consistency. The papers cited in this section used the terms bias and consistency without defining them. In the following paragraphs, I assume that bias is a prejudice either for or against a student and that consistency is a broader term referring to repeatability of results that can vary due to either bias or human error (e.g. adding marks or transcribing incorrectly, or differing judgments).

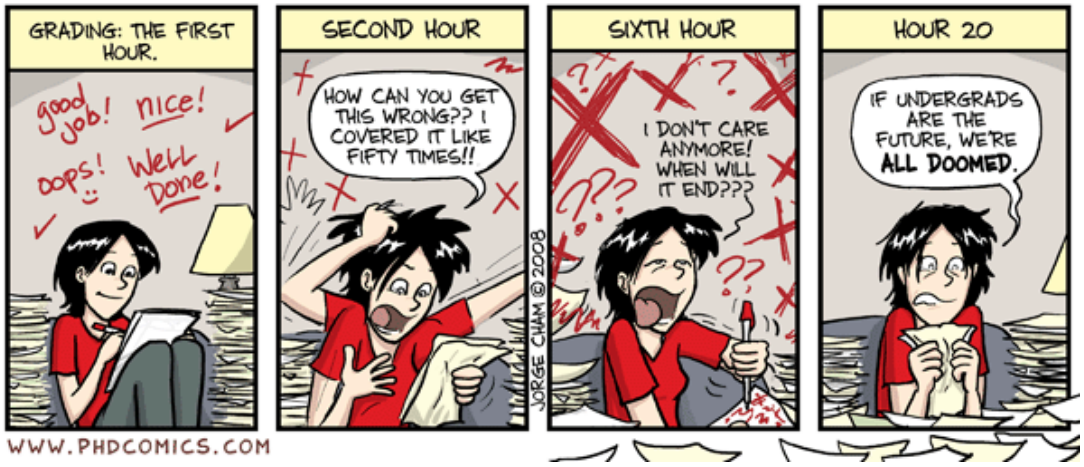


Figure 1-1 Human fallibility: A source of bias and inconsistency in marking

Used by permission: "Piled Higher and Deeper" by Jorge Cham www.phdcomics.com

Figure 1-1 is a humorous depiction of how human fallibility can cause marker bias and lack of consistency.

Christie (2003) gave a comprehensive list of causes leading to lack of consistency. (Although Christie mentions essays, his comments generalize to short answers, which is the focus of this dissertation.) The comic strip exemplifies some of these factors.

> "Manual marking is prone to several adverse subjective factors, such as:
>
> The length of each essay,
>
> The size of the essay set,
>
> The essay's place in the sequence of the essays being marked,
>
> The quality of the last few essays marked affecting the mark awarded to the essay currently being marked,
>
> The effect of the essayist's vocabulary and errors (spelling and grammar) on the marker,
>
> The marker's mood at the time of marking
>
> Marker's expectations of the essay set and of each essayist."

A thoughtful paper discussing a survey on bias (Sabar, 2002) reported that educators employ a wide range of solutions to the problem of how to resolve assessment difficulties arising from favouritism, implicitly acknowledging the ubiquity of possible bias in marking.

One study found bias in manual marking due to "inter-tutorial or intra-tutorial marking variations" (Summons, Coldwell, Henskens & Bruff, 1997). They claimed that reducing bias would have been "extremely difficult" without their CAA due to the large number of tutors and that most of their tutors "would have varied from the marking scheme". Thus, CAA led to more consistent marking.

The developers of Ceilidh (Benford, Burke, Foxley & Higgins, 1996) reported increased consistency using their CAA:

> "… hand marking of any form of coursework can lead to a student being treated less fairly than others. For instance, coursework marked by more than one person will lead to inconsistencies in marks awarded due to differing ideas of what the correct answer should be. This coupled with other problems such as racism, sexism and favouritism can lead to certain students achieving poorer marks than they deserve. We believe that such explicit discrimination is reduced, if not eliminated, by the use of the Ceilidh system since it marks each solution consistently."

Joy & Luck (1998) argued that CAA provides consistency in marking: "… while the accuracy of marking, and consequently the confidence enjoyed by the students in the marking process, is improved. In addition, consistency is improved, especially if more than one person is involved in the marking process." Three years later, the consistency argument was still being made (Davies, 2001). An international survey (Carter, Ala-Mutka, Fuller, Dick, English, Fone & Sheard, 2003) reported that CAA is widely perceived to increase consistency in marking. Conole & Warburton agree with the survey that CAA "offers consistency in marking" (2005 p. 26). Tsintsifas (2002 p. 19) states:

> "Reliability and fairness increase by automating the assessment process because the same marking mechanism is employed to mark each piece of work. There is no possibility of discrimination and students are well aware of the fact that everyone is treated equally by the system."

The Open University follows formal procedures to address marker bias and inconsistency. The OU is particularly susceptible to these problems given the huge number of students and tutors involved in every presentation of a course. For example, almost 3,000 students took the computing course that this dissertation uses for data. Part of the work involved in preparing a course is producing detailed Tutor Notes and Marking Schemes to help ensure marking consistency. Every exam undergoes moderation, that is, trained markers re-mark the exams and conflicting marks are investigated and resolved. A sample of all homework assignments is monitored to verify accuracy and consistency. These procedures are implicit evidence that The OU believes human marking can suffer from bias and inconsistency.

The papers cited in this subsection claimed, but did not provide evidence, that CAA improves marking consistency. Brown, Bull & Pendlebury (1997 p. 234) cite literature on general assessor inconsistency from 1890 to 1963. Newstead (2002), in an update of the classic article on the reliability of examiners (Newstead & Dennis, 1994) provides evidence of poor marker reliability in the field of psychology. Despite these examples, I could find no CAA literature that backed up, with evidence, the claim that CAA improves marking consistency. To do so, the researchers would need to present evidence that human markers are not consistent either with each other and/or with themselves over time and that using CAA leads to improvement. This dissertation provides evidence (see Chapter 3) that human markers are far from consistent, at least in the domain of computer science.

1.1.4.4 Provide rapid feedback

A national survey (Bull, 1999) reported a widespread consensus that rapid feedback is one of the main advantages of using CAA. An international survey in 2003 (Carter, Ala-Mutka, Fuller, Dick, English, Fone & Sheard) confirmed Bull's findings and added a comment from one of the participants about the benefit: "… the immediacy of feedback, although not necessarily the quality of feedback … poor feedback from CAA is preferable to detailed manual feedback which arrives weeks after the work was completed…".

In addition to the surveys of educators cited in the previous paragraph, a number of surveys of students reported that students appreciate the rapid feedback they experienced from using CAA (Summons, Coldwell, Henskens & Bruff, 1997; Ricketts & Wilks, 2002a p. 478; Osborne & Winkley, 2006).

1.1.4.5 Improve pedagogy

Various researchers have mentioned pedagogical reasons for adopting CAA. Carter, Al-Mutka, Fuller, Dick, English, Fone & Sheard (2003) stated that CAA frees students to work at their own pace. Tsintsifas claimed that CAA enforces deadlines by both "students and educators" (2002 p. 19). Venables & Haywood (2003) agreed with Tsintsifas and concluded that their students respected the computer's deadline more than their tutor's deadline because more of them turned in their submissions on time when using the *submit* system.

Mulligan (1999) concluded that students worked harder when using CAA based on survey responses from tutors even though the students did not report working harder. (Mulligan questioned the honesty of the students.) A later survey (Croft, Danson, Dawson & Ward, 2001) agreed that CAA caused students to work harder and added that it "encouraged them to work consistently". Wood & Burrow (2002) reported that their students believed that CAA "encouraged independent student learning". A national survey (McKenna, 2001) reported "There's far more effort over a prolonged period rather than a short intense burst at the end".

An international survey found that "those with no experience of CAA suggest that it cannot be used to test higher-order learning outcomes and that the quality of the immediate feedback is poor; these negative opinions diminish as experience is gained" (Carter, Ala-Mutka, Fuller, Dick, English, Fone & Sheard, 2003). Woit & Mason, in a five year study (2003), found that "online evaluation … can result in increased student motivation and programming efficacy".

1.1.4.6 Analyse a database of marks

Several papers mention the advantages that can accrue from the collection and analysis of marks that CAA makes easy to achieve. Hopkins, in a book first published in 1941 and now in its 8th edition (Hopkins, 1998) as cited by (Preston & Shackelford, 1999), says research using

stored marks can lead to "customized courses", "better use of class time", and "student trend analysis". Summons, Coldwell, Henskens & Bruff (1997) claim that their CAA system resulted in "assessment management advantages" although did not detail what they were. Bull (1999) was more specific, citing "statistical analysis of results". The developers of Ceilidh reported that the "general progress monitoring facilities" were helpful (Benford, Burke, Foxley & Higgins, 1996).

1.1.5 Cost Effectiveness of CAA

Subsection 1.1.4.1 gave examples of papers that claimed a time-savings for the markers. The weakness of these arguments, however, is that the researchers looked at only the actual time spent marking and not the time spent learning the system and developing the questions. Wood & Burrows (2002) reported that staff time and the cost of the hardware and software were the major expenses in using CAA, and that the majority of the staff effort was creating questions. Sclater & Howie (2003) stressed the point that "constructing high quality questions is difficult, time consuming and expensive" and proposed lowering the cost by creating item banks: "Developing items and assessments across a subject area or sector can bring economies of scale in the development process and a considerable reduction in duplication of effort in different colleges and universities" (Sclater & MacDonald, 2004). Wood & Burrows (2002) expected a cost savings by reusing questions. A more general warning comes from Stephens (1994) as quoted by (Sim, Holifield & Brown, 2004): "The perceived benefits of CAA of freeing lecturers' time can be illusive if no institutional strategy or support is offered". One of the points this dissertation makes is that the amount of human effort required to set up and use a CAA is an important item to consider when evaluating the cost of a CAA and thus is part of the description and evaluation framework I developed as part of the work done for this dissertation (see Chapter 5).

In contrast to the researchers cited in 1.1.4.1 who claimed time and cost savings by using CAA, one respondent to McKenna's survey (2001) noted "…although CAA eventually saves time, all the time costs are at the beginning, whereas, with essay questions, the setting of the

paper is much faster at the start of the process." Another respondent claimed that it had taken over 10 years to realize a cost savings (McKenna, 2001). Brown, Bull & Pendlebury (1997 p. 220) invoke the 2/3 rule to caution about the cost-effectiveness of CAA:

> "In the short term, CAA is unlikely to save time and resources. All innovations require detailed planning and organisation. The 2/3 rule applies. An innovation can be cheap, fast, and of high quality. At best, it can only have two of these three characteristics. If it is cheap and fast, it will not be of high quality. If it is fast and high quality, it will not be cheap. Make your choice."

Bull (2000 p. 10) nicely summarised the difficulty of quantifying the cost effectiveness of CAA:

> "We have learnt that it is problematic to identify costs related to learning technology systems as many of the costs are hidden and relevant technologies and infrastructure may already exist in whole or part. The technical and pedagogical support which is required to effectively implement learning technology is costly because it requires a long term commitment. … In addition there are costs associated with staff development which should be considered but again are interdependent with other activities and problematic to measure."

Subsection 1.1.4 explored the literature relating to the perceived pros and cons of using CAA, one of which is its ability to assess higher order learning (HOL). The next subsection focuses on this often-stated concern.

1.1.6 Assessing higher order learning (HOL) with CAA

The HOL areas comprise these three levels of Bloom's Taxonomy (1956): Analysis, Synthesis, and Evaluation. Table 1.1 is taken from Carneson, Delpierre & Masters (1996); it describes the six levels of Bloom's Taxonomy and gives examples of the kinds of activities that assess each of these learning levels.

Table 1-1 Bloom's Taxonomy with examples[1]

Knowledge - remembering previously learned material. This may involve the recall of a wide range of material, from specific facts to complete theories, but all that is required is the bringing to mind of the appropriate information. Knowledge represents the lowest level of learning outcomes in the cognitive domain.
examples: know common terms, know specific facts, know methods and procedures, know basic concepts, know principles
Comprehension - the ability to grasp the meaning of material. This may be shown by translating material from one form to another (words to numbers), by interpreting material (explaining or summarizing), and by estimating future trends (predicting consequences or effects).
examples: understand facts and principles, interpret verbal material, interpret charts and graphs, translate verbal material to mathematical formulae, estimate the future consequences implied in data, justify methods and procedures
Application - the ability to use learned material in new and concrete situations. This may include the application of such things as rules, methods, concepts, principles, laws, and theories. Learning outcomes in this area require a higher level of understanding than those under comprehension.
examples: apply concepts and principles to new situations, apply laws and theories to practical situations, solve mathematical problems, construct graphs and charts, demonstrate the correct usage of a method or procedure.
Analysis - the ability to break down material into its component parts so that its organizational structure may be understood. This may include the identification of parts, analysis of the relationship between parts, and recognition of the organizational principles involved. Learning outcomes here represent a higher intellectual level than comprehension and application because they require an understanding of both the content and the structural form of the material.
examples: recognise un-stated assumptions, recognise logical fallacies in reasoning, distinguish between facts and inferences, evaluate the relevancy of data, analyse the organizational structure of a work (art, music, writing).
Synthesis - the ability to put parts together to form a new whole. This may involve the production of a unique communication (theme or speech), a plan of operations (research proposal), or a set of abstract relations (scheme for classifying information). Learning outcomes in this area stress creative behaviours, with major emphasis on the formulation of new patterns or structure.
examples: write a well organized theme or creative short story, poem, or music, give a well organized speech, propose a plan for an experiment, integrate learning from different areas into a plan for solving a problem, formulates a new scheme for classifying objects, or events, or ideas.
Evaluation - the ability to judge the value of material (statement, novel, poem, research report) for a given purpose. The judgments are to be based on definite criteria. These may be internal criteria (organization) or external criteria (relevance to the purpose) and the student may determine the criteria or be given them. Learning outcomes in this area are highest in the cognitive hierarchy because they contain elements of all the other categories, plus conscious value judgments based on clearly defined criteria.
examples: judge the logical consistency of written material, judge the adequacy with which conclusions are supported by data, judge the value of a work (art, music, writing) by the use of internal criteria, judge the value of a work (art, music, writing) by use of external standards of excellence.

[1] taken from Carneson, Delpierre, & Masters (1996).

Multiple choice questions (MCQs) are the most common form of question used in CAA (Stephens & Mascia, 1997; Bull, 1999; McKenna, 2001; Warburton & Conole, 2003a). However, researchers and educators (Bull, 1999; Croft, Danson, Dawson & Ward, 2001 p. 13; McKenna, 2001; Davies, 2002; Warburton & Conole, 2003a; Conole & Warburton, 2005) question whether MCQs can adequately assess HOL and consequently doubt that CAA can assess HOL (Sim & Holifield, 2004).

Sim & Holifield (2004) claim "CAA still suffers from a perceived inability to test 'deep learning' in a higher education context…". Many of those educators who question whether CAA can assess HOL believe that "CAA is almost synonymous with multiple-choice testing" (McKenna, 2001). However, there are a variety of opinions on the suitability of MCQs to assess HOL: some believe they are inappropriate (NCFOT, 1998), others believe MCQs are appropriate for assessing only lower-order levels (Farthing & McPhee, 1999), and others believe MCQs are appropriate for assessing HOL if "sufficient care" is taken in their construction (Duke-Williams & King, 2001).

Brown, Bull & Pendlebury (1997 p. 44) contrast MCQs and essays as follows:

> "Broadly speaking, essays are better at estimating understanding, synthesis and evaluation than multiple choice questions whereas multiple choice questions are better at sampling a wider range of knowledge."

The assessment literature tends to agree that essays are the best way to assess HOL. Race (1995) stated "…essays can reflect the depth of student learning. Writing freely about a topic is a process which demonstrates understanding and grasp of the material involved". Brown, Bull & Pendlebury (1997 p. 59) made the following claim about essays: "A good case could be made for arguing that they are the most useful way of assessing deep learning." One paper argued "…in the majority of instances Synthesis and Evaluation promote divergent thinking and answers cannot be determined in advance" (Sim, Holifield & Brown, 2004 p. 218). Conole & Burton (2005) agreed with Sim, Holifield, & Brown and added that "divergent assessment has traditionally relied on longer written answers or essays".

The classic book that introduced what has come to be known as Bloom's taxonomy (Bloom, Engelhart, Furst, Hill & Krathwohl, 1956) implied that objective questions are not suitable to assess Synthesis and Evaluation, the top two levels of the taxonomy. It stated (1956 p. 106) that "essay exercises can be used in the evaluation of interpretation ability" (which is a type of Comprehension) but adds "… objective exercises can also be used in the evaluation." For the level of Analysis, Bloom states "The student may show his ability by making a series of free or guided responses, or by selecting the best answers to objective questions (1956 p. 149)". But he makes no mention of objective questions being suitable for assessing Synthesis or Evaluation. Bloom states that Synthesis has an "emphasis on uniqueness and originality" and that a student "must draw upon elements from many sources and put these together into a structure or pattern not clearly there before" (1956 p. 149). Evaluation, according to Bloom (1956 p. 185) includes the lower levels of the taxonomy but "what is added are criteria including values."

The previous paragraphs show the wide-spread perception that essays and short answers are appropriate for testing HOL. Unfortunately, marking these types of questions is difficult, time-consuming and prone to error. Race (1995) claimed:

> "Essays are demonstrably the form of assessment where the dangers of subjective marking are greatest. Essay-marking exercises at workshops on assessment show marked differences between the mark or grade that different assessors award the same essay – even when equipped with clear sets of assessment criteria".

Brown, Bull & Pendlebury (1997 p. 60) agreed and suggested some reasons:

> "Essay questions are deceptively easy to set and disturbingly hard to mark objectively. At the very least one needs an idea of what counts as a good answer, an indifferent answer and a poor answer. One also needs to know one's values and to be able to distinguish between views that are only different from one's own and those that are both different and wrong."

The previous source (Brown, Bull & Pendlebury, 1997 p. 65) added "… differences in marks can owe more to variations among examiners than to the performance of students" and that "variations among examiners can be high" (1997 p. 46).

Brown, Bull & Pendlebury (1997 p. 203) reported that "Computers cannot yet cope with creative essays, short answer questions that are capable of a wide range of interpretations, or problems that have several potential pathways." Preston & Shackelford (1999) suggested that automated essay marking could be used to assess HOL but that such systems required "significant resources in terms of skills and time". Whittington & Hunt (1999) went so far as to claim that "The automated marking of student's essays is regarded by many as the Holy Grail of computer aided assessment".

Various CAA systems have gone beyond marking MCQs to assess HOL. CourseMarker is an automated assessment tool for marking programs (http://www.cs.nott.ac.uk/~ceilidh/). E-rater (Burstein, Chodorow & Leacock, 2003) grades general knowledge essays, AutoTutor (Wiemer-Hastings, Graesser & Harter, 1998) is a tutoring system that evaluates short answers, IAT (Mitchell, Aldridge & Broomhead, 2003) has a system that marks short answers, and the Intelligent Essay Assessor (IEA) (Landauer, Laham & Foltz, 2003) marks essays. It is unfortunate for researchers that all of these systems are proprietary, and thus one can find only incomplete information about them (Landauer, Laham & Foltz, 2003 p. 296). Some of the open questions about these proprietary systems are:

- What algorithms does the CAA system use for assessment?
- How close does the CAA system come to matching human markers and how was the evaluation carried out?
- How much human effort is required to set up and train the CAA system?

IEA is an automatic assessment system based on Latent Semantic Analysis (LSA) (Landauer, Laham & Foltz, 2003) that has been used to mark essays. Although incomplete in all the necessary details, the literature concerning LSA (see Chapter 2) suggests that LSA can be used to assess short answers.

1.2 Motivation for the research

Having presented the ideas in subsection 1.1, I can now motivate the research undertaken for this dissertation. To summarise the previous points of this chapter:

- Assessment is important for learning.
- Rapid feedback is desired by students.
- Assessment is time-consuming for educators.
- CAA is perceived to save time and therefore costs.
- MCQs are the most common form of CAA.
- Educators doubt the suitability of MCQs for evaluating higher order learning and prefer essays.
- Essays are notoriously difficult to mark objectively.
- A few systems exist, although they are proprietary, that attempt to assess higher order learning by marking essays and short answers.
- LSA might be a useful CAA tool to assess short answers in the domain of computer science.

The motivation for this research was to contribute to the field of automatic assessment by investigating a particular type of CAA capable of assessing short answers – one based on Latent Semantic Analysis (LSA) (see Chapter 2). I chose LSA for a pragmatic reason: The Open University (OU) offered me a studentship to pursue this work in the context of a European project called ELeGI[1], whose purpose was to use the Grid Infrastructure to improve learning. The ELeGI project members were interested in LSA as an assessment technique because the Grid was seen as a solution to the heavy computational requirements of using LSA. The OU bid

[1] See www.elegi.org

proposed using LSA because it had been used successfully in the past to mark general knowledge essays (Landauer, Foltz & Laham, 1998) and a pilot study (Thomas, Haley, De Roeck & Petre, 2004) showed it had promise in our department's area of short answers in the domain of computer science.

1.3 Research questions

The motivation for the research and the preliminary literature review led to the following research questions. (It might be necessary to read Chapter 2 to understand them fully.) The contributions of this dissertation are answers to these questions (see Chapter 8).

- What does the literature say about LSA?
- To what extent can LSA be used to assess short answers in the domain of computer science (CS)?
- How can LSA results be reported to other researchers?
- What questions should be asked by those interested in adopting a CAA?
- How hard is it to build an LSA based CAA?
- How inaccurate are human markers?
- How accurate, compared to human markers, is LSA when used to assess short answers in the CS domain? How do you measure accuracy?
- Given a suitable metric, how can you evaluate an LSA based CAA?
- What calibrations need to be made to an LSA-based marking system to assess short answers in the CS domain?
    - Corpus related questions
        - On what corpus should the LSA system be trained?
        - What is a good size for the corpus?

- Pre-processing questions
  - Does it help to remove stop words?
  - Does stemming help?
  - Will using compound nouns improve the performance of LSA?
- What number in the dimension reduction step gives the best results?
- Which weighting function gives optimum results - log-entropy or tfidf?
- Does proportional averaging improve results?
- Does varying the amount of training data improve results?

1.4 Roadmap for the dissertation

This dissertation documents the work undertaken to answer the research questions and provides some answers to those questions. The following paragraphs give the major steps.

Chapter 2 describes the literature review and research taxonomy from which I developed and refined the research questions. The taxonomy (Appendix A) reveals gaps in the literature that the research questions address.

A major effort for this dissertation was locating an evaluation metric. Calibrating the marking system is a major part of using LSA. I needed an appropriate metric to evaluate and quantify the results of literally hundreds of experiments. Chapter 3 details this work and the metric I chose to evaluate my CAA system – AC1. To my knowledge, my work is the first time AC1 has been used in the field of CAA.

Chapter 4 reports the results of a study undertaken to quantify how consistent human markers are. A CAA system needs to be only as consistent with human markers as they are with each other. Another way of putting this idea is that if one looks at a set of marks given by humans and the CAA, one cannot tell which marks were given by the CAA. This chapter reports the number and variety of questions used for the study, explains the source of the test data, describes the human markers in the study, and quantifies the consistency of these markers.

Chapter 5 presents a framework for describing and evaluating CAA systems. The state of knowledge about CAA would be improved if researchers were able to share each others' experience in a meaningful way. It is difficult to compare research efforts and existing systems because there is no uniform procedure for reporting CAA results. My framework fills that gap by providing a coherent, compact, and comprehensive outline for reporting on and evaluating automated assessment tools. Although this dissertation uses an LSA-based CAA, the framework can be used for other types of CAA systems.

Chapter 6 describes EMMA (ExaM Marking Assistant), an LSA-based assessment tool I developed to test how well LSA could perform in the domain of computer science. A major requirement of using LSA is a substantial training corpus. Chapter 6 describes the EMMA database used for the training corpus and the marking engine that together comprise the EMMA CAA.

Chapter 7 describes a series of experiments to improve the performance of EMMA and presents an evaluation of EMMA.

Chapter 8 compares EMMA's results with the human markers discussed in Chapter 4. It lists future research followed by conclusions.

Chapter 2. LSA – Method and Related Work

2.1 Introduction

This chapter describes Latent Semantic Analysis (LSA) and how it can be used to assess short answers.

2.1.1 Recap of Chapter 1

Chapter 1 presented the evidence for the motivation of the research conducted for and described in this dissertation. It summarised the motivation as follows:

- Assessment is important for learning.
- Rapid feedback is desired by students.
- Assessment is time-consuming for educators.
- Computer Assisted Assessment (CAA) is perceived to save time and therefore costs.
- Multiple Choice Questions (MCQs) are the most common form of CAA.
- Educators doubt the suitability of MCQs for evaluating higher order learning and prefer essays.
- Essays are notoriously difficult to mark objectively.
- A few systems exist, although they are proprietary, that attempt to assess higher order learning by marking essays and short answers.
- LSA might be a useful tool to assess short answers in the domain of computer science.

2.1.2 Non-technical overview of how LSA can mark short answers

The problem of assessing a short answer can be equated to the problem of locating within a set of previously-human-marked-answers that particular answer that is closest to the answer-being-marked. In other words, which marked answer resembles most closely the student answer? Or in, yet again, other words, how can I retrieve the marked answer that is the closest match to the student answer? When I have located that closest answer, I can award its mark to the student answer-being-marked.

The reasoning in the previous paragraph leads to the hypothesis that LSA can be a useful tool to mark answers. LSA has its roots in Information Retrieval (IR). Marking an answer involves *retrieving* the *answer* (along with its *mark)* that is closest to the answer-being-marked and then *assigning* the retrieved mark to the answer-being-marked. (In practice, a number of close answers are retrieved and the average of their individual marks is assigned to the answer-being-marked.) A basic IR technique that simply matches words is not appropriate because it will fail to retrieve answers containing words that are similar, but not identical, in meaning. LSA was designed to incorporate the concept that different words can represent the same idea (synonymy) in addition to the concept that one word can have different meanings (polysemy) (Dumais, 2007 p. 294). Another way of stating these concepts is that the meaning of a word can be found within its context.

The mathematical underpinnings can be daunting to non-mathematicians (Hu, Cai, Wiemer-Hastings, Graesser & McNamara, 2007 p. 407). They state "Why and how it works is a very deep mathematical/philosophical question … " and point the interested reader to Martin and Berry (2007). What follows is a non-mathematician's understanding of how LSA can retrieve similar answers. (Section 2.2.1 provides a more formal description.) The LSA technique is essentially a method for solving a huge set of simultaneous equations that represent terms in documents (Landauer, 2007 pp. 13-14). For this dissertation, the terms are all of the words contained in all of the documents; the documents are marked answers to a question in the domain of Computer Science and paragraphs from CS textbooks.

A single equation represents the number of times each word in the set of words-to-be-considered appears in an answer. The answer-being-marked can be thought of as a query. The totality of these equations describe the answer-being-marked in addition to the set of previously-marked-answers. The equations that follow show the idea.

$$A_1 = c_{11} * w_1 + c_{12} * w_2 + \ldots + c_{1n} * w_n$$

$$A_2 = c_{21} * w_1 + c_{22} * w_2 + \ldots + c_{2n} * w_n$$

$$A_m = c_{m1} * w_1 + c_{m2} * w_2 + \ldots + c_{mn} * w_n$$

$$Q = c_{q1} * w_1 + c_{q2} * w_2 + \ldots + c_{qn} * w_n$$

where

A_i - the *i*th answer in the set of previously human-marked set of answers

w_j - the *j*th word in the set of all of the words in all of the answers

c_{ij} - the count, for answer *i,* of the *j*th word in the set of words in all of the answers; that is, word *j* appears *c* times in answer *i*; in practice, the count is modified by a weighting function

Q - the query document, or the answer-to-be-marked.

LSA takes the equations above and creates a matrix whose columns represent the previously-marked-answers as well a paragraphs from CS textbooks and whose rows represent the words. This matrix represents all of the answers in the set of previously marked answers and all of the words in the set. LSA then uses Singular Value Decomposition (SVD), which is a mathematical technique that transforms a possibly non-square matrix into three smaller matrices. By decreasing the dimensionality of one the matrices and then re-multiplying the three matrices, you get a smaller matrix that retains the most characteristic aspects of the answers and the words they contain. And most importantly for marking answers, the new, smaller matrix represents the answers in terms of words that did not appear in the original answer at all (Landauer, 2007 p. 15). This consequence of SVD followed by dimension reduction is the key to the ability of LSA to mark answers. It can retrieve marked answers that contain semantically

similar words to the answer-to-be-marked, not just the *same* words. An answer to a question about which breed of dog won the American Kennel Club competition in 1976 that contains, for example, the word *Alsatian*, could be retrieved by a query that includes the words, *German Shepherd*. But would LSA retrieve an answer containing *Alsatian wine* or *residents of Alsace-Lorraine*? No, because the set of previously-marked-answers about dogs and competitions would not, statistically speaking, contain many references to *wine* or the region of *Alsace-Lorraine*.

Having read this intuitive description of how LSA can be used to mark essays, the reader can move to the more theoretical description that follows.

2.2 Introduction to formal description of LSA

Latent Semantic Analysis (LSA) is a statistical method for capturing meaning from a text. A seminal paper (Landauer, Foltz & Laham, 1998) gives a more formal definition: "Latent Semantic Analysis is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text". LSA was first used as an information retrieval technique in the late 1980s, when it was called Latent Semantic Indexing (LSI) (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990). Later, the developers found that LSI could be useful to analyse text and created the term LSA to describe LSI when used for this specialised area.

By 1997, Landauer and Dumais (1997) asserted that LSA could serve as a model for the human acquisition of knowledge. They claimed that LSA solves Plato's problem, that is, how do people learn so much when presented with so little? The answer, oversimplified but essentially accurate, is an inductive process: LSA "induces global knowledge indirectly from local co-occurrence data in a large body of representative text" (Landauer & Dumais, 1997).

2.2.1 The LSA method

To use LSA, researchers amass a suitable corpus of text. They create a term-by-document matrix where the columns represent documents and the rows represent terms (Deerwester,

Dumais, Furnas, Landauer & Harshman, 1990). A term is a subdivision of a document; it can be a word, phrase, or some other unit. A document can be a sentence, a paragraph, a textbook, or some other unit. In other words, documents contain terms. The elements of the matrix are weighted term counts of how many times each term appears in each document. More formally, each element, a_{ij} in an $i \times j$ matrix is the weighted count of term i in document j.

LSA decomposes the matrix into three matrices using Singular Value Decomposition (SVD), a well-known technique that is the general case of factor analysis, which is restricted to a square matrix. SVD is able to decompose a rectangular matrix, i.e., the number of rows and columns may be of any size (Miller, 2003). Following Deerwester et. al., (1990) the process is:

Let t = the number of unique terms in the corpus - rows in the matrix

d = the number of documents in the corpus - columns in the matrix

X = a t by d matrix

Then, SVD calculates the 3 matrices TSD such that $X = TSD^T$, where

m = the number of dimensions, m <= min(t,d)

T = a t by m matrix

S = an m by m diagonal matrix, i.e., only diagonal entries have non-zero values

D^T = an m by d matrix

Up to this point, LSA is just the vector space method of information retrieval (Salton, Wong & Yang, 1975). The LSA innovation is to reduce S, the diagonal matrix created by SVD, to an appropriate number of dimensions k, where $k << m$, resulting in S'. The product of $TS'D^T$ is the least-squares best fit to X, the original matrix (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990): $X' = TS'D^T \approx X$. Figure 2-1 explains the LSA method using diagrams.

The literature often describes LSA as analysing co-occurring terms. Landauer and Dumais (1997) argue that it does more; they explain that the new matrix captures the "latent transitivity relations" among the terms. Terms not appearing in an original document are represented in the

new matrix as if they actually were in the original document (1997). Landauer and Dumais (1997) consider LSA's ability to induce transitive meanings to be especially important given their finding that fewer than 20% of paired individuals will use the same term to refer to the same common concept (Furnas, Gomez, Landauer & Dumais, 1982).

LSA exploits what I think of as the transitive property of semantic relationships:

let → stand for *is semantically related to*

If A→B and B→C, then A→C.

However, the similarity of the transitive property of semantic relationships to the transitive

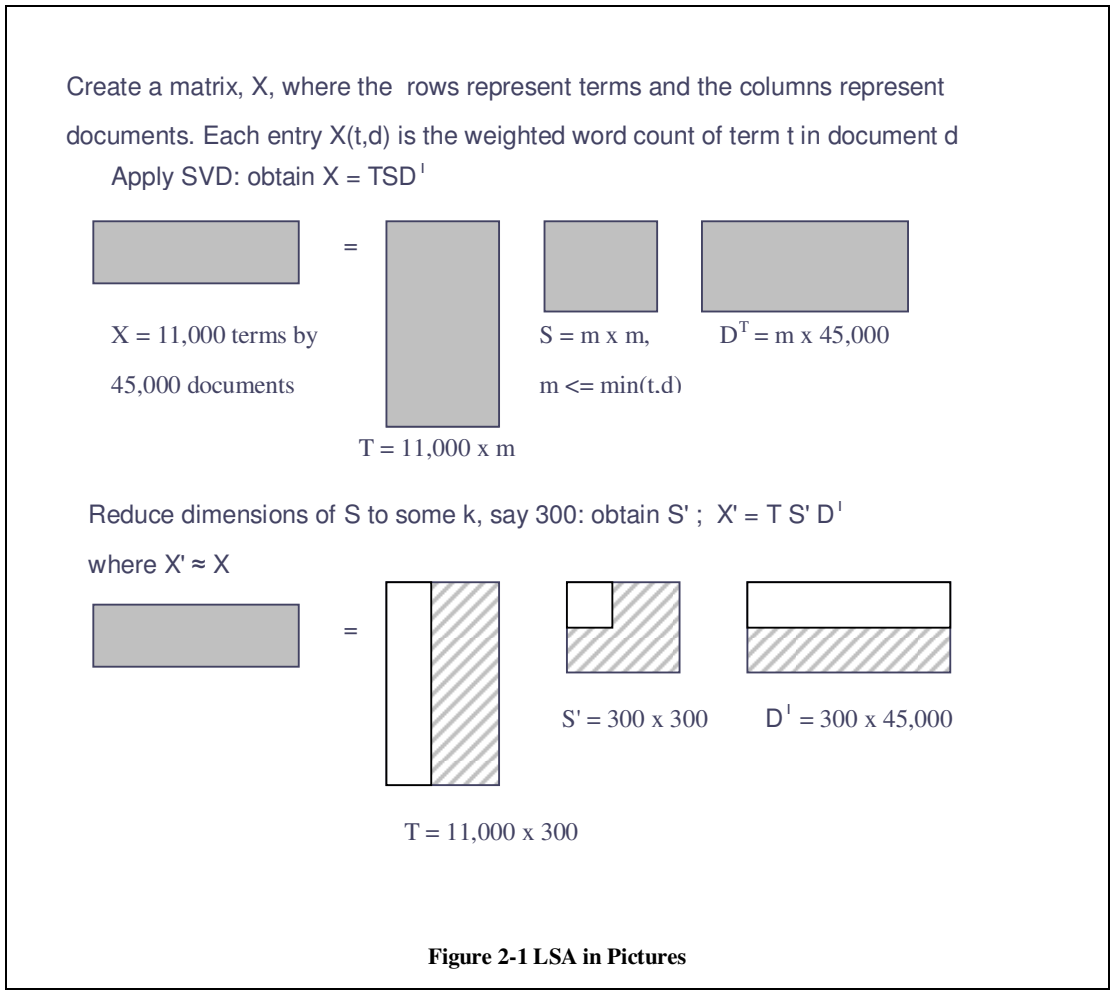


Figure 2-1 LSA in Pictures

property of equality is not perfect. Two words widely separated in the transitivity chain can have a weaker relationship than closer words. For example, LSA might find that copy → duplicate → double → twin → sibling. *Copy* and *duplicate* are much closer semantically than *copy* and *sibling*. LSA explicitly handles this variation in similarity by using a similarity measure, the cosine. Words and documents are represented in the LSA space as vectors. The higher the cosine between two vectors, the closer is the meaning of the underlying terms or documents.

Finding the correct number of dimensions for S' is critical; if it is too small, the structure of the data is not captured. Conversely, if it is too large, sampling error and unimportant details remain, e.g., grammatical variants (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990; Miller, 2003; Wade-Stein & Kintsch, 2003). Empirical work yields different figures for the correct number of dimensions. Some researchers report that about 300 dimensions is correct (Landauer & Dumais, 1997; Foltz, Laham & Landauer, 1999; Kintsch & Bowles, 2002; Wade-Stein & Kintsch, 2003). Others report that they achieve better results using other numbers including 10 (Nakov, Valchanova & Angelova, 2003), 50-150 (Furnas, Deerwester, Dumais, Landauer, Harshman, Streeter & Lochbaum, 1988) and 1500 (Landauer, Laham, Rehder & Schreiner, 1997). The figures for the optimum number of dimensions range from 10 to 1500. Note that Landauer reports two different figures in the same year: 300 (Landauer & Dumais, 1997) and 1500 (Landauer, Laham, Rehder & Schreiner, 1997). Because of the lack of consistency within these results, any individuals developing an LSA-based marking system must experiment to determine the number of optimum dimensions themselves.

Training the system, i.e., creating the matrices from a huge corpus of training data using SVD and reducing the number of dimensions, requires massive computing power. It was reported in 2003 that it can take hours or days to complete the processing (Miller). For example, in 2003 (on unspecified hardware) it took about five hours to process a huge corpus of 500 million words comprising 2.5 million documents and 725,000 unique words (Dennis, Landauer, Kintsch & Quesada, 2003). Fortunately, once the training is complete, it takes just seconds for

LSA to evaluate a text sample (Miller, 2003). Foltz, Laham, & Landauer (1999) reported that it took 20 seconds to calculate a mark. The shorter time for marking an answer once the training is complete reflects the fact that the SVD step is the major time-consuming piece of performing LSA. Once the SVD has finished, marking an answer involves simply converting it to the same space as the training data and then comparing it to the columns in the reduced matrix created by LSA.

2.3 Using LSA for assessment

2.3.1 The relationship to Information Retrieval

This subsection describes how LSA is used for assessment. One simple way to think about it is to recall that LSA was developed originally as a tool for information retrieval (IR). The group of previously-graded-answers in the training data is the database of documents to be searched. The answer-to-be-graded is the query. The LSA-based CAA system searches the database and retrieves a pre-determined number of documents (previously-marked-answers) that are *close enough* to the query (answer-to-be-graded), and assigns a grade based on a weighted average of the grades previously awarded. LSA uses a similarity measure to determine how close the answers are and assigns a threshold value to decide if the answers are *close enough*.

2.3.2 The mathematics

The SVD step and the reduction to k dimensions generate what is known as an LSA semantic space (Foltz, Laham & Landauer, 1999). In order to compare an answer-to-be-graded to the previously-marked-answers, it must be transformed to the same semantic space. First a query vector q of length t terms, is created of weighted word counts, i.e., each value q_i is the number of times the term t_i exists in the answer-to-be-graded modified by the same term weighting function used in the matrix of weighted word counts. Then q is transformed into a pseudo-document, $\hat{q}$, by Equation 2-1 (Berry, Dumais & O'Brien, 1995):

Equation 2-1 Transforming a query into a pseudo-doc $\hat{q} = q^T T_k S_k^{-1}$

The derivation for Equation 2-1 comes from (Dennis, Landauer, Kintsch & Quesada, 2003). In their notation, $[X:X_q]$ is the matrix X with the query document appended at the end.

$[X : X_q] = TS[D : D_q]^T$

$T^T[X : X_q] = S[D : D_q]^T$ (since T and D are orthonormal, $TT^T = I$, the identity matrix)

$S^{-1}T^T[X : X_q] = [D : D_q]^T$ (since S is a diagonal matrix, $S^{-1}S = I$)

$[D : D_q] = [X : X_q]^T TS^{-1}$

$D_q = X_q^T TS^{-1}$

At this point, $\hat{q}$ can be compared with every column (previously-graded-answer) in the D matrix scaled by the singular values in S_k. The column of D that is closest to $\hat{q}$ is used to mark the answer. The cosine or the dot product can be used as the similarity measure (Furnas, Deerwester, Dumais, Landauer, Harshman, Streeter & Lochbaum, 1988; Deerwester, Dumais, Furnas, Landauer & Harshman, 1990; Berry, Dumais & O'Brien, 1995). (See Equation 2-2 for the cosine similarity measure, where $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots y_n)$ are two n-dimensional vectors.) The cosine is preferred over the dot product because the cosine counteracts the effect of vector length on the distance calculated by the dot product by normalising the vectors. Berry, Dumais & O'Brien (1995 p. 10) suggest a threshold of 0.9, that is, only documents with a cosine of at least 0.9 should be considered *close enough* to the query (answer-to-be-graded).

Equation 2-2 The Cosine Similarity Measure $$\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x^2{}_i \sum_{i=1}^{n} y^2{}_i}}$$

2.4 Introduction to the research taxonomy

A major product of this dissertation is the research taxonomy (see Appendix A) resulting from an in-depth, systematic review of the literature concerning Latent Semantic Analysis (LSA) research in the domain of educational applications. The taxonomy presents the key points from a representative sample of the literature in a format that is comprehensive, relatively compact, and useful to other researchers. It exposes 5 main research themes (see subsection 2.6.1) and emphasises the point that even after more than 15 years of research, some details needed to build an assessment system based on LSA remain unreported. Researchers and developers implementing LSA-based educational applications will benefit by studying the taxonomy because it brings to one place the techniques and evidence reported in the vast LSA literature.

I realized the need for a taxonomy while building EMMA, an LSA-based assessment system for use in computer science courses. Although the original assessment results were encouraging, they were not good enough for the intended task of summative assessment (Thomas, Haley, De Roeck & Petre, 2004). I conducted a comprehensive, in-depth literature review to find techniques to improve my system. I wanted to assimilate all of the LSA literature and fully understand the state of the art in LSA theory to improve my system. The taxonomy documents my findings and supports the insights gained by studying the literature.

There are well over a hundred published papers on LSA[3]. Some of them involve educational uses e.g., Steinhart (2001); some concentrate on LSA theory e.g., Landauer and Dumais (1997); and some of the newer papers suggest applications of LSA that go beyond analysing prose e.g., Quesada, Kintsch & Gomez (2001) look at complex problem solving and Marcus, Rajlich & Maletic (2004) study document to source code traceability.

[3] Benoit Lemaire maintains a website (http://www-timc.imag.fr/Benoit.Lemaire/lsa.html) listing over 75 LSA-related papers and gives the homepages of some of the major LSA researchers.

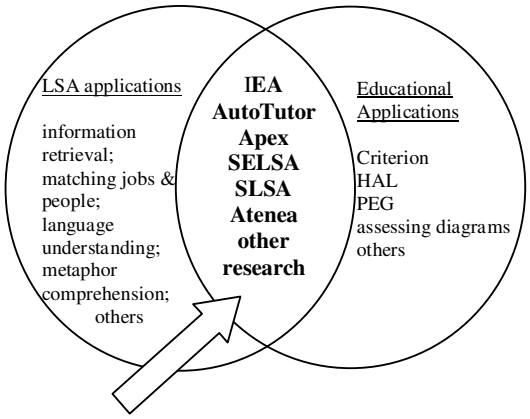


Figure 2-2 Scope of the taxonomy - the intersection of LSA and educational applications

2.4.1 The scope of the research taxonomy

The taxonomy summarises and highlights important details from the LSA literature (See Figure 2-2). Because the literature is extensive and my interest is in the assessment of short answers, the taxonomy includes only those LSA research efforts that overlap with educational applications. Therefore, LSA research into such areas as information retrieval (Nakov, Valchanova & Angelova, 2003), metaphor comprehension (Lemaire & Bianco, 2003) and source code analysis (Marcus & Maletic, 2003) do not appear in the taxonomy. Similarly, the taxonomy ignores various non-LSA techniques that have been used to assess essays (Burgess, Livesay & Lund, 1998; Burstein, Chodorow & Leacock, 2003) and diagrams (Anderson & McCartney, 2003; Thomas, Waugh & Smith, 2005).

The next subsections discuss the rationale for choosing certain articles over others and the meaning of the headings in the taxonomy.

2.4.2 Method for choosing papers

The literature review found 150 papers of interest to researchers in the field of LSA-based educational applications. In order to reduce this collection to a more reasonable sample, I

constructed a citer – citee matrix of articles. That is, each cell entry (i,j) was non blank if article i cited article j. I found the twenty most-cited articles and placed them, along with the remaining 130 articles, in the categories shown in Table 2-1.

I chose the twenty most-cited articles for the taxonomy. Some of these most-cited articles were early works explaining the basic theory of Latent Semantic Indexing (LSI). Although not strictly in the scope of the intersection of LSA and educational applications, a representative sample of these articles appear in the taxonomy because of their seminal nature. Next, I added articles from the category that combined educational applications with LSA that were of particular interest, either because of a novel domain or technique, or an important result. Finally, I decided to reject certain heavily cited articles because they presented no new information pertinent to the taxonomy. This left 28 articles in the taxonomy.

Table 2-1 Categories of articles in the literature review and those that were selected for the taxonomy

Type of Article	Number in Literature Review	Number in Taxonomy
most cited	20	13
LSA and educational applications	43	15
LSA but not educational applications	13	0
Latent Semantic Indexing	11	0
theoretical / mathematical	11	0
reviews / summaries	11	0
educational applications but not LSA	41	0
Total	**150**	**28**

2.5 The taxonomy categories

The taxonomy organises the articles involving LSA and educational applications research into three main categories: an *Overview*, *Technical Details*, and *Evaluation*. Figure 2-3 through Figure 2-5 describe the headings and sub-headings.

2.5.1 Category A: Overview

The *Overview* (Figure 2-3) is an "at a glance" summary of an article. The *System Name* column shows the name of a system or technique and gives the URL if available. It is blank if the work described in the article is unnamed. The *Reference* column contains a pointer to the references section at the end of this dissertation. The entries are of the form XXXnn where

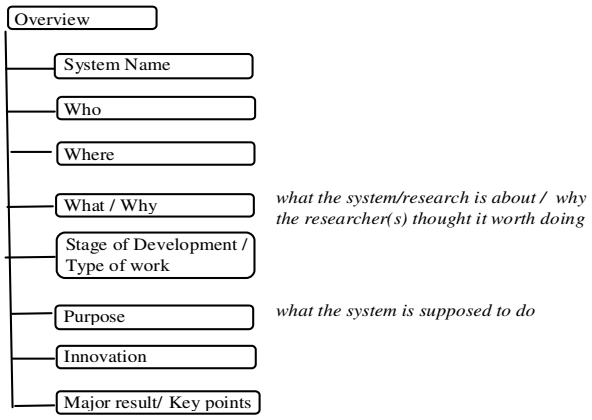


Figure 2-3 Category A: Overview

XXX are the initials of up to three of the authors. If capitalised, they represent different authors; if the first is capitalised and the second two are lower case, the article has one author. The lower case letters, nn, stand for the 2-digit year of publication. The *Who* and *Where* columns give the authors' last names and their affiliations, which can be useful to trace the development of an idea or system. The *What/Why* column briefly explains what the work is about and why the authors considered it worth doing. The next column, S*tage of development / Type of work*, indicates whether or not the system is a deployed application and gives a synopsis of the work

described in the article. *Purpose* is what the project or system is attempting to do. *Innovation* explains what is new about the work described in the paper. The final column, *Major result / Key points*, is a short summary of the important outcomes reported on in the paper.

2.5.2 Category B: Technical Details

The second major category (Figure 2-4), *Technical Details*, has three subcategories: *Options*, *Corpus*, and *Human Effort*. *Options* refers to the choices that an LSA researcher must make when implementing the LSA algorithm. The necessity of making these choices leads some researchers to call LSA an art (Nakov, 2000). *Pre-processing* is anything done to the text before running it through the system, e.g., stemming and spelling correction. The number of dimensions in the reduced LSA matrix is one of the crucial implementation choices. It is in the column *# dimensions*. The *Weighting function* column gives the method chosen by the author to indicate the relative importance of the term-counts in the matrix. Nakov, Valchova & Angelova

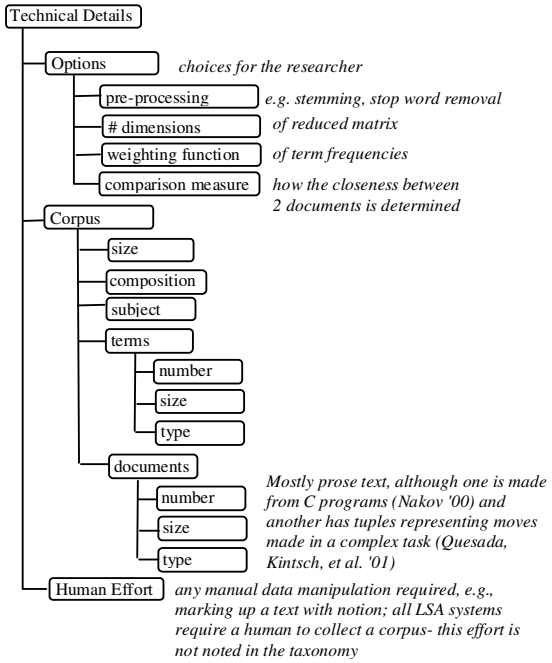


Figure 2-4 Category B: Technical Details

(2003) claim that the weighting function is one of the most important choices. *Comparison measure* shows how the system determines the closeness between two documents.

Corpus, the next subsection in *Technical Details*, comprises three more subsections: *Subject*, *Terms*, and *Documents,* as well as *size* and *composition* of the corpus as a whole. The S*ubject* column shows the main topics covered by the corpus. *Terms and Documents* show similar information: the number of terms and documents in the weighted word count matrix, the size of the terms and documents, and the type. Most of the corpora are prose text, although one is made from C programs (Nakov, 2000) and another has tuples representing moves made in a complex task (Quesada, Kintsch & Gomez, 2001).

Human effort describes any manual data manipulation required before using the LSA assessment system. Computerised pre-processing, such as stemming, is not listed under *human effort*. All LSA systems require human intervention to collect a corpus; this effort is not noted in the taxonomy.

2.5.3 Category C: Evaluation

Evaluation (Figure 2-5) explains how three types of system appraisal are done: *accuracy*, *effectiveness*, and *usability*. *Accuracy* pertains to how well the LSA system works. The subsections are *method used*, *granularity of marks*, *item of interest*, *number of items assessed*, and *results*. These categories apply only when the system assesses some kind of artefact. Not all of the articles in the taxonomy evaluate artefacts – those cells are shaded out. *Method used* describes how the researchers evaluated their system and which success measure they used. *Granularity* gives the maximum number of possible marks, which is important because the finer the granularity, the harder it is to match human markers. The *Item of interest* is the artefact to be marked, e.g., an essay or a short answer. The *Number of items* assessed is the number of questions or essays involved, that is, if one exam has ten essays and 300 students complete the exam, then the *Number of items* assessed is 3,000. The *Results* subsection compares the average correspondence between the marks given by human graders to the correspondence between the LSA system and the average human to human correspondence. To be successful, an LSA-based

assessment system should correspond to human markers as well as human markers correspond with each other.

The final two subsections are *effectiveness* and *usability*. *Effectiveness* refers to whether or not the system improved the learning of students using the system. *Usability* refers to how easy the system is to use. Both of these terms apply only to those articles that describe a deployed system. The cell in the taxonomy is shaded for articles that describe other kinds of research.

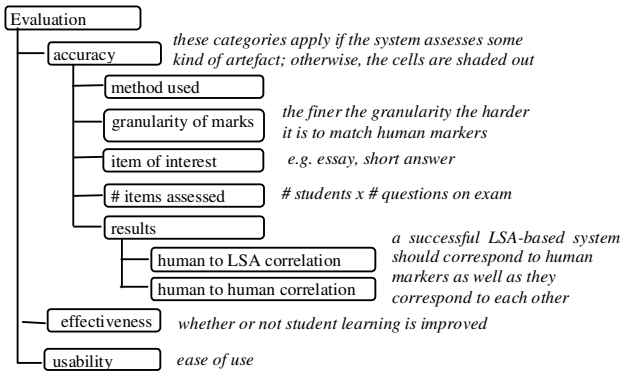


Figure 2-5 Category C: Evaluation

2.5.4 How to read the research taxonomy

Appendix A presents the taxonomy. When looking at it, the reader should keep a few points in mind. First, the taxonomy is three pages wide by three pages high. Pages A1 - A3 cover the overview for all of the articles in the taxonomy. Pages A4 - A6 list the technical details. Pages A7 - A9 give the evaluation information. Second, each line presents the data relating to one study. However, one article can report on several studies. In this case, several lines are used for a single article. The cells that would otherwise contain identical information are merged. Third, the shaded cells indicate that the data item is not relevant for the article being categorised. Fourth, blank cells indicate that I was unable to locate the relevant information in the article. Fifth, the information in the cells was summarised or taken directly from the articles. Thus, the *Reference* column on the far left holds the citation for the information on the entire row.

Organising a huge amount of information in a small space is not easy. The taxonomy in the appendix is based on an elegant solution in (Price, Baecker & Small, 1993).

2.6 Discussion

This section discusses the insights revealed by the taxonomy. Sections 2.6.1 and 2.6.2 describe what can be found in the literature, and section 2.6.3 highlights some of the gaps in the literature.

2.6.1 Main research themes

A researcher has access to a vast literature about LSA and about educational applications. However, the taxonomy reveals 5 main research themes:

1. seminal literature describing the new technique named LSI, which was later renamed LSA when used for semantic analysis rather than strictly information retrieval,
2. attempts to reproduce the results reported in the seminal literature, which for the most part failed to achieve the earlier results,
3. attempts to improve LSA by adding syntax information,
4. applications that analyse non-prose text,
5. attempts to improve LSA by experimenting with corpus size and composition, weighting functions, similarity measures, number of dimensions in the reduced LSA matrix, and various pre-processing techniques – exactly those items in Category B of the taxonomy.

2.6.2 Diversity in the research

The taxonomy reveals variety in the research. Researchers in North America, Europe, and Asia work on both deployed applications and continuing research. They use a wide variety of options for pre-processing techniques, number of dimensions in the reduced matrix, weighting

functions, and composition and size of corpus. They use English, French, Spanish, and Bulgarian corpora. The researchers report their evaluation methods with different specificity.

2.6.3 Gaps in the literature

The great variety of techniques used by researchers mentioned in the previous subsection leads to difficulty in comparing the results. Other researchers need to know all of the details to fully evaluate and compare reported results.

Much information is missing on page 2 of the taxonomy – Category B: Technical Details. These missing data concern the choices researchers must make when they implement their systems. Page 3 of the taxonomy, Category C: Evaluation, shows that some researchers have not evaluated the effectiveness or usability of their deployed systems. Of course, if the system is still in the research phase, evaluation and usability may not apply and thus the cells are shaded.

The *Method used* subheading under *Accuracy* in Category C is a major area for gaps. Although many researchers report correspondences between LSA and human graders, they usually do not mention whether they are using Pearson, Spearman, Kendall's tau, or some other correspondence measure. One of the contributions of this dissertation is the application of a new and not very well known method for evaluating assessment systems. See Chapter 3 for a description, justification, and example of using the AC1 statistic introduced by Gwet (2001a).

The existence of the blank cells in the taxonomy is troubling. They imply that researchers often neglect to report critical information, perhaps due to an oversight or page length restrictions. Nevertheless, the ability to reproduce results would be enhanced if more researchers provided more detailed data regarding their LSA implementations.

2.7 Value of the Taxonomy

The previous subsections demonstrate the value of the taxonomy. First, it captures a large amount of data in a relatively compact form. Second, it provides an easy way to compare and contrast various LSA research papers. Finally, it gives a graphic picture of the gaps in the LSA literature. Table 2-2 quantifies these gaps. The items in the first column are the column

headings of the taxonomy. The second column shows the number of papers that did not reveal the given information. The third column gives the percentage of papers with missing information. (Recall that there are 28 papers in the taxonomy.)

The table shows some surprising gaps. Ten of the items important for understanding a CAA are described by less than half of the 28 papers in the taxonomy. Half of the papers did not discuss a most important piece of data – how the accuracy of their systems compare with human markers. Overall, 40% of the cells in the taxonomy were blank, indicating that many of the papers did not give enough detail to understand LSA on the one hand, and the marking system in question on the other hand.

Table 2-2 Gaps in the literature as revealed by the taxonomy

Item	# of blank cells	% of blank cells
innovation	3	11%
purpose	1	4%
major result	1	4%
pre-processing	14	50%
# dimensions	12	43%
weighting function	18	64%
comparison measure	9	32%
size of training data	17	61%
composition of training data	4	14%
subject	5	18%
term information	10	36%
document information	14	50%
human effort required	14	50%
method	4	14%
granularity	14	50%
item of interest	12	43%
# items assessed	10	36%
human to computer comparison	14	50%
human to human comparison	14	50%
effectiveness	21	75%
use-ability	24	86%
average	11	40%

2.8 Summary of findings from the literature review

The taxonomy revealed that others were having difficulty matching the results reported by the original LSA researchers (Landauer, Foltz & Laham, 1998). I found ambiguity in various critical implementation details (especially weighting function and number of dimensions) as well as unreported details. I speculate that the conflicting or unavailable information explains at least some of the inability to match the success of the original researchers.

I hope that future LSA researchers will keep the taxonomy in mind when presenting their work. Using it will serve two main purposes. First, it will be easier to compare various research results. Second, it will ensure that all relevant details are provided in published articles, which will lead to improved understanding and the continued development and refinement of LSA.

The variability in the results documented in the taxonomy shows that LSA is still something of an art. More than 15 years after its invention, the research issues suggested by Furnas, et. al. (1988) are still very much open.

This chapter described the theory and method of LSA, highlighted the gaps in the literature, and listed the calibrations that need to be made in a CAA based on LSA. In order to judge the effects of these calibrations, one needs an adequate success metric. The next chapter discusses various metrics in use and explains why they were not useful for my purposes. It concludes with an explanation and justification of Gwet's AC1, the inter-rater reliability statistic that I used as a success metric.

Chapter 3. Evaluation Metrics

3.1 Introduction

This chapter discusses methods to evaluate Computer Assisted Assessment (CAA) systems, including some commonly used metrics as well as unconventional ones. I found that most of the methods to measure automated assessment reported in the literature were not useful for my purposes. After much research, I found a new metric, the Gwet AC1 inter-rater reliability (IRR) statistic (Gwet, 2001a), that is a good solution for evaluating CAAs. Section 3.7 discusses AC1, but first I describe other possible metrics to motivate why I think that AC1 is the best available for evaluating an automated assessment system.

I focus on two types of metrics that I label external and internal metrics. External metrics can be used for reporting and sharing results. Internal metrics are used for comparing results within a research project.

Producers of CAAs need an easily understandable external metric to report results to consumers of CAAs, i.e., those wishing to use a particular system. In addition to reporting results to potential consumers, researchers may wish to share their results with other researchers. Finally, and perhaps most important for this dissertation, producers need an internal metric to quickly compare the results of selecting different parameters of the assessment algorithm. Many choices need to be made when implementing an LSA-based marking system. The LSA literature frequently leaves many of these choices unspecified, including number of dimensions in the reduced matrix, amount and type of training data, types of pre-processing, and weighting functions (see the taxonomy in Appendix A). The choice of these parameters is an intrinsic aspect of building an LSA marking system. Therefore, researchers need an adequate

way to measure and compare the results of the various selections, as I shall explore in this chapter.

Section 3.2 describes a simple metric that is often used for external reporting of results. Section 3.4 discusses existing ways to measure the success of LSA-based assessment systems and motivates the need for new metrics. Sections 3.4 and 3.5 discuss several standard statistical tests that could be used to measure the success of automated assessment systems and argue that none of them is suitable for my purposes. Section 3.6 discusses possible metrics that use the distance between two vectors for comparing automated assessment systems - the Manhattan distance (L1) and the Euclidean distance (L2). Finally, Section 3.7 explains and justifies the metric I chose to evaluate EMMA (the LSA-based assessment system created for this dissertation) – the Gwet AC1 inter-rater reliability statistic and discusses how it overcomes the flaws of the better-known kappa statistic.

3.2 A simple metric (SM)

A simple success measure is to determine the percentage of marks where two markers give identical marks. However, this simple metric (SM) gives an incomplete picture of the results. Consider the hypothetical case illustrated in Table 3-1. It shows how closely two markers agree with a third marker assumed to be the gold standard, i.e., the correct answer. Eighty percent of

Table 3-1 Hypothetical results for two markers that show the simple metric of the percent of identical scores for a four-point question hides important details

	Marker A	Marker B
Point Difference between Markers and a "Gold Standard"	% of Questions	
0	80	75
1	0	25
2	0	0
3	5	0
4	15	0

the answers marked by Marker A agreed exactly with the gold standard, 5% differed by 3 points, and 15% disagreed by 4 points. Seventy-five percent of the answers marked by Marker B agreed exactly with the gold standard and 25% differed by 1 point. The SM awards 80% to Marking System A and 75% to Marking System B. Clearly, both markers have a high percentage of agreement, but which is the better marker - A or B? The SM says that A is better than B. However, even though A has a higher percentage of identical answers than does B, the latter has 100% of its marks disagreeing with the human by at most one point while A has only 80% of its marks disagreeing by at most one point and 20% that differ by three or more points. One of the flaws in the SM is that it gives no indication of the spread, or distribution, of the marks. The standard deviation (SD) is the widely known statistic to indicate the spread of values. The better marker is the one that has the highest number of identical marks with the lowest SD.

What happens to the SM if the SD is given to indicate the spread of the marks? Table 3-2 shows the SDs of the two markers from Table 3-1 and adds two more markers. The mean is the average difference between the marker and the Gold Standard. The SD indicates the spread of the differences between the marker and the Gold Standard. The most accurate marker will have the highest number of identical marks compared with the gold standard *combined* with the lowest mean and lowest standard deviation of the difference between the marks. It is not easy to compare these sets of thee numbers to determine the best marker. Markers A and B have SMs of 80 and 75 respectively. Marker B is clearly the better marker and its lower SD seems to support this conclusion. But how can you quantify whether Marker B's SD of .44 is low enough compared to Marker A's SD of 1.52 to justify calling it a better marker when its SM is also lower? Marker C and D are more straight forward. Both of them have a SM of 40 but have SDs of 1.68 and .75 respectively. Their means are 2 and .8 respectively. Is this difference in the means and SDs enough to just that Marker D is better than Marker C given their identical SMs ? It would be very time-consuming to determine a better marker by examining these three numbers. What is needed is a single number that gives the answer. Subsection 3.7.2 shows that Gwet's AC1 statistic is just such a metric.

Table 3-2 Hypothetical results for four markers with Means and Standard Deviations of the differences between the marks

	Marker A	Marker B	Marker C	Marker D
Point Difference between Markers and a "Gold Standard"	% of Questions			
0	80	75	40	40
1	0	25	0	40
2	0	0	0	20
3	5	0	40	0
4	15	0	20	0
Mean	.75	.25	2	.8
Standard Deviation	1.52	.44	1.68	.75

Perhaps the SM is an acceptable external metric to use for reporting results to consumers, but, even with the addition of standard deviations, it is inadequate for internal comparison purposes.

3.3 Attempt to improve the SM

This subsection describes an attempt to overcome the flaws of the SM by enlarging it to include the all of the point differences, not just the ones equal to zero, as in the SM. Table 3-3 shows the results of a study to determine the optimum amount of training data (the amount that gives the best results), which is one of the parameters for calibrating an LSA-based marking system. Note that this study is somewhat different than the hypothetical study shown in the previous section. The previous study attempted to determine the more accurate marker; this study used the same (non-human) marker but varied the amount of training data to determine

the amount of training data that produced results that match most closely with the human marks previously given.

EMMA requires two types of training data - general training data comprising textbooks and specific training data comprising previously-marked-answers. I had 960 previously-marked answers; the study used 333 to be marked and the remaining 627 as specific training data.

The question being studied was worth four points, thus, the worst results occurred when the human marker and the tutor (considered the gold standard) disagreed by ±4 marks.

Table 3-3 shows that when, for example, 10 previously-marked-answers were used as specific training data, 65% of the 333 answers-to-be-marked were marked identically by the computer and by a human.

Finding the optimum amount of training data by studying this table is very difficult. Unfortunately, studying a graph turned out to be no easier. Indeed, this difficulty in interpreting the data was a major motivation for finding an alternative success metric. As stated earlier, I needed a way to compare quickly the results of literally hundreds of experiments.

Table 3-3 Percentages of Agreement between Human and Computer when varying amount of training data

# of Marked Answers	% Equal Scores	Tutor and Computer differ by ±1mark	Tutor and Computer differ by ± 2 marks	Tutor and Computer differ by ± 3 marks	Tutor and Computer differ by ± 4 marks
10	65	20	13	2	1
20	68	15	13	3	1
30	60	27	11	2	1
40	60	27	11	2	1
50	57	31	10	2	0
60	61	25	11	1	1
70	67	18	12	3	1
80	67	18	11	3	1
90	67	18	11	3	1
100	67	17	11	3	2
200	35	54	7	4	1
300	45	39	12	2	2
400	62	22	12	2	2
500	61	24	12	2	2
600	47	38	12	3	1
627	59	27	11	2	1

Figure 3-1 shows the information from Table 3-3 in graphical form. The figure is in two parts - the left part shows the results when the amount of training data varied from 10 to 90 increasing by 10 each time. The right part shows the results when the amount of training data varied from 100 to 627, the maximum available, increasing by 100 each time.

The figure contains a great deal of information, making it difficult to understand and interpret. How can one determine the best amount of training data by looking at this chart? The y-axis shows the percentage of marks. The x-axis shows the amount of training data. The data points show the percentage of marks where the human and the computer agree, or differ by from zero to four points. The first set of data points (marked by an open 0) indicates the cases where there was zero difference between the tutor mark and the computer mark, i.e., they are identical. The second set of data points (marked by a +) is where there was a difference of plus or minus one point. The fifth set (marked with an open square) indicates those questions with the worst results: either the tutor awarded four points and the computer awarded zero points, or vice versa. The legend below the graph shows the correspondence between each set of data points and the amount by which the human and computer scores differ.

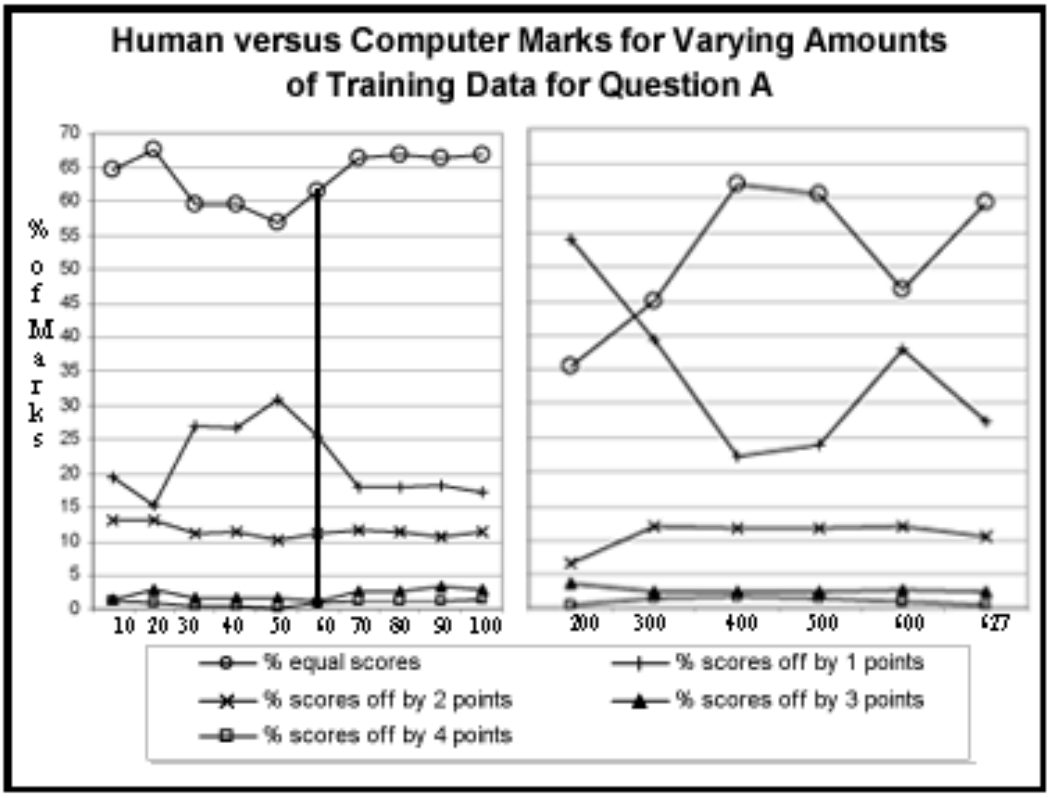


Figure 3-1 Comparison of human and computer marks for various amounts of training data

The viewer can see all of the results for a particular amount of training data by looking at a vertical slice of the graph. For example, the vertical line in the graph shows that when 60 training examples were used, EMMA matched the human about 61% of the time, differed by one point about 25% of the time, differed by two points about 11% of the time, differed by three points about 1% of the time, and differed by 4 points about 1% of the time.

Looking at a vertical slice of the graph shows the performance of EMMA for a particular amount of training data. Looking at a horizontal set of data points gives another point of view. A data set shows how much the performance varies over different amounts of training data. For example, the set indicated with a vertical bar shows that the marks that differed by plus or minus one point ranged from about 15% for 20 training data items to about 54% for 200 training data items.

I tried several different graphical ways to display the data – Figure 3-1 shows the clearest way I found. Even so, it is difficult to evaluate the overall effectiveness of varying the amount of training data by analysing this figure because it contains a lot of information, all of which is necessary to measure the results. Thus, the SM is not adequate for the internal purpose of evaluating various calibrations of the LSA algorithm. The next two sections discuss several other metrics and explain why I found them, like the SM, to be not useful for my work.

3.4 The inadequacy of existing success measures

The literature offers two widely used techniques to evaluate marking systems – precision and recall, and correlation. The following subsections describe them and suggest why they are inappropriate for evaluating CAAs.

3.4.1 Precision and recall

The first technique is the use of precision and recall; these measures are used widely in LSI and LSA research (Dumais, 1991; Manning & Schütze, 1999; Graesser, Wiemer-Hastings, Wiemer-Hastings, Harter & The Tutoring Research Group, 2000; Nakov, Valchanova & Angelova, 2003). Precision looks at how relevant the collection of retrieved documents is; it is the ratio of correctly retrieved, i.e. relevant, documents to all retrieved documents. Recall is a measurement of completeness. It is the ratio of correctly retrieved documents to all relevant documents i.e., those that were retrieved plus those that the retrieval system failed to retrieve (Foltz, 1990). As recall goes up, precision tends to go down; in the trivial case, a system achieves 100% recall if all the documents are retrieved, which would give the lowest precision. Information retrieval (IR) researchers plot values of precision for various levels of recall to provide a good picture of the effectiveness of their techniques (Dumais, 2003). The relevance to LSA and marking is that LSA retrieves the marked answers from the training data that are closest to the answer being marked.

It is important to have a good metric to measure success when calibrating a marking system. Dumais, in a widely cited study (1991), used precision and recall to justify the use of log-

entropy as the weighting function in the term-frequency matrix. (The decision of a weighting function is a critical choice to be made by LSA researchers.) Nakov, Valchanova & Angelova (2003) used precision and recall figures to argue that the choice of a weighting function is the most crucial of all calibration techniques. Many researchers continue to justify the use of the log-entropy weighting factor (Foltz, Kintsch & Landauer, 1998) by relying on the early work of Dumais (1991). Although log-entropy *may* be the best weighting function, it should be justified for LSA-based assessment systems on research done with LSA-based assessment systems instead of IR systems. Researchers need to remember that Dumais is primarily interested in information retrieval rather than essay assessment.

Although precision and recall are useful for evaluating IR techniques, they are largely irrelevant when measuring automated marking systems. Recall is not important – it makes no difference how many documents are returned because the marking system looks at only a pre-determined number that are the closest matches to the document being marked. Precision, on the other hand, is very important – the documents judged by the marking system to be relevant must actually be relevant. Precision, however, is a binary measure; it assumes that the documents are relevant or not. EMMA uses the cosine similarity measure to rank the documents in terms of how similar they are to the answer being marked. It then awards a mark by calculating the weighted average (using the cosine measure) of the five most similar answers. This feature of LSA provides a finer-grained measure than the technique of using precision and recall, which is better suited to information retrieval.

3.4.2 Correlation

The second technique to evaluate marking systems is statistical correlation, which is used by many researchers (Wiemer-Hastings, 1999; Foltz, Gilliam & Kendall, 2000; Perez, Gliozzo, Strapparava, Alfonseca, Rodriquez & Magnini, 2005). The most widely known correlation measures are Pearson's r, Spearman's rho, and Kendall's tau_b (Dancey & Reidy, 2002). The formulas for the Pearson and Spearman measures given by Daniel (1977) are shown below.

Spearman's rho calculates the correlation by calculating the difference between each pair of data points, or ranks. Daniel gives a correction if there are many tied ranks (1977 p. 364).

Equation 3-1 The Pearson correlation coefficient

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}$$

X and Y are the two variables being correlated and n is the number of cases.

Equation 3-2 The Spearman rank correlation coefficient

$$r_s = \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$

n is the number of cases, d is the difference between the ranks

Kendall's tau is based on concordant and discordant pairs (Stegmann & Lucking, 2005). Equation 3-3 gives the formula. Given 2 observations: (x_i, y_i) and $\left(x_j, y_j\right)$ they are:

concordant if when $x_j > x_i$ then $y_j > y_i$

discordant if when $x_j > x_i$ then $y_j < y_i$

tied if $x_i = x_j$ and/or $y_i = y_j$

Equation 3-3 Kendall's tau

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n\left(n-1\right)}$$

where n_c = number of concordant pairs

and n_d = number of discordant pairs

Pearson's r is used when data are normally distributed. For many marking schemes, the marks are negatively skewed, i.e., the tail on the distribution graph goes to the left (Rowntree, 2004 p. 59). This trait occurs because markers tend to give high, rather than evenly distributed, marks. Figure 3-2 gives an example of a skewed distribution. The data are taken from the corpus and are typical of all of the questions I have examined. The figure shows that the data

are non-normally distributed and thus a non-parametric test (e.g., Spearman's rho or Kendall's tau_b) is the appropriate choice (Dancey & Reidy, 2002; Rowntree, 2004 p. 125).

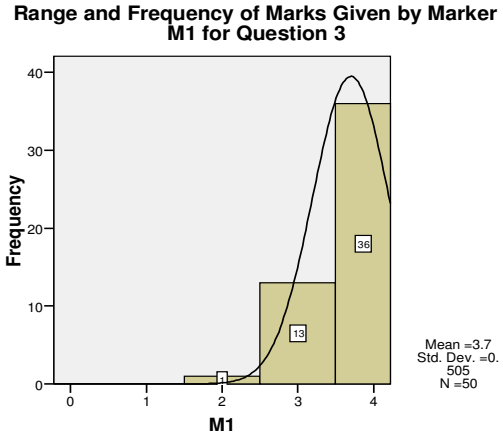


Figure 3-2 Histogram showing that marks are non-normally distributed

Correlation statistics indicate how well one variable can be used to predict another variable. If human-assigned marks and computer-assigned marks agree, the correlation would be perfect. Even if the human marks were always twice the computer marks, the correlation would once again be perfect. In this case, a good correlation would not mean a good marking system.

Another problem with the correlation statistic is that it would be low in the case where computer marks are off by plus-or-minus one point. In this situation, the computer mark could not be used to predict the human mark even though I argue that the overall results of the marking system would be very good if all the inconsistent marks differ by only one point in either direction. The computer would be a good marker overall if sometimes it marks a bit high and other times it marks a bit low. For these reasons, the standard correlation statistics may not be useful for evaluating automated marking systems.

Table 3-4 shows various correlation coefficients for the case where two markers have 96% identical marks, and 4% where they differ by two points. For one mark, marker 1 scored higher than marker 2 and in the other case, scored lower than marker 2. The commonly held understanding of correlation would lead one to expect a high correlation between these two markers because they agree exactly on 96% of the marks, but SPSS calculates essentially zero correlation between the two markers. This example shows that the traditional correlation statistics fail the common sense test and are not applicable to the problem of comparing the similarity between human and computer markers.

Section 3.5 looks at a traditional statistical test (and its non-parametric variations) and explains why it, also, failed to help evaluate my automated marking system.

Table 3-4 Output from SPSS that shows no correlation for two markers who have 96% identical answers

m1 * m2 Crosstabulation

Count

		m2		
		2	4	Total
m1	2	0	1	1
	4	1	48	49
Total		1	49	50

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-b	-.020	.014	-.722	.470
	Kendall's tau-c	-.002	.002	-.722	.470
	Spearman Correlation	-.020	.014	-.141	.888[c]
Interval by Interval	Pearson's R	-.020	.014	-.141	.888[c]
N of Valid Cases		50			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

3.5 Problems with the traditional t-test

Having considered and rejected the metrics described in Section 3.4, I turned to the traditional t-test as a candidate for evaluating automated marking systems. It comes in parametric and non-parametric versions. The parametric tests are more powerful than the non-parametric versions but the data must meet three assumptions to use them: normally distributed populations, approximately equal variations of the populations, and no extreme scores.

The t-test compares the means of two groups. For marking systems, one group is the human-assigned scores and the other group is the computer-assigned scores. When all participants take place in both conditions (short answer marked by tutor and short answer marked by EMMA), the study design is known as within-participants (also called repeated measures or related design) and the appropriate parametric statistical test for comparing the groups is the t-test (Dancey & Reidy, 2002). The SPSS output of the t-test includes the mean scores for each group, the difference between them, and the standard deviations. With these values, one can compute the effect size, which is the difference of the means divided by the mean of the standard deviations. Confidence intervals around the effect sizes are an additional tool for evaluating results (Aberson, 2002). Therefore, if the data meet the three assumptions for using parametric tests, employing effect sizes with confidence intervals could be a good way to evaluate automated marking systems.

Unfortunately, I cannot use the t-test and effect sizes with confidence intervals because my data are not normally distributed, as suggested by Figure 3-2. The marks given by tutors are highly negatively skewed because the marks tend to cluster towards the high end of the marking scale. If the marks are not normally distributed, the effect sizes and confidence intervals will be incorrect (Thompson, 2002) and the t-test is not applicable. I can, however, use the Wilcoxon signed ranks test, which is the non-parametric version of the t-test. This test statistic is calculated by ranking the differences between the two scores. But the scores with zero difference are ignored because "they do not give us any information" (Dancey & Reidy, 2002).

There are two problems with the Wilcoxon t-test. The first problem is the elimination of those cases where the difference is zero. Dancey & Reidy (2002) claim that these cases do not give us any information, which may be true when trying to establish that there *is* a difference between two groups. However, when evaluating marking systems, I want to establish that there is *no* difference between two groups or that the difference is *very small*. If, for example, a marking system produces marks that agree with the human 95% of the time, that figure is informative, contradicting one of the assumptions of the Wilcoxon test. I need a test statistic that takes into account the number of cases where the difference between two marks is zero.

The second problem with the Wilcoxon t-test is that it shows whether two groups are different but not by how much. To solve that problem, I can look at the mean difference given by the descriptive statistics – no difference or very small differences would allow me to conclude that there is no significant difference between two groups. However, as mentioned earlier, calibrating an LSA-based marking system is critical. How should I compare the results of calibrating the system? I cannot use mean differences by themselves; I must consider the standard deviations. This requirement leads me back to effect sizes, but the results will be invalid because my data are not normally distributed.

For the reasons given above, I cannot use correlation statistics, t-tests, or effect sizes with confidence intervals. Section 3.6 presents possible alternative metrics and provides an example of using them to evaluate test results.

3.6 Success metrics using the distance between two vectors

The inability to locate an appropriate metric, as described in previous sections, combined with the difficulty in interpreting Figure 3-1 and Table 3-3, led me to investigate two metrics from the field of vector space theory. The Manhattan Distance measure (L1) and the Euclidean Distance measure (L2) are two metrics used to calculate the distance between two vectors (Gerald & Wheatley, 1970). Their application to marking exams is as follows. One vector is the list of marks given to answers to a question by one marker; the other vector is the list of marks

assigned by another marker. If the vectors are identical, the distance between the vectors is zero and the two markers would agree perfectly.

The measures are calculated using the well-known formulas (Gerald & Wheatley, 1970) shown below. These formulas compute the distance between the vectors in slightly different ways. The L1 computes the sum of the differences between each point in the two vectors; the L2 computes the square root of the sum of the squares of the differences.

Equation 3-4 The Manhattan Distance (1 norm, or L1):

$$M(X,Y) = \sum_{i=1}^{n} |x_i - y_i|$$

Equation 3-5 The Euclidean Distance (2-norm, or L2):

$$M(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

where X = (x_1, x2, xn) and Y = (y_1, y2, yn) are two n-dimensional vectors

The L1 and L2 metrics provide a figure that could be used to assess the results of the different experiments quickly. Table 3-5 shows the same information as Table 3-3 except that it includes L1 and L2 and is sorted by L2. This table shows that 50 is the amount of training data that corresponds to the best outcome for the question being marked using either L1 or L2 as the metric. This evaluation agreed with a careful hand analysis, taking into consideration the number of marks that differ by 1 or more points, of all of the numbers given in Table 3-3.

Table 3-5 Result of varying the amount of training data - sorted from best to worst

# of Marked Answers	% Equal Scores	Tutor and Computer differ by ±1	Tutor and Computer differ by ± 2	Tutor and Computer differ by ± 3	Tutor and Computer differ by ± 4	Manhattan Distance L1	Euclidean Distance L2
50	57	31	10	2	0	93	9.6
60	61	25	11	1	1	95	9.7
30	60	27	11	2	1	97	9.8
40	60	27	11	2	1	98	9.9
627	59	27	11	2	1	100	10.0
10	65	20	13	2	1	105	10.2
80	67	18	11	3	1	107	10.3
70	67	18	12	3	1	108	10.4
20	68	15	13	3	1	109	10.5
90	67	18	11	3	1	110	10.5
100	67	17	11	3	2	114	10.7
500	61	24	12	2	2	116	10.8
400	62	22	12	2	2	119	10.9
200	35	54	7	4	1	122	11.1
600	47	38	12	3	1	124	11.1
300	45	39	12	2	2	133	11.5

The Manhattan and Euclidean distance measures at first seemed to be promising tools for automated marking researchers to evaluate their systems. Unlike the simple metric (SM) these two metrics take into consideration the values of agreement between human and computer over the whole range of possibilities. That is, they evaluate the results where human and computer marks are identical, where they are off by plus or minus one point, plus or minus two points, and so on until the worst result which is where the human and computer differ by the maximum point value of the question. The SM uses just the value where the human and computer marks are identical and can lead to ambiguity, as demonstrated in Table 3-1. L1 and L2 give a richer picture of the effectiveness of an automated marking system than the SM and are no more difficult to analyse than the SM. There are, however, three problems with them. The first problem is that they are not widely used for evaluating CAAs and no agreed upon cut-off levels exist. The second and more serious problem arises when comparing answers with differing point values. The distance between vectors for questions of differing point values cannot be compared in a sensible manner. Finally, neither of these metrics considers chance agreement by markers.

The next subsections discuss metrics that compensate for these problems.

3.7 The inter-rater reliability statistics

Inter-rater reliability (IRR) statistics attempt to quantify the consistency between two raters, e.g. radiologists interpreting an x-ray or humans marking an exam. This section discusses two IRR statistics – Cohen's kappa and Gwet's AC1. Kappa is the better known of the two. AC1, first introduced in 2001 (Gwet), corrects some of the deficiencies of kappa.

3.7.1 The problem with the kappa inter-rater reliability statistic

Cohen's kappa statistic is used for inter-rater reliability (Cohen, 1960). I tried this measure and then discarded it because it gave me non-sensible results. I had instances where EMMA and the human raters agreed by as much as 97% using the SM but the kappa statistic was close to zero, indicating no correspondence. I was reassured to find a paper by two researchers that gave an example of what they called an "absurd kappa value" (Stegmann & Lucking, 2005). Table 3-6 shows the example they used – it was taken from another paper (DiEugenio & Glass, 2004). In each of the experiments illustrated, the observers agreed in 90% of the cases but one case

Table 3-6 Tables illustrating a balanced (left) and a skewed (right) distribution

Observer B	Observer A		Total
	1	2	
1	45	5	50
2	5	45	50
Total	50	50	100

kappa = .8

Observer B	Observer A		Total
	1	2	
1	90	5	95
2	5	0	5
Total	95	5	100

kappa = an "absurd" -0.0526

showed a high kappa figure of 0.8 while the other showed essentially no agreement at kappa = -0.0526. The problem with the kappa statistic suggested by Table 3-6 was first documented by Feinstein and Cicchetti (1990) as summarised in their abstract:

> In a fourfold table showing binary agreement of two observers, the observed proportion of agreement, P0 can be paradoxically altered by the chance-corrected ratio that creates κ as an index of concordance. In one paradox, a high value of P0 can be drastically lowered by a substantial imbalance in the table's marginal totals either vertically or horizontally. In

> the second pardox, (sic) κ will be higher with an asymmetrical rather than symmetrical imbalance in marginal totals, and with imperfect rather than perfect symmetry in the imbalance. An adjustment that substitutes Kmax for κ does not repair either problem, and seems to make the second one worse.

DiEugenio & Glass (2004) explain the problem in more accessible language: "κ is affected by skewed distributions of categories (the **prevalence problem**) and by the degree to which the coders disagree (the **bias problem**)."

Researchers in several disciplines have noted the problems with the kappa statistic and have begun to use Gwet's AC1 statistic. Chan, in a statistics tutorial for the medical profession (2003), provides an example where kappa gives strange results because one of the rater categories has a small percentage. He recommends the use of AC1. Several researchers in the field of software process improvement (Huo, Zhang & Jeffrey, 2006) suggest the use of AC1. Two computational linguists interested in the automatic classification of documents (Purpura & Hillard, 2006) suggest the use of AC1. Another group of computational linguists interested in classifying documents (Yang, Callan & Shulman, 2006) use AC1. Two researchers at the Dartmouth Medicine School (Blood & Spratt, 2007) recommend the use of AC1 and have created and made freely available a macro for the statistical package SAS. However, they caution that AC1 is still a new statistic:

> "Although the AC1 and AC2 statistics are about five years old now, they remain infants in the statistical world, especially since so few people have been exposed to them. With greater usage will come greater scrutiny, and with greater scrutiny may come identification of problems inherent in these statistics. Therefore, as is always the case with new statistics, caution should be exercised in their use and further examination should occur before they are adopted as the standard."

The next subsection describes the Gwet AC1 statistic.

3.7.2 The Gwet AC1 inter-rater reliability statistic

Kilmer Gwet has written extensively about the problems of the kappa statistic and has proposed AC1 (2001a; 2002a; 2002b), which he claims overcomes the problems with kappa. What follows, except where noted, comes from (Gwet, 2002a). The explanation is for the simplified case of two raters and two categories of ratings. Gwet uses the following table in his formulas. "A" is the number of times both raters gave a rating of "1". "B" is the number of times rater A gave a "2" when rater B gave a "1". "A1" is the total number of times Rater A gave a "1" and "A2" is the total number of times Rater A gave a "2". N is the total number of observations.

Table 3-7 Distribution of subjects by rater and response category

Rater B	**Rater A**		
	1	**2**	**Total**
1	A	B	**B**1=A+B
2	C	D	**B**2=C+D
Total	**A**1=A+C	**A**2=B+D	N

According to Gwet (2001a), the kappa formula takes the form of Equation 3-6. It is equivalent to the SM discussed in Section 3.2 corrected by the probability of chance agreement. He shows an example similar to Table 3-6 and claims that kappa can be misleading. He claims that kappa incorrectly overstates the correction for chance agreement.

Equation 3-6 The kappa formula

$$kappa = \frac{p - e(\kappa)}{1 - e(\kappa)}$$

where p = the overall agreement = $\frac{A + D}{N}$

and e(κ) = the chance agreement probability = $\frac{A1}{N} * \frac{B1}{N} + \frac{A2}{N} * \frac{B2}{N}$

and A, A1, A2, B1, B2, D, and N are the figures in Table 3-7.

Equation 3-7 shows Gwet's AC1 statistic.

Equation 3-7 Gwet's AC1

$$AC1 = \frac{p - e(\gamma)}{1 - e(\gamma)}$$

where $e(\gamma) = 2P_1(1 - P_1)$

$$P_1 = \frac{(A1 + B1)/2}{N}$$

and $p = \frac{A + D}{N}$

and

AC1 = the first order agreement coefficient

e (γ) = the chance-agreement probability

P_1 = the approximate chance that a rater classifies a subject into category 1

A1 = the number of times a rater, A, classifies a subject into category 1

A = number of times both raters classifies a subject into category 1

D = number of times both raters classifies a subject into category 2

p = the overall agreement

Equation 3-7 shows how to calculate AC1 for the simple case of two raters and two rating categories. Blood & Spratt (2007) give the formula for the general case. I implemented this formula in Java and used it to evaluate the results I have from EMMA. I chose not to use the version in the statistical package, SAS, mentioned in subsection 3.7.1 for a few reasons. First, I am not familiar with SAS having done my work for this dissertation using SPSS. Second, it is tedious and time-consuming to export raw data from EMMA to SAS and import results back to EMMA. And finally, AC1 is relatively easy to implement. Writing and testing the code took

less time than I estimated locating and obtaining the SAS macro and learning how to use SAS would have taken.

Equation 3-8 The AC1 formula for the general case

$$AC1 = \frac{p_a - p_{e\gamma}}{1 - p_{e\gamma}}$$

where $p_a = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{q=1}^{Q}\frac{r_{iq}\left(r_{iq}-1\right)}{r\left(r-1\right)}\right\}$

and $p_{e\gamma} = \frac{1}{Q-1}\sum_{q=1}^{Q}\pi_q\left(1-\pi_q\right)$

and $\pi_q = \frac{1}{n}\sum_{i=1}^{n}\frac{r_{iq}}{r}$

P_a = the overall agreement probability

$P_{e\gamma}$ = the chance-agreement probability

r_{iq} = the number of raters who classified the ith object into the qth category. The index i *ranges* from 1 to n and q ranges from 1 to Q

n = the number of objects rated

Q = the number of categories in the rating scale

r = the total number of raters

π_q = the probability that a rater classifies an object into category q

3.8 A worked example

I demonstrate the use of the kappa and AC1 formulas, first with the balanced example and then with the skewed example given in Table 3-6.

Table 3-8 summarises the results.

3.8.1 Balanced distribution

$$p = \frac{A+D}{N} = \frac{45+45}{100} = .9$$

$$\text{e}(\kappa) = \frac{A1}{N} * \frac{B1}{N} + \frac{A2}{N} * \frac{B2}{N} = \frac{50*50}{100*100} + \frac{50*50}{100*100} = \frac{2500+2500}{10000} = .5$$

$$kappa = \frac{p - e(\kappa)}{1 - e(\kappa)} = \frac{.9-.5}{1-.5} = \frac{.4}{.5} = .8$$

$$P_1 = \frac{(A1+B1)/2}{N} = \frac{(50+50)/2}{100} = \frac{50}{100} = .5$$

$$e(\gamma) = 2P_1(1-P_1) = 2*.5(1-.5) = 1*.5 = .5$$

$$AC1 = \frac{p - e(\gamma)}{1 - e(\gamma)} = \frac{.9-.5}{1-.5} = \frac{.4}{.5} = .8$$

3.8.2 Skewed distribution

$$p = \frac{A+D}{N} = \frac{90+0}{100} = .9$$

$$\text{e}(\kappa) = \frac{A1}{N} * \frac{B1}{N} + \frac{A2}{N} * \frac{B2}{N} = \frac{95*95}{100*100} + \frac{5*5}{100*100} = \frac{9025+25}{10000} = .905$$

$$kappa = \frac{p - e(\kappa)}{1 - e(\kappa)} = \frac{.9-.905}{1-.905} = \frac{-.005}{.095} = -.0526$$

$$P_1 = \frac{(A1+B1)/2}{N} = \frac{(95+95)/2}{100} = \frac{95}{100} = .95$$

$$e(\gamma) = 2P_1(1-P_1) = 2*.95(1-.95) = 1.9*.05 = .095$$

$$AC1 = \frac{p - e(\gamma)}{1 - e(\gamma)} = \frac{.9-.095}{1-.095} = \frac{.805}{.905} = .8895$$

Table 3-8 Comparison of kappa and AC1 for balanced and skewed distributions shown in Table 3-6 showing that kappa gives a strange result for a skewed distribution

	Balanced Distribution	Skewed Distribution
kappa	0.8	-0.05
AC1	0.8	0.89

This worked example shows that kappa gives a result for a skewed distribution that fails the common sense test and confirms the work of various researchers (Feinstein & Cicchetti, 1990; Gwet, 2002a; Stegmann & Lucking, 2005; Blood & Spratt, 2007). It also supports Gwet's claim that AC1 is a "more robust chance-corrected statistic that consistently yields reliable results" (Gwet, 2002a).

Chapter 4. How Well Do Human Markers Agree?

4.1 The Study

4.1.1 The purpose of the study

A Computer Assisted Assessment system (CAA) is *good enough* if it agrees with human markers as well as human markers agree with each other. Thus, in order to evaluate my LSA-based CAA, I needed to quantify how well human markers agree with each other. While it is often assumed that marking variability exists (see Chapter 1), it is difficult to find supporting evidence. I use the results of this study as a baseline against which to compare my CAA. If the results of my CAA closely match or exceed the baseline, then I can be assured that my CAA is *good enough.*

Inter-rater reliability (IRR) is the technical term used to describe how closely raters agree with each other. Gwet (2001a p. vii) states "Virtually anything that is used to generate explicitly or implicitly a measure for classifying a subject into a predefined category can be considered as a rater." He uses nurses diagnosing psychiatric patients (2001a p. 53) and scientists classifying fish according to colour (2001a p. 98) as examples of raters. In this dissertation, the raters are both human and computer markers; this chapter discusses human markers. The subjects, analogous to Gwet's patients or fish, are student answers. The AC1 statistic, discussed in subsection 3.7.2, was created to establish the level of agreement among raters (Gwet, 2001a p. vii). I have chosen AC1 to report the IRR obtained by this study for the reasons given in Chapter 3.

4.1.2 The participants

I recruited five expert markers from the Open University (OU) staff. They have an average of 7.5 years experience as markers at the OU with an average of 3.5 years experience marking for the course from which I took the answers-to-be-marked. The OU markers are highly trained – they go through a training course, mark to a detailed marking scheme, and are accustomed to having their marks moderated. As a sign of their conscientiousness, they often use a course on-line bulletin board to discuss intricacies of marking particular questions.

The reader should note that the marks collected for this study are un-moderated, that is, they were not checked, verified, and re-marked in the event of a disagreement between markers. Had the marks been intended for actual marking, they would have been moderated. Because OU courses can have thousands of students, multiple markers mark one course. The OU has procedures in place, including moderating marks and double-marking for high stakes assessments, to ensure a high level of consistency.

4.1.3 The Data

I used 18 different questions for this study (see Table 4-1). There are several types of questions; however, they are all from the first two homework assignments of the February 2004 presentation of M150 – Data, Computing and Information, which is an introductory course offered by the OU Computing Department. Some of the questions (e.g. 13, 14, 16) require quite concise, short, straight-forward answers while others (e.g. 4, 20) require longer, more open-ended answers. Some (e.g. 1 and 2) are multi-part and worth 8 and 12 points respectively while others are worth just 2, 3, or 4 points. Five questions (8-12) are about html. Thus, there is a variety of question types, although the main point is that they are all short answer, rather than multiple choice or true/false type questions. Table 4-1 shows the text of the 18 questions for which the human markers evaluated the student answers. (Note that the 18 questions are numbered 1 to 21. I removed questions 5, 6, and 7 from the study because they involved converting numbers from binary to octal; being numeric, rather than textual, they were unsuited for an LSA-based assessment system.)

Table 4-1 Text of questions

	Question Text	points
Q1	Name 2 elements of the course materials that will be distributed via the M150 course website?	8
	What is the role of the Study Calendar? What is the cut-off date for TMA02?	
	Find the learning outcomes for M150 which are listed in both the Course Companion and the Course Guide. Write down the learning outcome that you feel you are most interested in achieving and one or two sentences to describe why you have chosen that learning outcome.	
	What does eTMA stand for? What is the name of the document you should read to prepare yourself for submitting an eTMA? Who should you contact with queries about course software?	
Q2	Find the UK AltaVista site. What is its URI? What is the name of the large aquarium in Hull?	12
	Which query led you to the answer? What is the URI of the site?	
	What is the minimum number of intervening web pages you have to visit between the main site and the page that contains the information on the ballan wrasse?	
	List the URI of each intervening web page. How big can a ballan wrasse grow?	
	Does the ballan wrasse page tell you anything about the age a ballan wrasse can reach?	
	What age can a ballan wrasse reach?	
	What is the URI of the web page where you found the information?	
	Which search engine, and which query got you to the page that contained your answer?	
Q3	Explain, with examples, the difference between an analogue and a discrete quantity.	4
Q4	Give an example of a computer standard, explaining its purpose. Why is there a general need for standards in computing?	4
8-12	For each case; write the correct HTML and write one or two sentences about the problem with the original HTML. (The first line is the original HTML. The second line is the desired appearance.)	
Q8	<B>Always look left and right before crossing the road.	4
	Always look left and right before crossing the road.	
Q9	<B>Important!<B>Do <B> not place metal items in the microwave.	4
	Important! Do **not** place metal items in the microwave.	
Q10	<I>It is <B>very</I> </B> important to read this text carefully.	4
	It is ***very*** important to read this text carefully.	
Q11	Things to do: Things to do:	4
	Pack suitcase, </BR>	
	Book taxi. Pack suitcase,	
	Book taxi.	
Q12	More information can be found here.	4
	More information can be found here.	
13-21	Victoria uses her computer to write up a report. When complete, she saves it to the hard disk on her computer. Later she revises her report and saves the final version with the same document name.	
Q13	Considering the contents of the report as data, at what point does the data become persistent?	2
Q14	What happens to the first saved version of the document?	2
Q15	Suggest an improvement in Victoria's work practice, giving a reason for your answer.	2
Q16	Give two examples of persistent storage media other than the hard disk.	2
Q17	Victoria then wishes to email a copy of her report, which includes data on identifiable individuals, to John, a work colleague at her company's Birmingham office. Write two sentences to explain the circumstances under which, within UK law, she may send the report.	2
Q18	Explain briefly the property of internet email that allows the contents of the report to be sent as an attachment rather than as text in the body of the email message.	2
Q19	John's email address is John@Birmingham.office.xy.uk Which parts of the address are: the user name, the name of the domain, the top-level domain?	2
Q20	Victoria then prepares her report for publication on a website. In no more than 100 words, explain what she has to take into account when making her report public.	3
Q21	Which of the following should she publish on the website with her report and why? Company address, personal telephone number, email address	3

The student answers came from the actual student scripts to questions given in the introductory computer literacy course mentioned above. Appendix C gives some examples of the answers. Each of the five markers (with exceptions noted below) marked the same set of 60 random student answers to the 18 questions using the marking scheme created for the presentation of the course used in this study. I discarded the marks for the first 10 answers to each question so that the markers could become familiar with the marking scheme before I recorded their marks. To calculate the IRR of the 5 markers, I paired each of them with the other four for a total of ten human to human comparisons (markers 1 and 2, 1 and 3, and so on). These individual comparisons give an idea of the range of variation in human marking on these questions. Chapter 8 uses the comparisons to evaluate the results of my LSA-based assessment system.

4.1.4 Validity

The study has good validity for several reasons. First, the participants were expert markers experienced in exactly the type of marking required by the study. In addition, the 18 questions were designed for an actual course presentation with no previous knowledge that they would be used to test the accuracy of human markers. The 50 answers marked for each question were genuine student answers. Finally, the large quantity of authentic data provides reassurance that the results can be generalised.

However, there are four possible threats to the validity of this study. One threat is the motivation of the markers, who were guaranteed anonymity and were paid for their work. (I was advised that it would be easier to recruit the markers if they knew their marking would not have negative repercussions, thus I promised I would not reveal their marking statistics. Also, it is standard practice to guarantee anonymity to test subjects.) Thus, if they were interested in completing the job as quickly as possible, they could have been careless with their marking. Unfortunately, I have no way of gauging the likelihood of this occurrence. This situation is somewhat analogous to real marking - markers are paid for their work. However, the guaranteed

anonymity removed one reason for conscientious marking – in real marking situations, markers are monitored and one who consistently mismarks is likely not be rehired.

The second threat to validity is that the web interface between the markers and the marks database prevented the markers from reviewing their marks to adjust them, unlike their normal marking procedures. This inability to revise their marks was due to an uninformed design decision on my part. It didn't occur to me that the markers would want to go over their mark. This oversight could have resulted in less consistency than normal due to the inability of the markers to double-check their work. However, at least two of the markers were conscientious enough to *want* to review their marks. This fact may counterbalance the threat in the previous paragraph - that markers may have been careless because they were guaranteed anonymity.

The third point is that the results obtained from this study might show an unusually high level of agreement because all of the markers are experienced. Less experienced markers might not be as consistent as these markers. The OU markers have years of experience carefully following a marking scheme to produce justifiably correct marks. In short, the OU markers are good. Less experienced or less well-trained markers might not do as well.

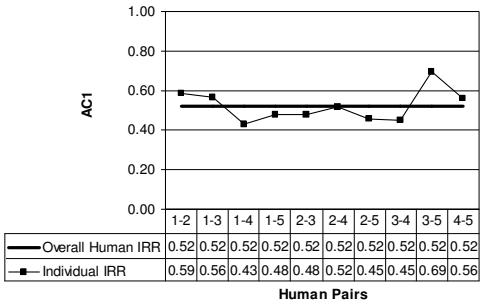
Finally, due to a database overflow problem, two of the markers were unable to complete all of the marking. Thus, Question 17 was marked by just four humans and Questions 19-21 were marked by only three humans. Although this problem does not invalidate the results, it does mean that different questions have differing number of markers requiring care to be taken when comparing the results for the affected questions. One of the strengths of this study, the vast amount of data collected and analysed, still holds.

Despite the four problems mentioned in the previous paragraphs, I believe the study provides valuable results. The markers were professional and experienced (in contrast to most studies e.g. (Foltz, 1996) which use graduate students as markers), and the variety and authenticity of the questions as well as the expertise of the markers support the generalise-ability of the findings.

4.2 The results

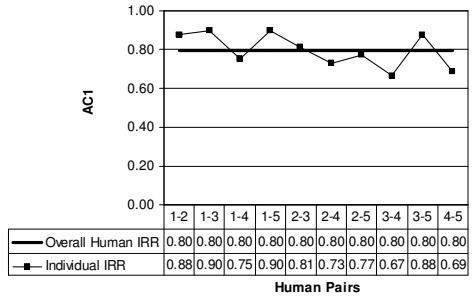
Appendix D gives the raw data collected for this study from which I computed the IRR figures using Gwet's AC1 statistic described in Chapter 3. For this metric, a higher AC1 number indicates that the relevant markers are closer in agreement than those with a lower AC1 number. Figure 4-1 through Figure 4-18 display the IRR figures (discussed in subsection 3.7.2) for each question. Questions 1-16 and 18 were marked by five humans yielding ten pairs for each question. Question 17 was marked by four humans resulting in six pairs. Questions 19-21 were marked by three humans giving three pairs for each question. In addition, I calculated the overall IRR for all five markers (four for question 17 and three for questions 19-21). In each of the figures, the horizontal line is the IRR for all of the markers; the segmented line shows the IRR for each pair of markers.

Figure 4-19 summarises the previous 18 figures; it shows the average IRR for each of the questions sorted from worst to best. This graph shows a wide range of values, from a low of 0.15 to a high of 0.97. The average IRR is 0.59 with a standard deviation of 0.27. By inspecting this figure, one can determine which questions show better agreement. Q19 shows the highest level of agreement while Q17 show the lowest level of agreement.



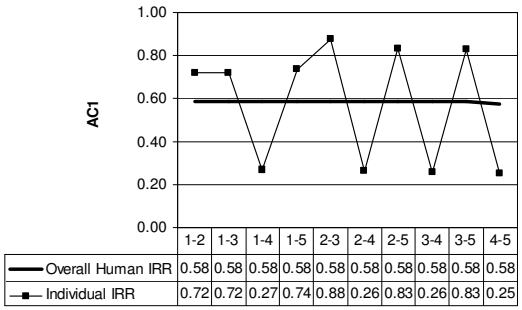
	1-2	1-3	1-4	1-5	2-3	2-4	2-5	3-4	3-5	4-5
Overall Human IRR	0.52	0.52	0.52	0.52	0.52	0.52	0.52	0.52	0.52	0.52
Individual IRR	0.59	0.56	0.43	0.48	0.48	0.52	0.45	0.45	0.69	0.56

Figure 4-1 Inter-rater Reliability for Question 1



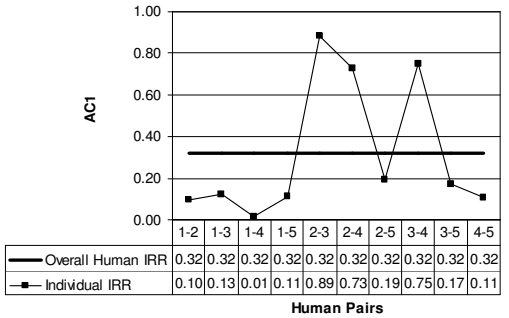
	1-2	1-3	1-4	1-5	2-3	2-4	2-5	3-4	3-5	4-5
Overall Human IRR	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
Individual IRR	0.88	0.90	0.75	0.90	0.81	0.73	0.77	0.67	0.88	0.69

Figure 4-2 Inter-rater Reliability for Question 2



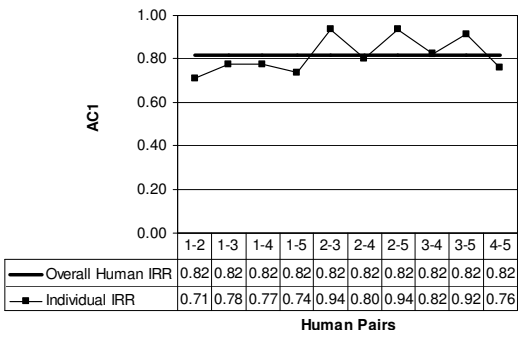
	1-2	1-3	1-4	1-5	2-3	2-4	2-5	3-4	3-5	4-5
Overall Human IRR	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58
Individual IRR	0.72	0.72	0.27	0.74	0.88	0.26	0.83	0.26	0.83	0.25

Figure 4-3 Inter-Rater reliability for Question 3



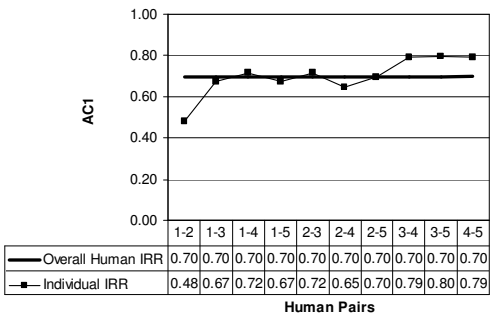
	1-2	1-3	1-4	1-5	2-3	2-4	2-5	3-4	3-5	4-5
Overall Human IRR	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32
Individual IRR	0.10	0.13	0.01	0.11	0.89	0.73	0.19	0.75	0.17	0.11

Figure 4-4 Inter-rater Reliability for Question 4



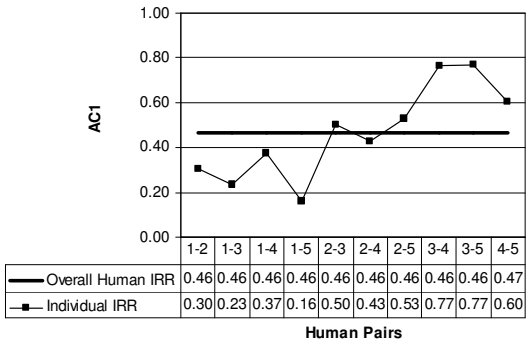
	1-2	1-3	1-4	1-5	2-3	2-4	2-5	3-4	3-5	4-5
Overall Human IRR	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82
Individual IRR	0.71	0.78	0.77	0.74	0.94	0.80	0.94	0.82	0.92	0.76

Figure 4-5 Inter-rater Reliability for Question 8



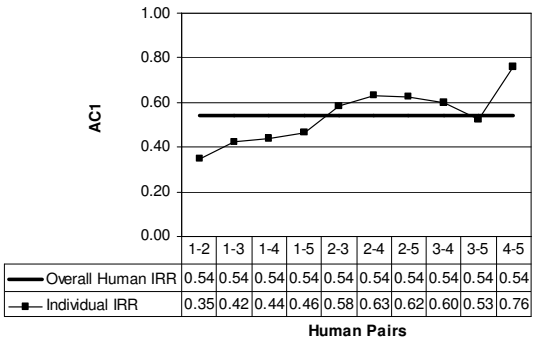
	1-2	1-3	1-4	1-5	2-3	2-4	2-5	3-4	3-5	4-5
Overall Human IRR	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70
Individual IRR	0.48	0.67	0.72	0.67	0.72	0.65	0.70	0.79	0.80	0.79

Figure 4-6 Inter-rater Reliability for Question 9



	1-2	1-3	1-4	1-5	2-3	2-4	2-5	3-4	3-5	4-5
Overall Human IRR	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.47
Individual IRR	0.30	0.23	0.37	0.16	0.50	0.43	0.53	0.77	0.77	0.60

Figure 4-7 Inter-rater Reliability for Question 10



	1-2	1-3	1-4	1-5	2-3	2-4	2-5	3-4	3-5	4-5
Overall Human IRR	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54
Individual IRR	0.35	0.42	0.44	0.46	0.58	0.63	0.62	0.60	0.53	0.76

Figure 4-8 Inter-rater Reliability for Question 11

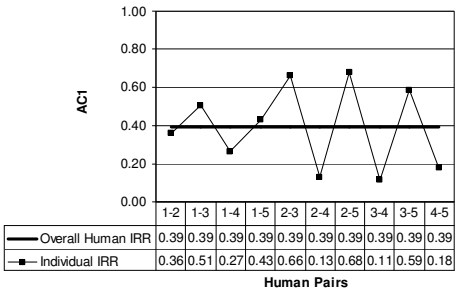


Figure 4-9 Inter-rater Reliability for Question 12

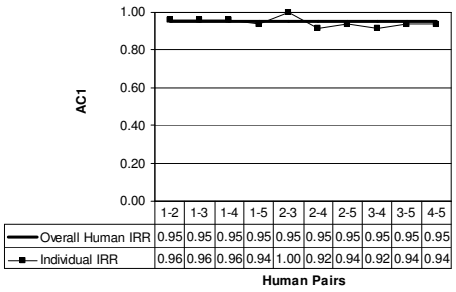


Figure 4-10 Inter-rater Reliability for Question 13

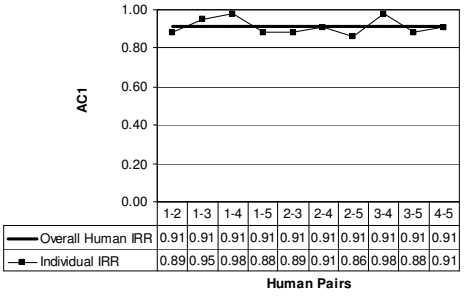


Figure 4-11 Inter-rater Reliability for Question 14

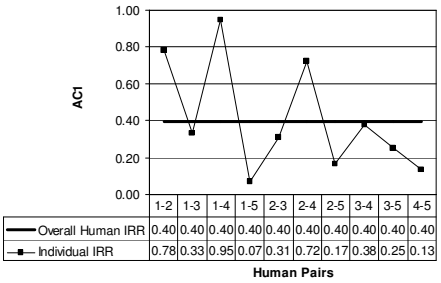


Figure 4-12 Inter-rater Reliability for Question 15

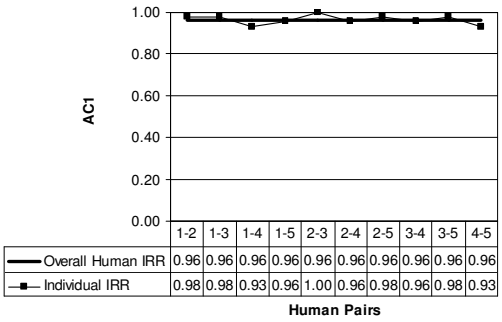


Figure 4-13 Inter-rater Reliability for Question 16

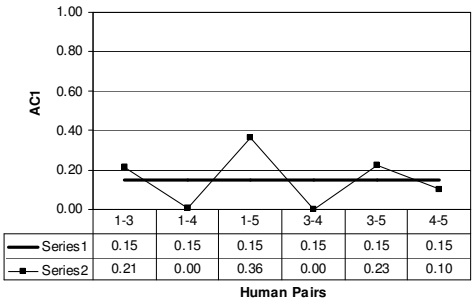


Figure 4-14 Inter-rater Reliability for Question 17

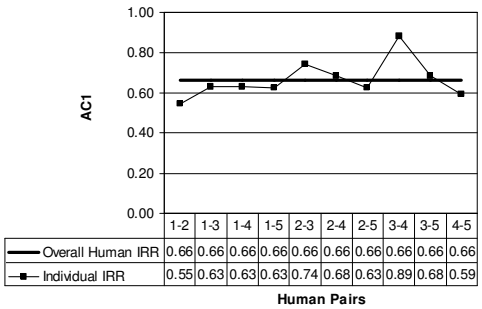


Figure 4-15 Inter-rater Reliability for Question 18

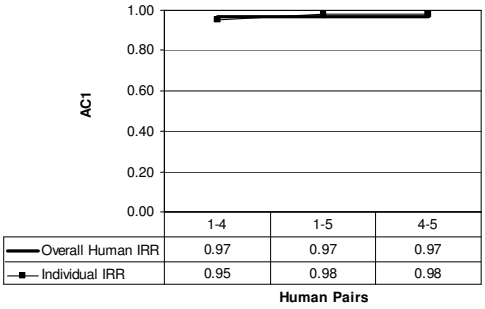


Figure 4-16 Inter-rater Reliability for Question 19

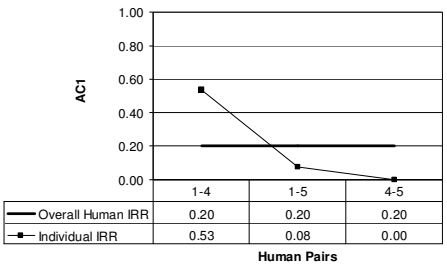


Figure 4-17 Inter-rater Reliability for Question 20

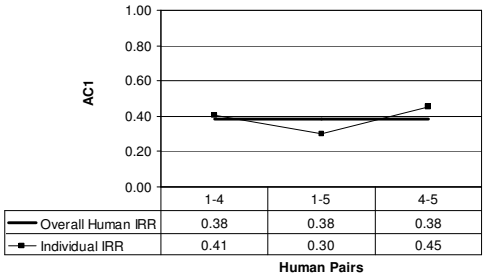


Figure 4-18 Inter-rater Reliability for Question 21

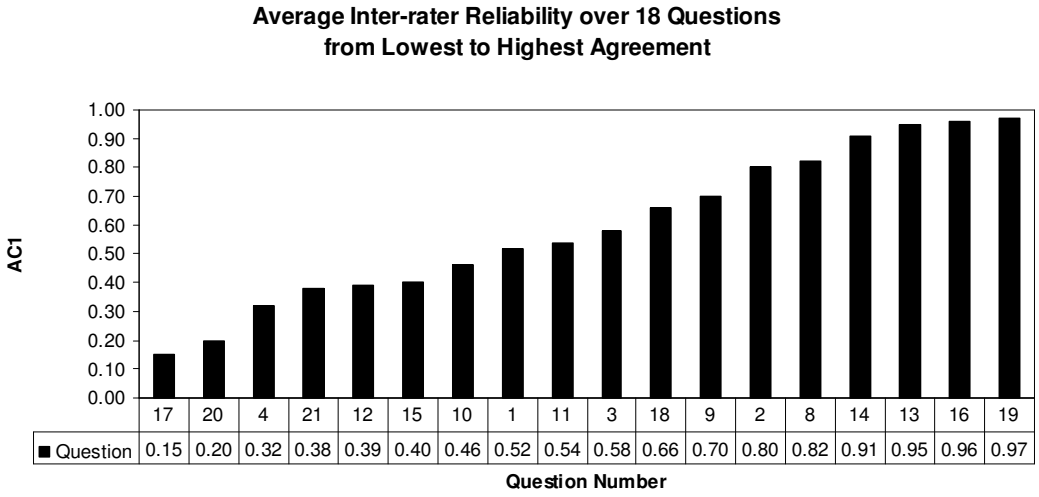


Figure 4-19 Average Inter-rater Reliability over 18 Questions from Worst to Best

4.3 Discussion and implications

By glancing at the first 18 figures, one can see that for many of the questions, there is a large amount of inconsistency in the IRR figures within a single question. Questions 3, 4, and 15 show dramatic differences among the pairs of markers. For example, in Q4 the IRR ranges from a low of 0.01 for pair 1 and 4 to a high of 0.89 for pair 2 and 3. The average IRR for Q4 is 0.34.

Seven pairs of markers were below this average and three pairs were substantially above the average.

Work undertaken by three OU researchers supports my findings that markers can vary widely. Thomas, et al. (2008) found that the IRR of humans and Auto Mark (their CAA) increased from 0.19 for un-moderated marks to 0.76 for moderated marks. The improvement from un-moderated to moderated marks implies that the original marks were not in close agreement.

In contrast to the questions with a wide variability in marking, in each of Questions 2, 13, and 16, the marker pairs are similar. For Q16, for example, the IRR ranges from 0.89 for pairs 1 and 4 and 4 and 5 to a high of 0.96 for pair 2 and 3; these ten pairs of markers have an average IRR of 0.92. These data suggest that Q16 is easy for human markers to mark at a high level of consistency.

For some of the questions, a particular marker or markers seem to lower the average IRR. For Questions 2, 3, 12, and 16, the worst four pairs contain marker 4; for Question 11, the worst four pairs contain marker 1, and for Question 15, the worst pairs contain marker 5. This observation has ramifications for evaluating the accuracy of a CAA system. If an observer can identify the CAA as giving the least consistent marks, then one might conclude that the CAA is not an adequate marker.

Figure 4-19 shows the average IRR for all of the 18 questions. They range from a low of 0.15 to a high of 0.97 with an average of 0.59. This huge difference from the lowest IRR to the highest IRR has a couple of implications. First, these data suggest that some questions are harder to mark than others. This difficulty could arise from an ambiguity in the question or a difference of opinion in how the marking scheme should be interpreted. Second, and more important for this dissertation, is the implication for the evaluation of a CAA system. Because the level of agreement among human markers depends on which question is being considered, it is necessary to compare the computer and human IRR figures for one question at a time. An inaccurate impression of the accuracy of an automatic marker would be given if, for example,

one reported that the average human IRR was 0.59 and the CAA achieved 0.57. The results of this study show that these two figures would overstate the CAA system's level of agreement with human markers for some questions and understate it for others.

4.4 Summary

The purpose of this study was to determine how well human markers agree with one another. By using Gwet's AC1 measure of inter-rater reliability, the study provides evidence that even very experienced and well trained markers often produce a wide range of IRR, both for the same question as well as for different questions.

The major conclusion from these data is that evaluating IRR is complex. It is not sufficient to report a single IRR figure. To gain a deeper understanding of the performance of raters, including automatic, computer-based raters, one needs to know the range and type of questions being marked as well as the IRR for each question. This conclusion will be further explored in Chapter 5, which discusses the evaluation framework that is a major result of the work undertaken for this dissertation.

Chapter 5. The Evaluation Framework

5.1 Background and usefulness

The framework for evaluating computer assisted assessment (CAA) systems is based on the research taxonomy (Haley, Thomas, De Roeck & Petre, 2005) I developed to compare LSA-based educational applications (see section 2.4). It was the result of an in-depth, systematic review of the literature concerning LSA research in the domain of educational applications. The taxonomy was designed to present and summarise the key points from a representative sample of the literature.

The taxonomy highlighted the fact that others were having difficulty matching the results reported by the original LSA researchers (Landauer & Dumais, 1997). I found ambiguity in various critical implementation details (e.g. weighting function used) as well as unreported details. I speculated that the conflicting or unavailable information explains at least some of the inability to match the success of the original researchers. The experience of creating and using the taxonomy served to crystallize my thinking about the important elements of reporting on a CAA and prompted me to create the evaluation framework. The framework simplifies the taxonomy and makes it more concise.

I designed the framework before the study to determine how well-human markers agree with each other (described in Chapter 4) was completed. The results of that study suggested that even well-trained and experienced human markers vary substantially in their marking, both within a single question and over a variety of questions. The study found that the IRR of the human markers ranged over the 18 questions being marked from a low of 0.15 to a high of 0.97 with an average of 0.59. The huge difference between the low and the high figures led me to conclude that the average IRR is a misleading, and certainly incomplete, measure of the accuracy of a

CAA system. The results of the study described in Chapter 4 strengthen my belief that the proposed evaluation framework is vital for anyone attempting to thoroughly understand an automatic marking system.

The framework is valuable to both producers and consumers of CAA. Producers are researchers and developers who design and build assessment systems. They can benefit from the framework because it provides a relatively compact yet complete description of relevant information about the system. If producers of CAA systems use the framework, they can contribute to the improvement of CAA state-of-the-art by adding to a collection of comparable data.

Consumers are organisations, such as universities, that wish to use a CAA system. CAA consumers are, or should be, particularly interested in two areas. The first and most important area is the accuracy of the results. But what does accuracy mean and how does one measure it? I contend that a CAA system is *good enough* if its marks correspond to human markers as well as human markers correspond with each other.

The second area that consumers should be interested in is the amount of human effort required to use the assessment system. Most natural language processing assessment systems, including those based on LSA, require a large amount of training data. Although the system might save time for markers, it may be impractical to use because it takes too much time to prepare the system for deployment (for example, to train the system for a specific data set). Some systems require human manipulation of the training data. An example of an LSA-based system requiring human effort beyond collecting the training data is called Apex (Lemaire & Dessus, 2001), which requires the user to annotate the training data.

5.2 Details of the framework

It is difficult to compare automated assessment systems because no uniform procedure exists for reporting results. This dissertation attempts to fill that gap by proposing a framework for reporting on and evaluating automated assessment tools. The framework for describing and evaluating a CAA can be visualised as the jigsaw puzzle in Figure 5-1. I contend that all the pieces of this puzzle must be present if a reviewer wants to see the whole picture.

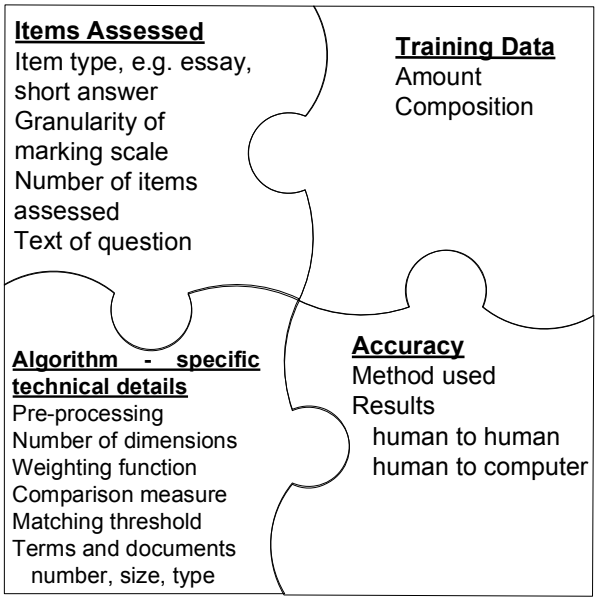


Figure 5-1 The framework for describing and evaluating Computer Assisted Assessment systems

The important categories of information for specifying a CAA are the items assessed, the algorithm-specific technical details, the training data, and the accuracy. The first category provides essential information about the items being assessed. The general type of question (e.g., short answer, multiple choice) is crucial for indicating the power of a system. Marking multiple choice questions is a well understood exercise - simply match up the choice given by a student to the choice on the marking scheme. The granularity of the marking scale provides important information about the accuracy – it is easier to obtain a higher level of marker agreement for a 3 point question than one worth 100 points. For example, two markers might

award marks of 95 and 97 if they are marking on a 100 point scale but both would probably award a mark of 3 on a 3 point scale. In the former case, they would not agree but would have perfect agreement in the latter case. The number of items assessed provides some idea of the generalise-ability and validity of the results. Both the number of unique questions and the number of examples of each question contribute to the understanding of the value of the results. The more items assessed, the more able one is to generalise the results. Also, knowing how many answers to a particular question were marked helps to know how one can generalise the results.

The second category of the framework comprises the technical details of the algorithm used. Haley, Thomas, De Roeck & Petre (2005) discuss why these options are of interest to producers of an LSA-based CAA. Essentially, the developers of an LSA system must choose many parameters that affect the accuracy of the system. The lower left piece of Figure 5-1 shows LSA-specific options, but these could be changed if the CAA is based on a different method.

The corpus used to train the CAA is the third crucial category. Both the type and amount of text help to indicate the amount of human effort needed to gather this essential element of CAA systems. Some systems (LSA for example (Haley, Thomas, De Roeck & Petre, 2007)) need two types of training data – general text about the topic being marked and specific previously marked answers. Researchers should include information about both these types of training data, if applicable.

The fourth category of the framework is the accuracy of the marks. A CAA system exhibiting poor agreement with human markers is of little value. Previous work (Haley, Thomas, De Roeck & Petre, 2005) showed that different researchers report their results using different methods. Ideally, all researchers would use the same method for easily comparable results. I propose the use of Gwet's AC1 statistic in this dissertation; but even if researchers disagree with my suggestion and choose to use another technique, they should at least clearly specify how they determined the accuracy of their results.

5.3 Using the framework for an LSA-based CAA

My framework for evaluating an automated assessment system is a refined version of the taxonomy discussed in Chapter 2. Table 5-1 is an example of how the framework could be used to compare different research results in tabular form. It starts with an overview and proceeds with the pieces in the puzzle of Figure 5-1. The first study, indicated by HTD07 in the far left column, attempted to quantify the optimum amount of training data to mark questions about html needed for best results (Haley, Thomas, De Roeck & Petre, 2007). Gwet's AC1 statistic was not used to report the results as I completed the 207 study before locating Gwet. The study uses the Simple Metric of Section 3.2.

All of the relevant information concerning that study is in the table. The next paragraph gives the information in prose form. It is easier to use the table to compare my results with other systems than it is to digest the text in the next paragraph. The table gives all of the information specified in the framework in a reasonably concise form.

The assessment system is called EMMA, which was developed to assess computer science short answers for summative assessment. EMMA is a research prototype – not yet a deployed system. The innovation of the study was to determine the optimum amount of training data and found that 50 marked answers were optimum for question A and 80 marked answers were optimum for question B. Each of the questions, which were about html, was worth 4 points and I evaluated 50 student answers per question. The table contains the text of the two questions. The table gives the information relating to LSA parameters. This may not be of interest to consumers of assessment systems but is vital for other researchers wishing to replicate the findings. I used 53,073 paragraphs from course textbooks to serve as general training data. To evaluate the results of EMMA, I compared the marks given by five humans and calculated the average. I then compared EMMA's marks with each of the five humans and calculated the average. I found that EMMA worked better for question A than it did for Question B. Fifty-three percent of EMMA's marks were identical to the human marks. Thirty-four percent of the marks differed by one point, 12% differed by two points, and 1% differed by three and four

Table 5-1 Filling in the framework Part 1

	OverView				
Refer-ence	Who / Where / What / Why	Stage of Development/ Type of work	Innovation	Major Result / Key points	Human Effort
HTD 07	Haley, Thomas, De Roeck, Petre; The Open University assess computer science short answers for summative assessment	research prototype	marked questions about html; determined the optimum amount of training data	amount of training data that works best: 50 marked answers for question A amount of training data that works best: 80 marked answers for B	gather training data, gather marked answers
SCS 08	Srihari, University at Buffalo, US // develop new algorithms to grade hand-written essays	research prototype	attempting to analyze handwriting before grading essays	could recognize 60% of handwritten words;	gather training data, gather marked answers

	Items Assessed			
Refer-ence	Type of Item	Granularity of Marking Scale	# of items assessed	text of question
HTD07	short answers about html	0 - 4 points	50	Correct the following fragments of HTML. For each case, write the correct HTML and write one or two sentences about the problem with the original HTML. HTML: <I>It is <B>very</I> </B> important to read this text carefully. The desired appearance: It is ***very*** important to read this text carefully.
HTD07		0 - 4 points	50	Correct the following fragments of HTML. For each case, write the correct HTML and write one or two sentences about the problem with the original HTML. HTML: Things to do: Pack suitcase, </BR> Book taxi. The desired appearance: Things to do: Pack suitcase, Book taxi.
SCS08	short answers for reading comprehension	0 - 6 points	150	How was Martha Washington's role as First Lady different from that of Eleanor Roosevelt?
		0 - 4 points	102	Write a newspaper article encouraging people to attend an art show where Alexandra Nechita is showing her paintings. Use information from BOTH articles that yu have read. In your article be sure to include (i) information from A. Nechita, (ii) Different ways people find art interesting, and (iii) Reasons people might enjoy A. Nechita's painting.

Reference	Algorithm-specific Technical Details											Training Data	
						Terms			Documents				
	preprocessing	# dimensions	weighting function	comparison measure	matching threshold	Number	Size	Type	Number	Size	Type	Size	Composition
HTD07	stemming, stop words	90	log / entropy	cosine	none	12k	1 word	text	45k	1 paragraph	text	1) 45k paragraphs 2) 50	1) course texts 2) human marked answers
HTD07		500	log / entropy	cosine	none							1) 45k paragraphs 2) 80	1) course texts 2) human marked answers
SCS08	stemming, stop words	47		cosine		2,078	1 word	text				1) 10 long passages 2)150	1) course texts 2) human marked answers
		213										1) 10 long passages 2) 103	1) course texts 2) human marked answers

Reference	Accuracy			
	method used		Human to LSA	Human to Human
HTD07	compared LSA marks with 5 human markers and calculated average; simple metric	average % identical off by 1 off by 2 off by 3 off by 4	53 34 12 1 1	54 32 11 1 1
HTD07	compared LSA marks with 5 human markers and calculated average, simple metric	average % identical off by 1 off by 2 off by 3 off by 4	43 45 6 3 3	61 28 9 1 1
SCS08	compared LSA marks with 2 humans, simple metric	average % identical off by 1 off by 2 off by 3 off by 4 off by 5	28 36 19 9 5 3	95% within one point
		average % identical off by 1 off by 2 off by 3 off by 4	31 48 14 7 0	96% within one point

points. This compares to the human average agreement, which was 54, 32, 11, 1, and 1 for the same point differences. These figures suggest that EMMA produced very similar results to what the humans did for question A. The results were not as good for question B. The table gives the relevant figures.

The second study, indicated by SCS08 in the far left column, involved handwriting recognition of essays followed by marking using LSA (Srihari, Collins, Srihari, Srinivasan, Shetty & Brutt-Griffler, 2008). It is included in the table to offer a comparison of how the framework can be used to compare research by two different groups of researchers.

5.4 Summary

The proposed evaluation framework serves the needs of both producers and consumers of CAA systems. Producers need all of the details of the system to understand and improve the state-of-the-art of CAA. Consumers need to understand the details of the system in order to make an informed choice about which CAA to use.

The four pieces of the framework are Items Assessed, Training Data, Algorithm-specific Technical Details, and Accuracy. I contend that all four of these pieces are necessary for a thorough understanding of any CAA system.

Chapter 6 discusses EMMA, the LSA-based CAA developed for this dissertation. Chapter 7 uses the framework to report the results of various experiments using EMMA. Chapter 8 compares the best results of EMMA with human Markers, gives conclusions and suggests further work.

Chapter 6. EMMA – An LSA-based Marking System

EMMA (ExaM Marking Assistant) is the name of the LSA-based marking system I developed to mark short answers to questions in the domain of computer science. It is written in Java and uses MYSQL as the database engine.

EMMA uses a database containing several types of information: basic course information (e.g. number of questions, question text), general training data in the domain being tested (e.g. course textbook) and previously marked answers.

Any LSA-based system requires a server with a huge amount of RAM and a fast processor. In addition, EMMA benefits from having multiple processors because the Java JVM naturally multithreads input and output, and EMMA uses a MySQL database management system that runs on the same computer as EMMA itself. The tests presented in this dissertation were run on a server-class computer equipped with:

- one 2 GHz quad-core Intel Xeon 5400 series processor (64 bit)
- ASUS Z7S WS motherboard
- 16 gigabytes of 667 MHz error-correcting memory (4 bars of 4GB)
- 7200 rpm hard disk
- Windows XP Professional x64 (64 bit)
- JDK 1.6.8 (AMD 64 bit version)
- Netbeans 6.1
- MySQL 5.0
-

With this computer and software, a single test requires between 2 minutes and 30 minutes, depending on the number of dimensions and the complexity of the test.

Computing the SVD (Singular Value Decomposition) of the matrix is the dominant factor in run time. The early versions of EMMA used a Java SVD package to compute the SVD. The first experiment using more than test data ran for 10 days without results, clearly unacceptable. To overcome this problem, EMMA now uses the SVDPACKC SVD program library (Berry, Do,

O'Brien & Krisna, 1993). In particular, EMMA uses the "las2" method, compiled using the 32-bit version of the Gnu C compiler (gcc) running on Windows under Cygnus. SVDPACKC is freely available from the University of Tennessee.

EMMA uses one other freely available piece of software - the Porter Stemmer (Porter, 2006), which is available in at least nineteen programming languages. EMMA uses the Java version.

6.1 The database

6.1.1 Introduction and motivation

An LSA-based assessment system requires a huge corpus. While many corpora are available online (e.g., the US Department of Defense, Reuters) they do not fulfil the need for corpora comprising Computer Science student essays and related textbook-type learning material. The LSA calibration experiments described in this dissertation would have been impossible had I not been given access to the course material and marked student work for the 2004 presentation of the Open University course: M150 - Data, Computing and Information. I used this corpus to populate a database for use with EMMA as well as for other interested researchers in the department.

To perform assessment, LSA requires two very large corpora: general text relating to the domain being studied and human-marked essays. For the former, I used the M150 course textbooks; for the latter, I used the marked answers from the 2,900 students enrolled in the 2004 presentation of M150.

6.1.2 Requirements for the database

- Multiple uses – The database must be general-purpose – other researchers might want to use the data differently from the LSA project.
- Anonymity – The raw data includes student and marker names. It is essential that all identifying information be removed from the database.

- The database should provide for multiple courses, multiple assessments for each course, multiple questions on each assessment, and multiple marks for each answer.
- There needs to be a provision for four types of information – short answer questions, short answers, text book passages, and comments from markers.
- The database needs to store marks by various markers, including those given by LSA.

6.1.3 Some challenges in creating the database

- The format of the student's submissions is not well specified - they had complete flexibility in how to submit their answers, although they had to be electronic. These raw data were produced during an actual delivery of the course; they were created before the onset of this research and were not structured to suit this research. For example, a file consisting of a student's submission to an exam may contain the questions in any order and may or may not contain the question text. The file itself may be of type rtf, MS Word, MS Works, or PDF. A file containing a tutor's comments may consist of a Word document containing the student's answers with tutor comments in red, or just the tutor comments. The question identifiers were in several formats – for example, "1a", "(1)a", (1a), "1.a.". Sometimes, a student used an incorrect identifier, such as "1a" when the question was actually "1b". Some students repeated the question text while others simply provided their answers. Converting the raw data to a useable format turned out to be a huge challenge. I used a program that attempted to isolate the questions and answers. However, I couldn't get good enough results automatically and so underwent a lengthy process of hand-checking 1,000 student answers for each of the

18 questions. Funding and time limitations prevented me from hand-checking all of the questions so the answers from the remaining 1,900 students were not used.

- Granularity of marks – For multi-part questions, some of the markers assigned a single mark for each sub-part; others give only the total for all the sub-parts. This inconsistency required me to abandon those answers with aggregated marks.

6.1.4 EMMA database details

Figure 6-1 shows the entity-relationship diagram for the database. The database is in 3rd normal form to eliminate data redundancy and simplify modifications to the data.

The EMMA database has thirteen tables; the primary keys in each table are non-meaningful, i.e., they are computer-generated sequential numbers. Appendix B gives the complete table descriptions and the entity relations among the tables. Following is a list of the tables with a brief description.

1. **MegaChunk** – the original source document
2. **Chunk –** a unit of text, can be any size
3. **LearningResourcePassage –** contains chunks from general documents, e.g. textbooks, or web sites used as resources for the learners
4. **LearningResource –** books or web sites where the chunk comes from
5. **EssayQuestion –** the question (or subpart if applicable) on an assessment
6. **Assessment –** the TMA (tutor marked assessment) or ECA (end of course assignment) or other exam
7. **Course –** the course that provided the essay answers
8. **EssayAnswer -** contains chunks that consist of answers to essay questions, can be either student or tutor generated

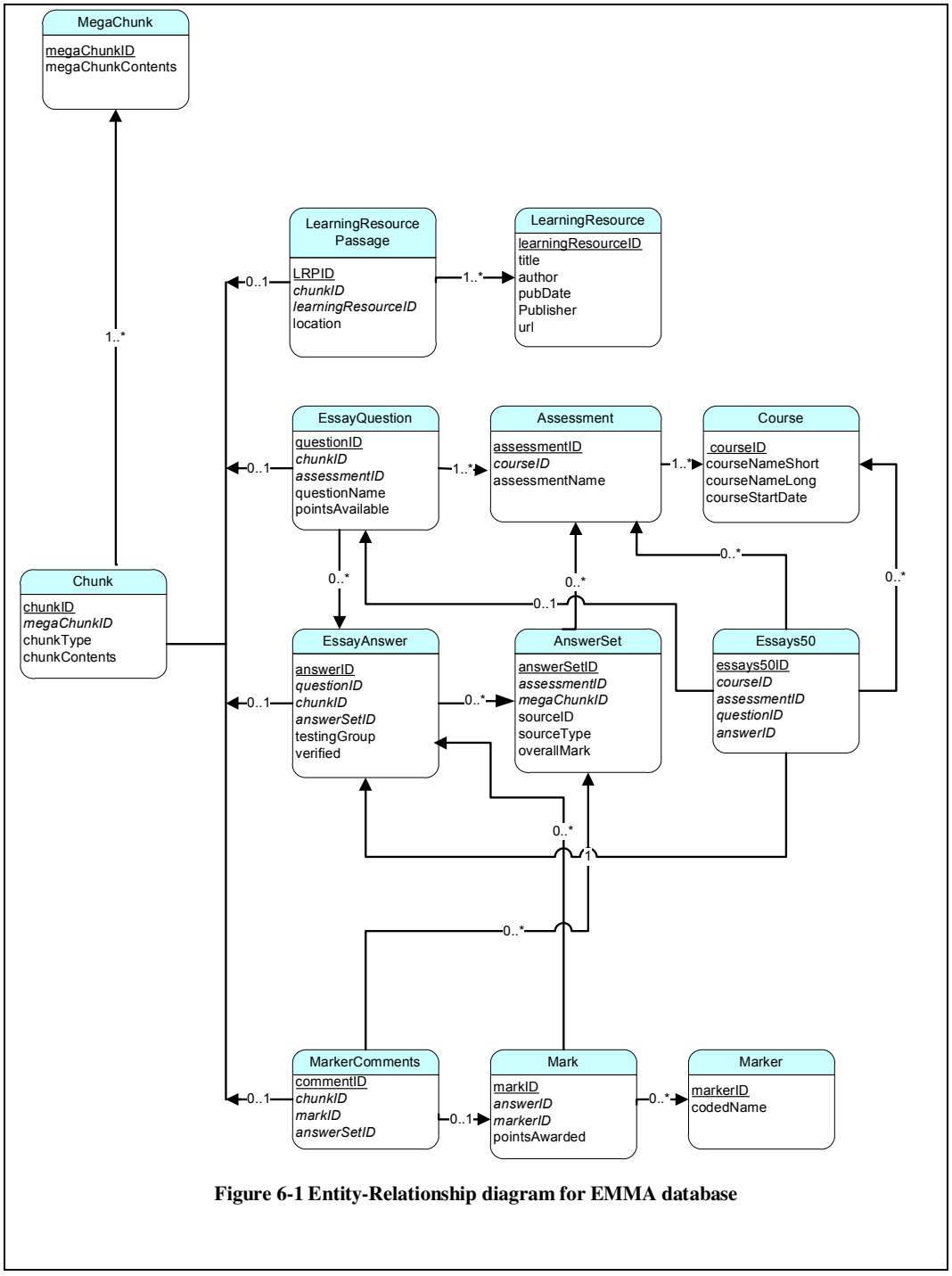


Figure 6-1 Entity-Relationship diagram for EMMA database

9. **AnswerSet -** all of the answers for a particular assessment by a particular student
10. **MarkerComments -** contains chunks that consist of feedback written by the marker for the learners
11. **Mark –** the mark given to an essay
12. **Marker –** the person (or EMMA) who marked the answer
13. **Essays50** – list of essays used in the human inter-rater reliability study done for this dissertation

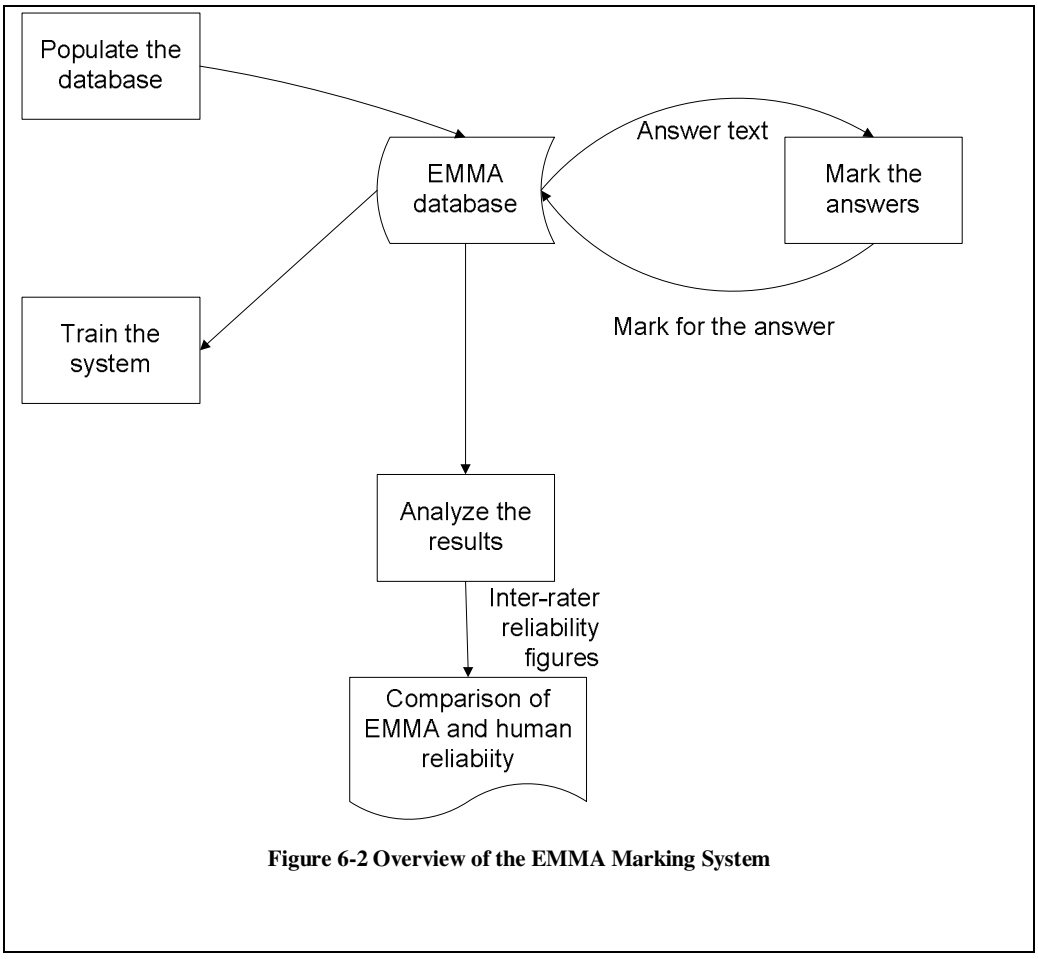


Figure 6-2 Overview of the EMMA Marking System

6.2 The architecture

EMMA consists of four main modules – populate the database, train the system, mark the answers, and evaluate the results. Figure 6-2 shows these modules as rectangles. The arrows to and from the database show how the modules either read from or write to the database.

The database is described in subsection 6.1.4. The *Train the system* module creates the LSA matrix M and uses SVD (singular value decomposition) to create 3 smaller matrices, T, S, and D. The steps involved in this process are described in detail in subsection 2.2.1. The *Mark the*

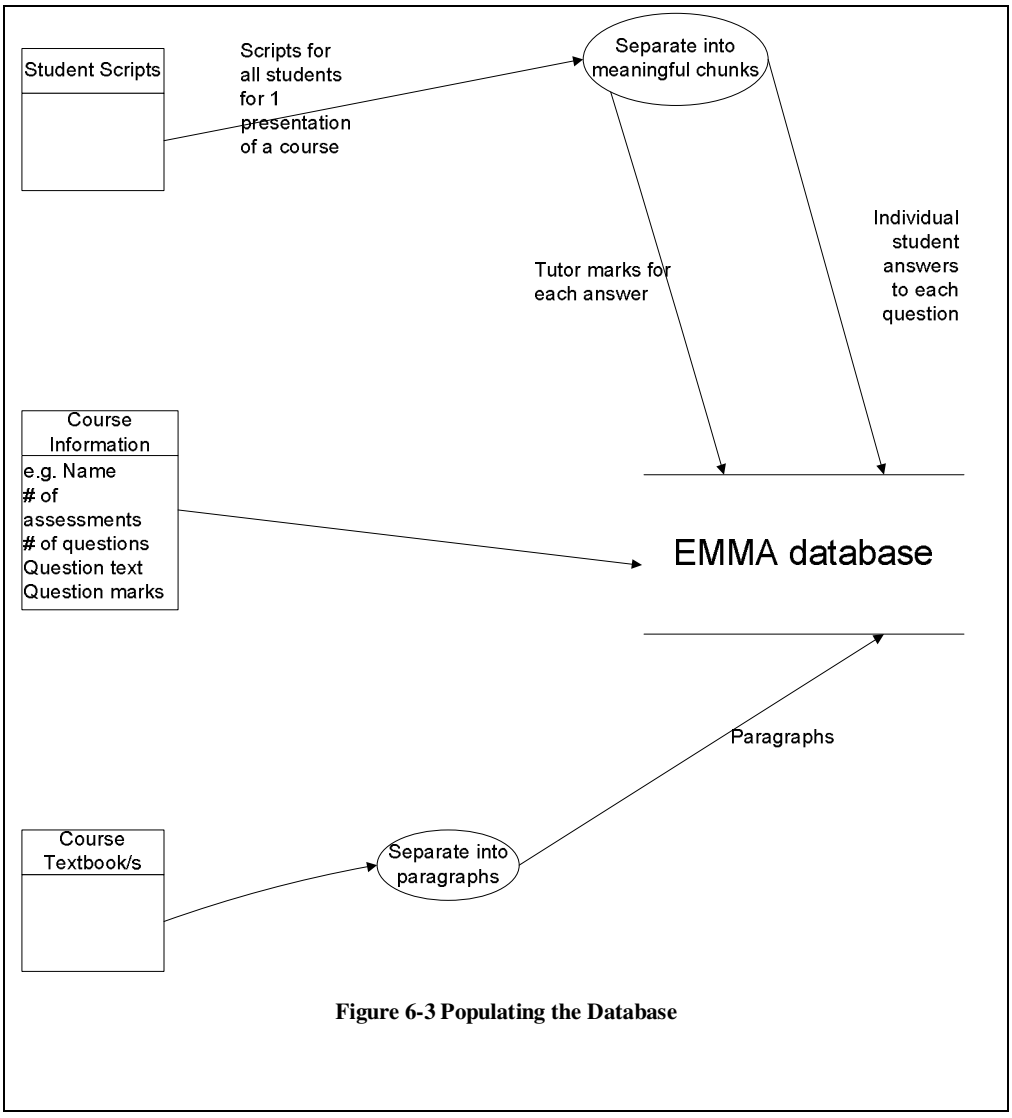


Figure 6-3 Populating the Database

answers module compares the answer-to-be-marked against the previously-marked-answers. It chooses the 5 previously-marked-answers that have the closest cosines to the answer-to-be-marked and that are above a specified threshold. It then calculates the average mark using a proportional weighting, i.e., the marks with higher cosines count more than marks with lower cosines. The final module, *analyze the markers,* calculates the Gwet AC1 inter-rater reliability agreement coefficient for each of the 15 pairs of the six markers (five tutors plus EMMA). Figure 6-3 through Figure 6-6 below show data flow diagrams (DFD) for the four modules. In the DFDs, the ovals represent processes and the arrows show data going into and coming out of a process. The rectangles represent external entities.

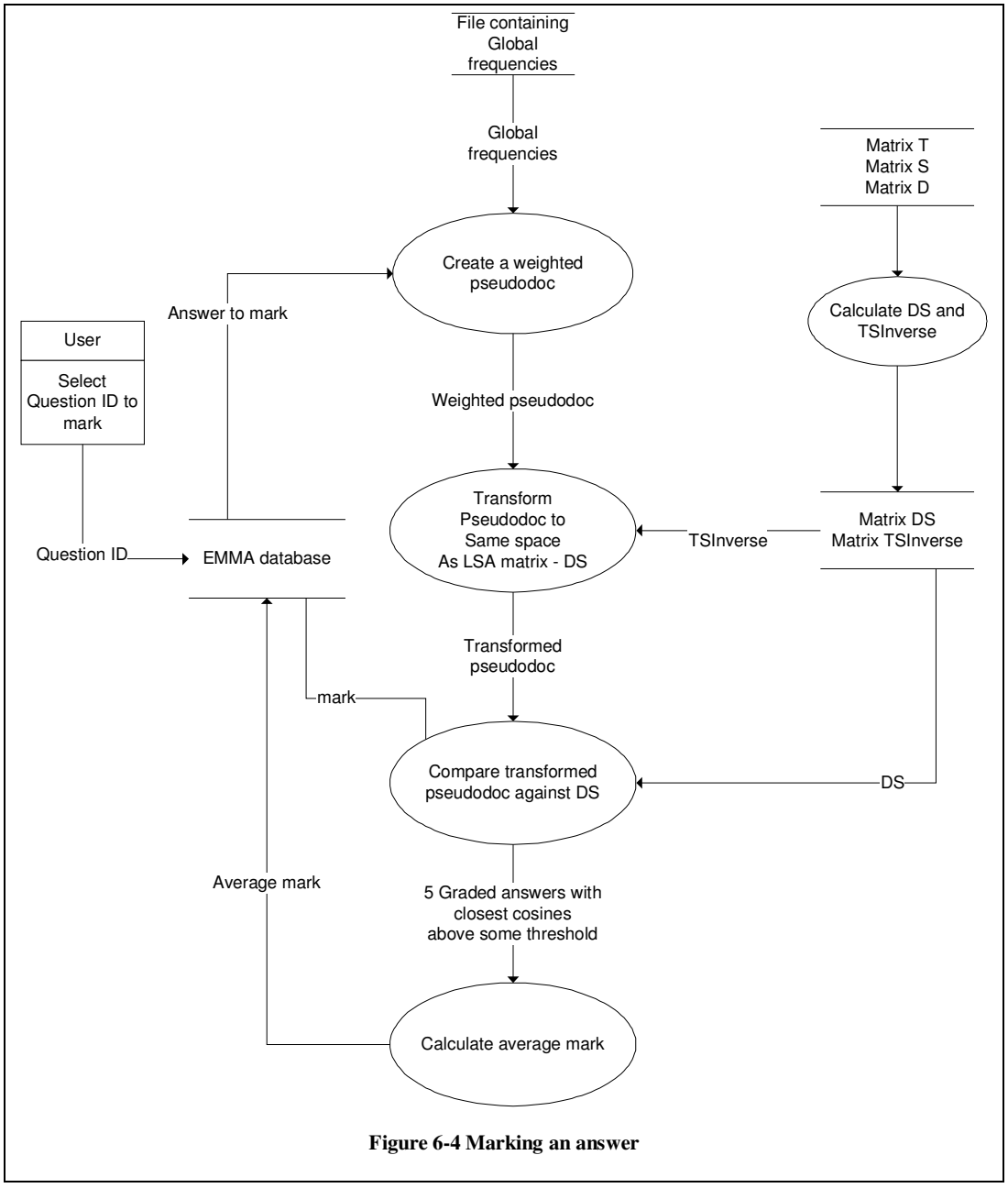


Figure 6-4 Marking an answer

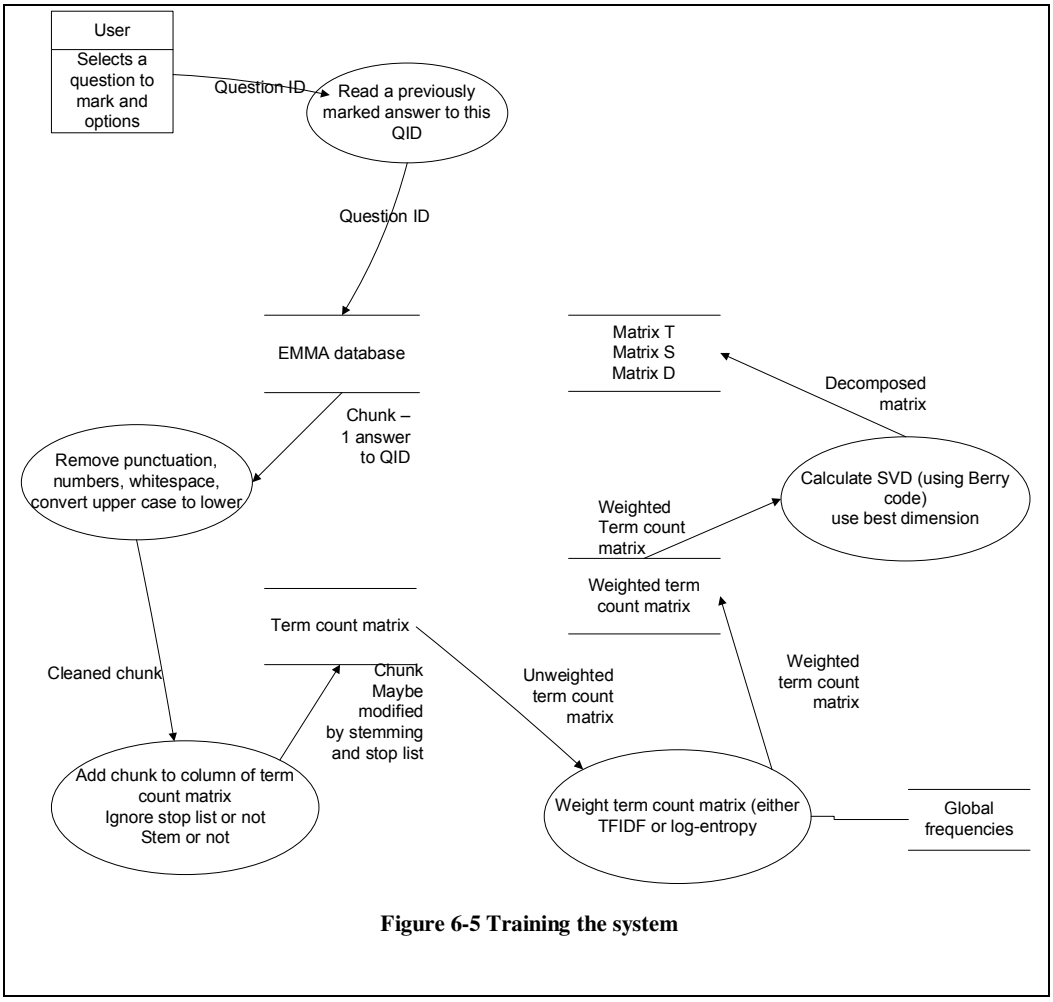


Figure 6-5 Training the system

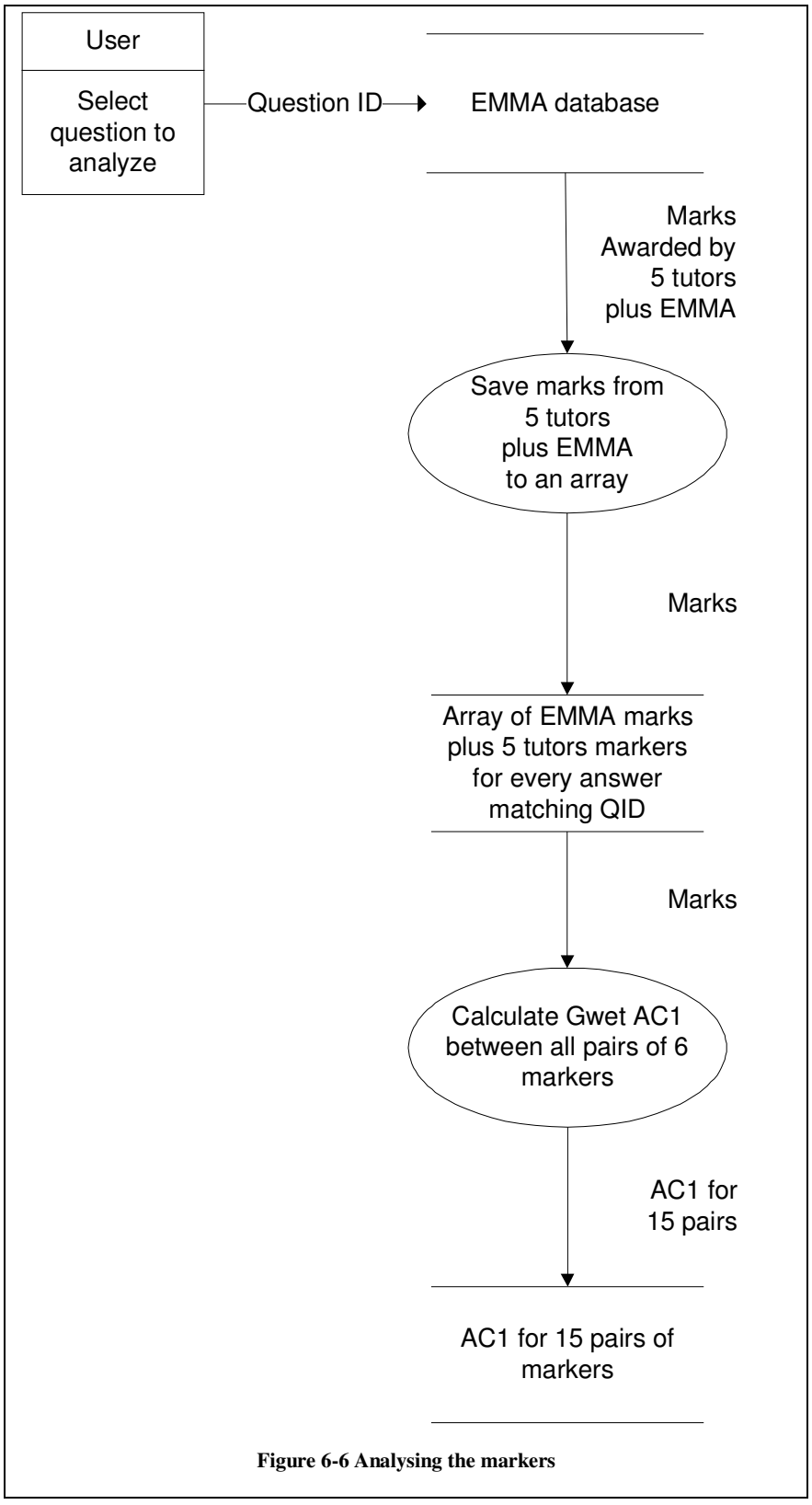


Figure 6-6 Analysing the markers

Chapter 7. Using the Framework to Evaluate LSA Calibrations to Improve the Performance of EMMA

This chapter brings together the various topics discussed so far:

- Chapter 1 – assessment, particularly Computer Assisted Assessment (CAA)
- Chapter 2 – Latent Semantic Analysis (LSA) as a tool for CAA
- Chapter 3 – using the AC1 statistic to evaluate inter-rater reliability (IRR)
- Chapter 5 – the framework for evaluating CAA systems
- Chapter 6 – EMMA, an LSA-based CAA system

The results of Chapter 4, establishing a baseline, i.e., determining how well humans agree, are reviewed in Chapter 8 where the best results obtained from EMMA are compared with the average human IRR.

This chapter uses the framework of Chapter 5 to evaluate the results of calibrating several parameters that are crucial to the operation of an LSA-based CAA system; these parameters were identified in the literature review of Chapter 2. The calibration results provide IRR figures measured by AC1 (presented in Chapter 3) from which I determined the most accurate version of EMMA. The calibration experiments determined which parameters produced the best results as measured by AC1. Figure 7-1 shows once again the framework visualised as a jigsaw puzzle with four areas of importance for evaluating a CAA system: Items Assessed, Training Data, Algorithm-specific Technical Details, and Accuracy. The next sections present all of the information necessary to understand how the experiments of this dissertation were carried out.

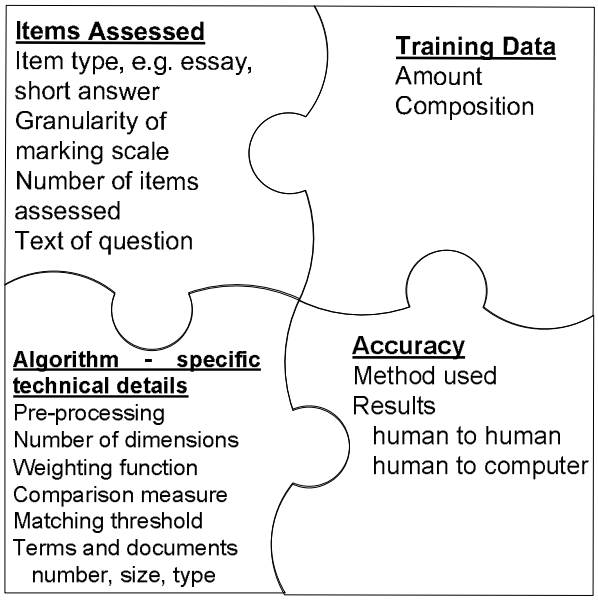


Figure 7-1 Framework for describing and evaluating Computer Assisted Assessment systems

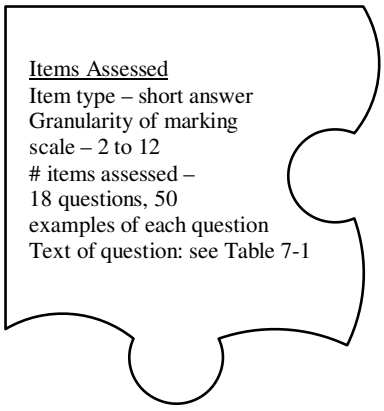


Figure 7-2 Characteristics of items assessed

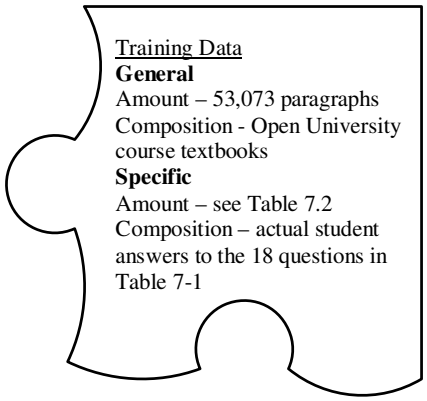


Figure 7-3 Characteristics of training data

The experiments discussed in the following sections calibrate EMMA by finding which parameters produce the best results, as measured by AC1:

- baseline - uses the parameters most often mentioned in the literature
- weighting factor - log-entropy or tfidf
- number of dimensions in the reduced matrix
- retaining or removing stop words
- stemming or not stemming
- the optimum amount of training data
- threshold for measuring if an answer is close enough
- number of answers to average
- proportional or non-proportional averaging

7.1 The Framework

7.1.1 Items assessed

The *Items Assessed* by EMMA, for all the experiments in this chapter, were the same items assessed for the study to compare human markers. The items comprise authentic student answers to 18 questions from an introductory Open University course in computing. Refer to subsection 4.1.3 for the details of the answers. Figure 7-2 summarises the information. Appendix C shows a sample of the questions.

7.1.2 Training data

EMMA uses two types of training data – general and specific. The general training data comprises 53,073 paragraphs (*documents* in LSA terminology) from Open University course text books for both the course being assessed and other computer science courses. The total

number of words (*terms* in LSA terminology) was 829,519. The average length of each paragraph was 16 words.

The specific training data was a set of human marked answers to each of the 18 questions. These marks were authentic tutor marks awarded during the presentation of the course. Table 7-2 shows, per question, the number of answers, the number of words, and the average length of each answer. A specific training document (previously graded answer) had, on average, 109 words.

The number of answers used as specific training data varied by question because of one of the challenges in creating the database mentioned in subsection 6.1.3: due to the wide variation in answer format, it was difficult to computerise the extraction of the answer to a particular question from a student script of all the answers. This problem necessitated manually checking the extraction results; time and cost constraints allowed only a limited number of answers to be verified.

Experiments 7.3 through 7.7 used all of the available specific training data, except for the 60 answers-to-be-marked. Thus, the training data and the testing data did not overlap. Experiment 7.8 varied the amount to find the optimum amount of specific training data per question, which was used for the remainder of the experiments.

7.1.3 Algorithm-specific Technical Details

One of the main claims of this dissertation is that all four pieces of the jigsaw puzzle must be in place to adequately describe and evaluate a CAA. The AC1 figures for EMMA are somewhat meaningless (at least for researchers, as opposed to consumers) without the technical details of the algorithm on which the assessment system is based. In fact, as discussed in Chapter 2, calibrating the system by adjusting the items listed in Figure 7-4 is an integral part of building an LSA-based CAA.

Table 7-1 Text of questions

	Question Text	points
Q1	Name 2 elements of the course materials that will be distributed via the M150 course website?	8
	What is the role of the Study Calendar? What is the cut-off date for TMA02?	
	Find the learning outcomes for M150 which are listed in both the Course Companion and the Course Guide. Write down the learning outcome that you feel you are most interested in achieving and one or two sentences to describe why you have chosen that learning outcome.	
	What does eTMA stand for? What is the name of the document you should read to prepare yourself for submitting an eTMA? Who should you contact with queries about course software?	
Q2	Find the UK AltaVista site. What is its URI? What is the name of the large aquarium in Hull?	12
	Which query led you to the answer? What is the URI of the site?	
	What is the minimum number of intervening web pages you have to visit between the main site and the page that contains the information on the ballan wrasse?	
	List the URI of each intervening web page. How big can a ballan wrasse grow?	
	Does the ballan wrasse page tell you anything about the age a ballan wrasse can reach?	
	What age can a ballan wrasse reach?	
	What is the URI of the web page where you found the information?	
	Which search engine, and which query got you to the page that contained your answer?	
Q3	Explain, with examples, the difference between an analogue and a discrete quantity.	4
Q4	Give an example of a computer standard, explaining its purpose. Why is there a general need for standards in computing?	4
8-12	For each case; write the correct HTML and write one or two sentences about the problem with the original HTML. (The first line is the original HTML. The second line is the desired appearance.)	
Q8	<B>Always look left and right before crossing the road.	4
	Always look left and right before crossing the road.	
Q9	<B>Important!<B>Do <B> not place metal items in the microwave.	4
	Important! Do **not** place metal items in the microwave.	
Q10	<I>It is <B>very</I> </B> important to read this text carefully.	4
	It is ***very*** important to read this text carefully.	
Q11	Things to do: Things to do:	4
	Pack suitcase, </BR>	
	Book taxi. Pack suitcase,	
	Book taxi.	
Q12	More information can be found here.	4
	More information can be found here.	
13-21	Victoria uses her computer to write up a report. When complete, she saves it to the hard disk on her computer. Later she revises her report and saves the final version with the same document name.	
Q13	Considering the contents of the report as data, at what point does the data become persistent?	2
Q14	What happens to the first saved version of the document?	2
Q15	Suggest an improvement in Victoria's work practice, giving a reason for your answer.	2
Q16	Give two examples of persistent storage media other than the hard disk.	2
Q17	Victoria then wishes to email a copy of her report, which includes data on identifiable individuals, to John, a work colleague at her company's Birmingham office. Write two sentences to explain the circumstances under which, within UK law, she may send the report.	2
Q18	Explain briefly the property of internet email that allows the contents of the report to be sent as an attachment rather than as text in the body of the email message.	2
Q19	John's email address is John@Birmingham.office.xy.uk Which parts of the address are: the user name, the name of the domain, the top-level domain?	2
Q20	Victoria then prepares her report for publication on a website. In no more than 100 words, explain what she has to take into account when making her report public.	3
Q21	Which of the following should she publish on the website with her report and why? Company address, personal telephone number, email address	3

Table 7-2 Specific training data – previously human-marked answers

Specific Training Data			
Question	# answers	Total Terms (words)	Average Terms (words) Per Answer
1	2,933	883,062	301
2	2,922	887,700	304
3	3,024	481,059	159
4	2,991	422,023	141
8	936	59,876	64
9	936	70,944	76
10	934	73,764	79
11	936	99,068	106
12	925	71,045	77
13	1,751	64,909	37
14	932	56,630	61
15	927	78,083	84
16	928	29,041	31
17	917	67,261	73
18	922	76,284	83
19	927	35,436	38
20	910	119,183	131
21	915	114,685	125
Total for all questions		3,690,053	1,970
Average terms per q		205,003	109
Standard Deviation		276,862	79

7.1.4 Accuracy

All of the studies of this chapter used the same technique to evaluate the accuracy. I used Gwet's AC1 statistic to measure the IRR by comparing EMMA's results to each of the five human markers of Chapter 4 and then averaged the results. Figure 7-5 gives the method; the AC1 figures, per question, are shown in Table 7-4 through Table 7-12.

In order to determine whether the results are significant, the experimenter must do the calculations described by Gwet (2001b p. 6). If AC1 divided by the standard error (Gwet, 2001a pp. 52-54) is greater than 1.5, then the AC1 obtained is significant at the 95% level. Table 7-3

shows that 88% of the rater pairs have statistically significant results at the 95% level. By studying the table, one can see that the results are not significant when the AC1 is 0.12 or less. This result suggests that very low AC1 figures should be ignored. Since this chapter uses higher AC1 amounts to establish which rater pairs are more closely related, the few cases yielding very low AC1 amounts can safely be ignored.

Table 7-3 The AC1 results are significant at the 95% level for 88% of the rater pairs

QID	Pair	AC1	std err	AC1 / std err	> 1.5?
1	0-1	0.40	0.08	5.16	yes
	0-2	0.29	0.07	3.93	yes
	0-3	0.39	0.08	5.11	yes
	0-4	0.28	0.07	3.83	yes
	0-5	0.31	0.07	4.09	yes
2	0-1	0.63	0.07	8.78	yes
	0-2	0.60	0.07	8.37	yes
	0-3	0.65	0.07	9.24	yes
	0-4	0.46	0.07	6.16	yes
	0-5	0.58	0.07	8.03	yes
3	0-1	0.72	0.07	10.56	yes
	0-2	0.88	0.05	18.15	yes
	0-3	0.83	0.05	15.18	yes
	0-4	0.24	0.08	3.01	yes
	0-5	0.79	0.06	12.94	yes
4	0-1	0.02	0.07	0.29	no
	0-2	0.54	0.08	6.57	yes
	0-3	0.58	0.08	7.37	yes
	0-4	0.52	0.08	6.46	yes
	0-5	0.14	0.08	1.80	yes
8	0-1	0.68	0.07	9.56	yes
	0-2	0.88	0.05	18.15	yes
	0-3	0.85	0.05	16.49	yes
	0-4	0.70	0.07	10.12	yes
	0-5	0.92	0.04	23.01	yes
9	0-1	0.43	0.08	5.33	yes
	0-2	0.88	0.05	18.04	yes
	0-3	0.67	0.07	9.44	yes
	0-4	0.58	0.08	7.52	yes
	0-5	0.67	0.07	9.40	yes
10	0-1	0.14	0.08	1.86	yes
	0-2	0.53	0.08	6.83	yes
	0-3	0.36	0.08	4.42	yes
	0-4	0.25	0.08	3.18	yes
	0-5	0.48	0.08	5.89	yes
11	0-1	0.28	0.08	3.30	yes
	0-2	0.33	0.08	4.04	yes
	0-3	0.51	0.08	6.22	yes
	0-4	0.44	0.08	5.22	yes
	0-5	0.48	0.08	5.71	yes
12	0-1	0.30	0.08	3.58	yes
	0-2	0.50	0.08	6.20	yes
	0-3	0.46	0.08	5.64	yes
	0-4	0.06	0.07	0.86	no
	0-5	0.54	0.08	6.76	yes

QID	Pair	AC1	std err	C1 / std e	> 1.5?
13	0-1	0.98	0.02	47.52	yes
	0-2	0.98	0.02	47.52	yes
	0-3	0.98	0.02	47.52	yes
	0-4	0.94	0.04	26.32	yes
	0-5	0.96	0.03	32.92	yes
14	0-1	0.84	0.06	15.32	yes
	0-2	0.89	0.05	18.86	yes
	0-3	0.84	0.06	15.32	yes
	0-4	0.87	0.05	16.99	yes
	0-5	0.84	0.06	15.16	yes
15	0-1	0.16	0.09	1.68	yes
	0-2	0.19	0.09	2.04	yes
	0-3	0.47	0.09	5.37	yes
	0-4	0.15	0.10	1.58	yes
	0-5	0.00	0.09	0.00	no
16	0-1	0.94	0.04	25.79	yes
	0-2	0.91	0.04	21.79	yes
	0-3	0.91	0.04	21.79	yes
	0-4	0.91	0.04	21.79	yes
	0-5	0.94	0.04	25.79	yes
17	0-1	0.28	0.10	2.76	yes
	0-3	0.12	0.09	1.24	no
	0-4	0.00	0.09	0.00	no
	0-5	0.24	0.10	2.36	yes
18	0-1	0.52	0.09	5.85	yes
	0-2	0.46	0.09	5.03	yes
	0-3	0.49	0.09	5.42	yes
	0-4	0.46	0.09	5.01	yes
	0-5	0.21	0.10	2.24	yes
19	0-1	0.10	0.09	1.12	no
	0-4	0.07	0.09	0.79	no
	0-5	0.08	0.09	0.86	no
20	0-1	0.00	0.07	0.00	no
	0-4	0.00	0.07	0.00	no
	0-5	0.51	0.08	6.15	yes
21	0-1	0.37	0.09	4.08	yes
	0-4	0.39	0.09	4.30	yes
	0-5	0.29	0.09	3.18	yes

Note: AC1 is statistically significant when
AC1 divided by the standard error > 1.5

7.2 The experiments

This section gives the results of the following calibrations to determine how close EMMA could match human markers:

- a baseline
- varying the weighting function
- varying the number of dimensions
- retaining and removing stop words
- stemming and non-stemming
- varying the amount of training data
- varying the threshold of similarity, i.e., how close must a submitted answer match a pre-marked answer to be considered similar?
- varying the number of similar answers whose marks are averaged to determine the mark awarded by EMMA
- using non-proportional averaging

Each of these calibration efforts is discussed in the rest of chapter.. The *Items Assessed* piece of the jigsaw puzzle remains the same as shown in Figure 7-2 and is omitted from the descriptions for the rest of the chapter. The remaining three pieces are reported for each calibration.

I evaluated each of the studies for statistical significance and effect size. If a result is statistically significant with a p value of .05, it means that there is a 95% probability that the test statistic is real and not due to chance (Field, 2005 p. 28). The effect size gives the importance of the result. It is "a standardized measure of the magnitude of the observed effect" (Field, 2005 p. 32). Appendix F gives details about these two statistics.

7.3 Study to establish a baseline

I began my calibration experiments by establishing a baseline using parameters gleaned from the literature review (Chapter 2) and shown in Figure 7-4. Table 7-4 shows the results of using these baseline parameters. The table reveals a wide range of IRR, from a low of 0.07 for question 20 (Q20) to a high of 0.92 for Q16. The mean IRR for the 18 questions is 0.45 with a standard deviation of 0.25.

7.4 Study to determine the optimum weighting function

The next experiment varied the weighting function. The two most common weighting functions in the literature are log-entropy and term frequency inverse document frequency (tfidf).

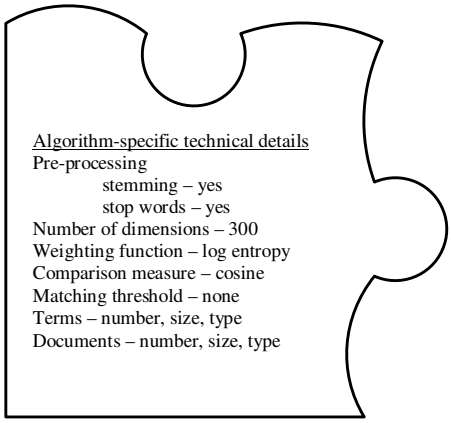


Figure 7-4 Algorithm - specific technical details - the choices for the baseline

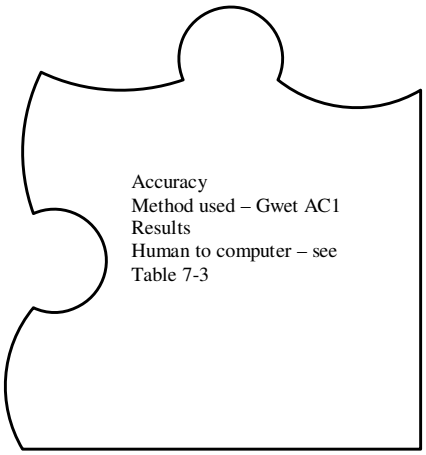


Figure 7-5 Accuracy

Table 7-4 Baseline results by question

System Name	Items Assessed				Algorithm-specific Technical Details												Accuracy	
										Terms			Documents					
	Type of Item	Granularity of Marking Scale	# of items assessed	Question Number (See Table 7.1 for text of question)	preprocessing	# dimensions	weighting function	comparison measure	matching threshold	Number	Size	Type	Number	Size	Type		method used	AC1 using log entropy
EMMA	short answer	0-8	50	1	stemming, stop words	300	log / entropy and tfidf	cosine	none	13,983	1 word	text	56,006	1 paragraph	text		used AC1 to compare EMMA with 5 human markers	0.37
		0-12		2						15,041			55,995					0.73
		0-4		3						13,674			56,073					0.62
		0-4		4						13,498			56,064					0.24
		0-4		8						12,431			54,009					0.70
		0-4		9						12,481			54,009					0.67
		0-4		10						12,517			54,007					0.29
		0-4		11						12,555			54,009					0.32
		0-4		12						12,658			53,998					0.37
		0-2		13						12,374			54,824					0.78
		0-2		14						12,365			54,005					0.69
		0-2		15						12,572			54,000					0.31
		0-2		16						12,380			54,001					0.92
		0-2		17						12,398			53,990					0.10
		0-2		18						12,422			53,995					0.32
		0-2		19						12,372			54,000					0.36
		0-3		20						12,564			53,983					0.07
		0-3		21						12,596			53,988					0.19
																	mean	0.45
																	std dev	0.25

7.4.1 Log-entropy

One of the original LSA investigators (Dumais, 1991) recommended using log-entropy weighting, which is a form of local weighting times global weighting. Local weighting is the most basic form of term weighting; it is defined as tf_{ij} (the number of times term i is found in document j) dampened by the log function: local weighting = 1 + log (tf_{ij}). This dampening reflects the fact that a term that appears in a document *x* times more frequently than another term is not *x* times more important. Global weighting quantifies the assumption that a term appearing in many documents is less important than a term appearing in fewer documents. The log-entropy term weight for term *i* in doc *j* is:

Equation 7-1 Log-entropy weighting function

$$TermWeightLogEntropy_{ij} = \log\left(1+tf_{ij}\right) * \left[1 - \frac{\sum \frac{tf_{ij}}{gf_i} * \log\frac{tf_{ij}}{gf_i}}{\log(numdocs)}\right]$$

where

tf_{ij} – term frequency – the frequency of term i in document j

gf_i – global frequency – the total number of times term i occurs in the whole collection

7.4.2 tfidf

More recently, Sebastiani claimed the most common weighting is tfidf, or term frequency inverse document frequency (2002). Like log-entropy, tfidf gives more weight to terms that appear frequently in a document and less weight to terms that appear frequently throughout the corpus. The tfidf weighting for term i in document j is:

Equation 7-2 tfidf - Term frequency inverse document frequency weighting function

$$tfidf_{ij} = (tf_{ij})\log\left(\frac{numDocs}{gf_i}\right)$$

where tf_{ij} - term frequency - denotes the number of times term i occurs in document j

numDocs denotes the number of documents in the collection

gf_i - global frequency of term i denotes the number of docs in which term i occurs

7.4.3 The term weighting study

Dumais recommended the use of log-entropy weighting for LSI based on her results from the field of Information Retrieval (IR). Sebastiani was reporting on text categorization, of which essay assessment can be seen as a sub-part. I think the choice of weighting function for LSA based CAA systems should be grounded in a comprehensive analysis of assessment rather than retrieval. LSI is a tool for IR and IR uses *precision* and *recall* as measures of success. Subsection 3.4.1 detailed why these measures are not relevant for assessment. In brief, *recall* measures how many correct matches are retrieved. For assessment, only one match is necessary. *Precision* answers the question "Is the retrieved document correct?" with a *yes* or *no*. LSA

Table 7-5 Results, by question, of term weighting study

System Name	Items Assessed				Algorithm-specific Technical Details												Accuracy		
										Terms			Documents						
	Type of Item	Granularity of Marking Scale	# of items assessed	Question Number (See Table 7.1 for text of question)	preprocessing	# dimensions	weighting function	comparison measure	matching threshold	Number	Size	Type	Number	Size	Type	method used	AC1 using log entropy	AC1 using tfidf
EMMA	short answer	0-8	50	1	stemming, stop words	300	log / entropy and tfidf	cosine	none	13,983	1 word	text	56,006	1 paragraph	text	used AC1 to compare EMMA with 5 human markers	0.37	0.28
		0-12		2						15,041			55,995				0.73	0.58
		0-4		3						13,674			56,073				0.62	0.62
		0-4		4						13,498			56,064				0.24	0.37
		0-4		8						12,431			54,009				0.70	0.81
		0-4		9						12,481			54,009				0.67	0.62
		0-4		10						12,517			54,007				0.29	0.40
		0-4		11						12,555			54,009				0.32	0.25
		0-4		12						12,658			53,998				0.37	0.40
		0-2		13						12,374			54,824				0.78	0.95
		0-2		14						12,365			54,005				0.69	0.81
		0-2		15						12,572			54,000				0.31	0.22
		0-2		16						12,380			54,001				0.92	0.92
		0-2		17						12,398			53,990				0.10	0.16
		0-2		18						12,422			53,995				0.32	0.32
		0-2		19						12,372			54,000				0.36	0.20
		0-3		20						12,564			53,983				0.07	0.05
		0-3		21						12,596			53,988				0.19	0.23
																mean	0.45	0.46
																std dev	0.25	0.28

incorporates the concept of *degrees* of precision by its cosine comparison measure; thus precision can be more fine-grained with LSA. For these reasons, I believe a study comparing log-entropy versus tfidf is needed. To my knowledge, no such study has been conducted. I filled this gap by using tfidf to repeat the experiments I had previously carried out for the baseline, which used log-entropy as the weighting function.

Table 7-5 shows the results of the term weighting study. Seven of the 18 questions had a better IRR when log-entropy was the weighting function; 8 questions showed the reverse, while 3 remained the same. The mean for tfidf weighting is slightly higher than for log-entropy (0.46 versus 0.45) and the standard deviation for tfidf is slightly higher (0.28 versus 0.25). The t test showed that these differences were not statistically significant ($t(17) = -.339$, $p = .3695$). Although not statistically significant, this study suggests that the tfidf weighting function gives slightly better results than does log-entropy. Therefore, the remaining studies use the tfidf weighting function. (This finding needs further study with different data.)

7.5 Study to determine the number of dimensions

The next step in calibrating EMMA was to vary the number of dimensions of the middle matrix in the LSA SVD calculation (M = TSD) as discussed in subsection 2.3. The most commonly mentioned number in the literature is 300; the baseline used this number. However, various figures are reported in the literature, even contradictory figures by the same researcher in the same year: 300 (Landauer & Dumais, 1997) and 1500 (Landauer, Laham, Rehder & Schreiner, 1997). I re-computed the experiments using dimensions ranging from 10 to 90 with increments of 10 and then from 100 to 900 with increments of 100. I used the tfidf weighting function because study 7.4 suggested that the results would be slightly better using tfidf than using log-entropy.

Table 7-6 shows the results for each question. The optimum number of dimensions ranged from a low of 10 to a high of 900. The improvements per question by using the optimum dimension are shown in the rightmost column; they range from a low of no improvement for Q8 and Q12 to 122% improvement for Q20 with an average of 26% improvement. The t test showed the results were statistically significant and the effect size was large ($t(17) = -5.274$, $p = 0$, $r = .79$). Clearly, determining the optimum number of dimensions on a per question basis can yield more accurate results.

Table 7-6 Results, by question, of varying the number of dimensions

Items Assessed				Algorithm-specific Technical Details												Accuracy				
									Terms			Documents								
Type of Item	Granularity of Marking Scale	# of items assessed	Quest'n (See Table 7.1 for text of question)	preprocessing	# dimensions	weighting function	comparison measure	matching threshold	Number	Size	Type	Number	Size	Type	method used	average AC1 using 300 dimensions	average AC1 using best dimension	Difference between 300 and Best	% Improvement
short answer	0-8	50	1	stemming, no stop words	80	tfidf	cosine	none	13,983	1 word	text	56,006	1 paragraph	text	used AC1 to compare EMMA against the baseline	0.28	0.37	0.09	32%
	0-12		2		80				15,041			55,995				0.58	0.63	0.05	9%
	0-4		3		40				13,674			56,073				0.62	0.68	0.06	10%
	0-4		4		400				13,498			56,064				0.37	0.40	0.03	8%
	0-4		8		300				12,431			54,009				0.81	0.81	0.00	0%
	0-4		9		100				12,481			54,009				0.62	0.65	0.03	5%
	0-4		10		40				12,517			54,007				0.40	0.43	0.03	7%
	0-4		11		60				12,555			54,009				0.25	0.37	0.12	48%
	0-4		12		40				12,658			53,998				0.40	0.40	0.00	0%
	0-2		13		100				12,374			54,824				0.95	0.97	0.02	2%
	0-2		14		400				12,365			54,005				0.81	0.83	0.02	2%
	0-2		15		400				12,572			54,000				0.22	0.35	0.13	59%
	0-2		16		90				12,380			54,001				0.92	0.94	0.02	2%
	0-2		17		10				12,398			53,990				0.16	0.24	0.08	50%
	0-2		18		10				12,422			53,995				0.32	0.48	0.16	50%
	0-2		19		900				12,372			54,000				0.20	0.30	0.10	50%
	0-3		20		70				12,564			53,983				0.09	0.20	0.11	122%
	0-3		21		30				12,596			53,988				0.23	0.26	0.03	13%
															mean	0.46	0.52	0.06	26%
															std dev	0.27	0.25	0.05	0.32

7.6 Study to determine if removing stop words is helpful

Stop words are frequently occurring words, such as *the* and *that*, that are often excluded by NLP researchers doing textual analysis. Appendix E lists the 571 stop words removed in all of the studies up to this point. These are the classic stop words from the Salton and Buckley SMART retrieval system (Onix, 2008). An often-cited reason for using stop words is that they reduce the size of the LSA matrix. I wanted to see if this benefit would affect the accuracy. Table 7-7 shows that removing stop words, in fact, improves the accuracy. The results of retaining stop words ranges from no change for Q17 to a decline in accuracy of 100% for Q19 with an average 19% loss in accuracy. In no case was the accuracy improved. The t test showed the results were statistically significant and the effect size was large ($t(17) = 4.454$, $p = 0$, $r = .73$). Therefore, this

Table 7-7 Results, by question, of removing versus retaining stop words

Items Assessed				Algorithm-specific Technical Details													Accuracy				
									Terms				Documents								
Type of Item	Granularity of Marking Scale	# of items assessed	Quest'n (See Table 7.1 for text of question)	preprocessing	# dimensions	weighting function	comparison measure	matching threshold	Number eliminating Stop Words	Number retaining Stop Words	Size	Type	Number	Size	Type	method used	average AC1 removing stop words	average AC1 retaining stop words	Difference between 300 and Best	% Improvement
short answer	0-8	50	1	stemming, eliminating stop words	80	tfidf	cosine	none	13,983	14,305	1 word	text	56,006	1 paragraph	text	used AC1 to compare EMMA against the baseline	0.37	0.36	-0.01	-3%
	0-12		2		80				15,041	15,370			55,995				0.63	0.55	-0.08	-13%
	0-4		3		40				13,674	13,997			56,073				0.68	0.55	-0.13	-19%
	0-4		4		400				13,498	13,823			56,064				0.40	0.26	-0.14	-35%
	0-4		8		300				12,431	12,756			54,009				0.81	0.78	-0.03	-4%
	0-4		9		100				12,481	12,808			54,009				0.65	0.63	-0.02	-3%
	0-4		10		40				12,517	12,843			54,007				0.43	0.31	-0.12	-28%
	0-4		11		60				12,555	12,883			54,009				0.37	0.25	-0.12	-32%
	0-4		12		40				12,658	12,986			53,998				0.40	0.37	-0.03	-8%
	0-2		13		100				12,374	12,704			54,824				0.97	0.93	-0.04	-4%
	0-2		14		400				12,365	12,693			54,005				0.83	0.67	-0.16	-19%
	0-2		15		400				12,572	12,899			54,000				0.35	0.26	-0.09	-26%
	0-2		16		90				12,380	12,708			54,001				0.94	0.92	-0.02	-2%
	0-2		17		10				12,398	12,726			53,990				0.24	0.24	0.00	0%
	0-2		18		10				12,422	12,751			53,995				0.48	0.43	-0.05	-10%
	0-2		19		900				12,372	12,700			54,000				0.30	0.00	-0.30	-100%
	0-3		20		70				12,564	12,888			53,983				0.09	0.07	-0.02	-22%
	0-3		21		30				12,596	12921			53,988				0.26	0.21	-0.05	-19%
									12827	13153			54498			mean	0.51	0.43	-0.08	-19%
									737	737			867			std dev	0.25	0.27	0.07	0.23

study suggests that removing stop words not only saves computer memory but is more accurate.

7.7 Stemming and non-stemming

Stemming consists of conflating word forms to a common string, e.g., *write, writing, writes, written, writer* would be represented in the corpus as *writ*. Stemming, like removing stop words, is used by NLP researchers doing textual analysis. By stemming terms, the size of the LSA matrix is greatly reduced, thus saving computer memory. The purpose of this experiment was to determine if stemming affected accuracy. All of the experiments thus far used stemming so this experiment did not stem. Table 7-8 shows the results. In two cases, not stemming increases accuracy: 2% for Q1 and 5% for Q10. There is no change for Q8. The remaining questions

Table 7-8 Results by question of not stemming

Items Assessed				Algorithm-specific Technical Details											Accuracy				
									Terms			Documents							
Type of Item	Granularity of Marking Scale	# of items assessed	Quest'n (See Table 7.1 for text of question)	preprocessing	# dimensions	weighting function	comparison measure	matching threshold	Number eliminating Stop Words	Size	Type	Number	Size	Type	method used	average AC1 using stemming	average AC1 without stemming	Difference between stemming and not stemming	% Improvement
short answer	0-8	50	1	removing stop words; stemming compared to non stemming	80	tfidf	cosine	none	13,983	1 word	text	56,006	1 paragraph	text	used AC1 to compare EMMA with and without stemming	0.37	0.38	0.01	2%
	0-12		2		80				15,041			55,995				0.63	0.54	-0.09	-14%
	0-4		3		40				13,674			56,073				0.68	0.61	-0.07	-10%
	0-4		4		400				13,498			56,064				0.40	0.27	-0.13	-32%
	0-4		8		300				12,431			54,009				0.81	0.81	0.00	0%
	0-4		9		100				12,481			54,009				0.65	0.63	-0.02	-3%
	0-4		10		40				12,517			54,007				0.43	0.45	0.02	5%
	0-4		11		60				12,555			54,009				0.37	0.30	-0.07	-18%
	0-4		12		40				12,658			53,998				0.40	0.38	-0.02	-4%
	0-2		13		100				12,374			54,824				0.97	0.86	-0.11	-12%
	0-2		14		400				12,365			54,005				0.83	0.70	-0.13	-16%
	0-2		15		400				12,572			54,000				0.35	0.26	-0.09	-26%
	0-2		16		90				12,380			54,001				0.94	0.92	-0.02	-2%
	0-2		17		10				12,398			53,990				0.24	0.16	-0.08	-35%
	0-2		18		10				12,422			53,995				0.48	0.33	-0.15	-31%
	0-2		19		900				12,372			54,000				0.30	0.25	-0.05	-16%
	0-3		20		70				12,564			53,983				0.09	0.07	-0.02	-20%
	0-3		21		30				12,596			53,988				0.26	0.24	-0.02	-8%
									12827			54498			mean	0.51	0.45	-0.06	-13%
									737			867			std dev	0.25	0.25	0.05	12%

show a decrease in accuracy so that overall, not stemming reduces accuracy by an average of 13%. The t test showed the results were statistically significant and the effect size was large ($t(17) = 4.760$, $p = 0$, $r = .76$).

7.8 Study to find the optimum amount of training data

The next study used the calibration results of the previous studies (best dimension per question, tfidf weighting function, removing stop words and stemming) and varied the amount of specific training data, i.e., human-marked answers. The number of required human-marked answers needed is an important criterion for evaluating the feasibility of using an LSA-based marking system and one which has not received much attention in the literature. One reason to use an automatic marker is to avoid the necessity of human marking. It would be hard to justify on the basis of cost savings the use of an automatic marker that requires more effort to create

Table 7-9 Results of varying the amount of training data

Training Data		Items Assessed				Algorithm-specific Technical Details												Accuracy				
											Terms				Documents								
Size of General Training Data - course text books	of Specific Training Data - human marke	Type of Item	Granularity of Marking Scale	# of items assessed	Quest'n (See Table 7.1 for text of question)	preprocessing	# dimensions	weighting function	comparison measure	matching threshold	Number max training data	Number best training data	Size	Type	Number max training data	Number best training data	Size	Type	method used	average AC1 using max training data	average AC1 using best amount of training data	Difference between max and best	% Improve-ment
53073 paragraphs	2,933	short answer	0-8	50	1	removing stop words; stemming	80	tfidf	cosine	none	13,983	same	1 word	text	56,006	same	1 paragraph	text	used AC1 to compare EMMA with and without stemming	0.37	0.37	0.00	0%
	2,922		0-12		2		80				15,041	same			55,995	same				0.63	0.63	0.00	0%
	3,000		0-4		3		40				13,674	13,611			56,073	55,873				0.68	0.70	0.02	3%
	2,991		0-4		4		400				13,498	13,030			56,064	54,873				0.40	0.43	0.03	7%
	936		0-4		8		300				12,431	same			54,009	same				0.81	0.81	0.00	0%
	936		0-4		9		100				12,481	same			54,009	same				0.65	0.65	0.00	0%
	934		0-4		10		40				12,517	same			54,007	same				0.43	0.43	0.00	0%
	936		0-4		11		60				12,555	12,424			54,009	53,473				0.37	0.48	0.11	30%
	925		0-4		12		40				12,658	same			53,998	same				0.40	0.40	0.00	0%
	1,751		0-2		13		100				12,374	same			54,824	same				0.97	0.97	0.00	0%
	932		0-2		14		400				12,365	same			54,005	same				0.83	0.83	0.00	0%
	927		0-2		15		400				12,572	same			54,000	same				0.35	0.35	0.00	0%
	928		0-2		16		90				12,380	same			54,001	same				0.94	0.94	0.00	0%
	917		0-2		17		10				12,398	same			53,990	same				0.24	0.24	0.00	0%
	922		0-2		18		10				12,422	same			53,995	same				0.48	0.48	0.00	0%
	927		0-2		19		900				12,372	same			54,000	same				0.30	0.30	0.00	0%
	910		0-3		20		70				12,564	12,335			53,983	53,273				0.09	0.12	0.03	33%
	915		0-3		21		30				12,596	12,447			53,988	53,473				0.26	0.35	0.09	35%
											12827				54498				mean	0.51	0.53	0.02	6%
											737				867				std dev	0.25	0.25	0.03	12%

the training data than it would to mark the answers manually. (This analysis ignores the other benefits of automatic markers given in subsection 1.1.4.) It takes less human effort and is thus more practical if *few* answers are required for *good results*.

Table 7-9 shows how using the optimum amount of training data can improve the results from EMMA. Thirteen questions showed the best results when the maximum amount of training data was used. Five questions improved: the best was Q21 which improved by 35%. Overall, using the optimum amount of training data improved results by 6%. The t test showed the results were statistically significant and the effect size was medium ($t(17) = -2.026$, $p = .0295$, $r = .44$).

A surprising result is that more is not always better. This is good news for consumers of LSA-based CAA systems and leads to a further research question – "Will hand-selecting the specific training data improve the results?". By hand-selecting, I mean doing a careful analysis

of the previously-marked-answers to ensure that they are correctly marked and include the full range of marks.

7.9 Varying the threshold of similarity

The experiments so far do not use a threshold of similarity: the marks given to the five closest answers are averaged without regard to the size of the cosine separating the vector of the answer-to-be-marked and the vector of the answer-being-marked. Various LSA researchers (Foltz, Kintsch & Landauer, 1998; Graesser, Wiemer-Hastings, Wiemer-Hastings, Harter & The Tutoring Research Group, 2000) use a threshold. If the cosine is too small (below the threshold) the answer is not used as one of the five closest answers. If no answers are above the threshold, the answer must be marked by hand. This experiment used thresholds of 0.2, 0.4, 0.6 and 0.8 to determine if the accuracy increased without an unacceptable number of answers needing to be marked by hand. The following 18 figures show the results.

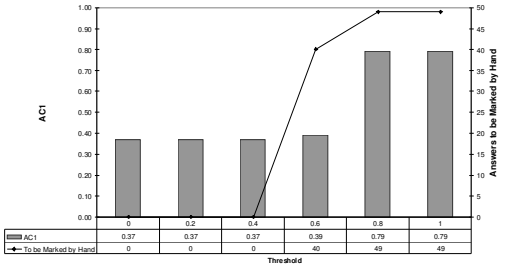


Figure 7-6 AC1 and # to mark by hand per threshold for Q1

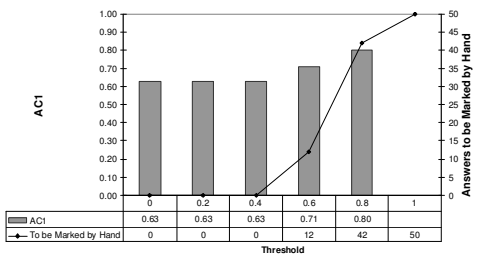


Figure 7-7 AC1 and # to mark by hand per threshold for Q2

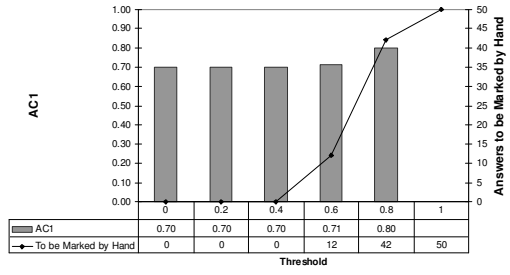


Figure 7-8 AC1 and # to mark by hand per threshold for Q3

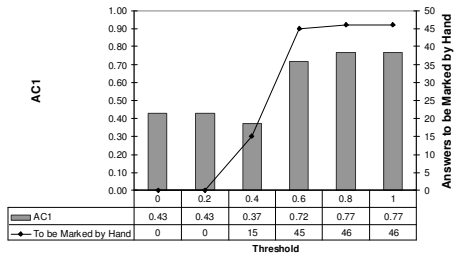


Figure 7-9 AC1 and # to mark by hand per threshold for Q4

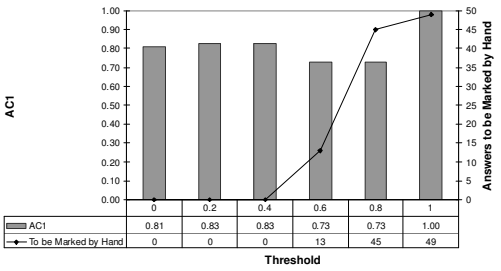


Figure 7-10 AC1 and # to mark by hand per threshold for Q8

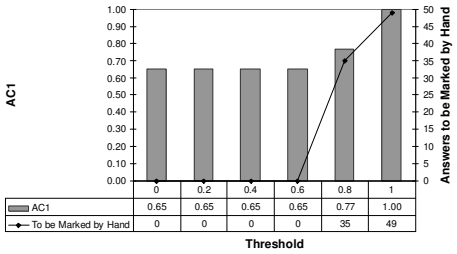


Figure 7-11 AC1 and # to mark by hand per threshold for Q9

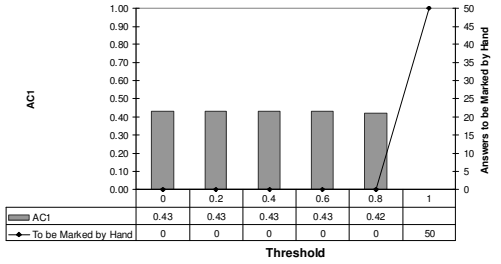


Figure 7-12 AC1 and # to mark by hand per threshold for Q10

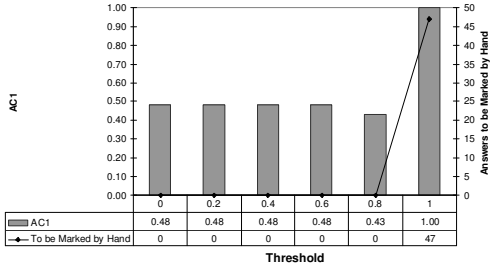


Figure 7-13 AC1 and # to mark by hand per threshold for Q11

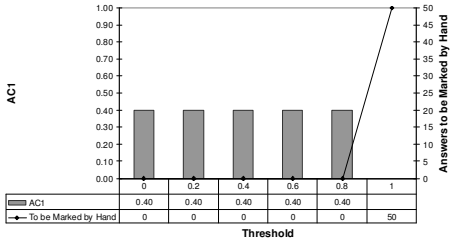


Figure 7-14 AC1 and # to mark by hand per threshold for Q12

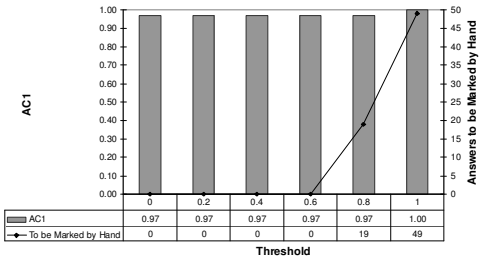


Figure 7-15 AC1 and # to mark by hand per threshold for Q13

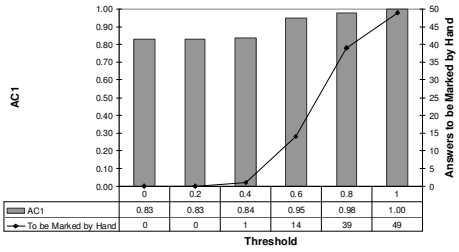


Figure 7-16 AC1 and # to mark by hand per threshold for Q14

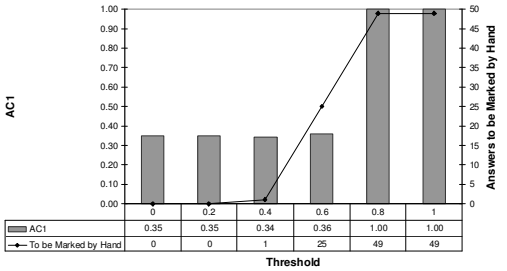


Figure 7-17 AC1 and # to mark by hand per threshold for Q15

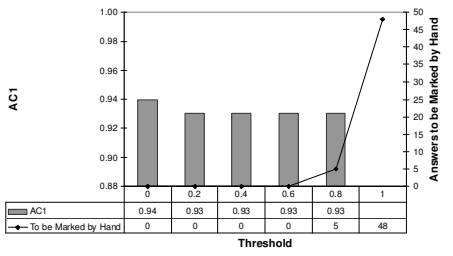


Figure 7-18 AC1 and # to mark by hand per threshold for Q16

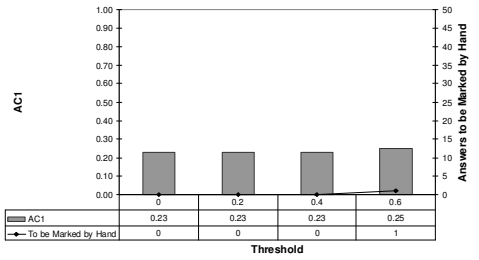


Figure 7-19 AC1 and # to mark by hand per threshold for Q17

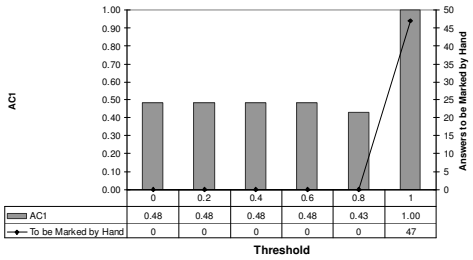


Figure 7-20 AC1 and # to mark by hand per threshold for Q18

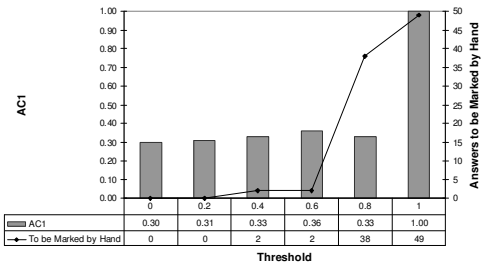


Figure 7-21 AC1 and # to mark by hand per threshold for Q19

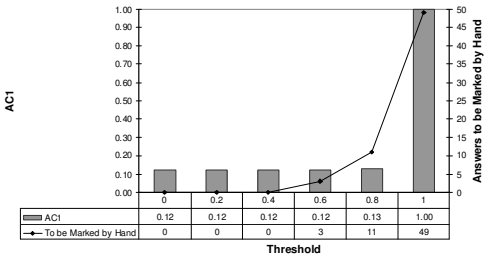


Figure 7-22 AC1 and # to mark by hand per threshold for Q20

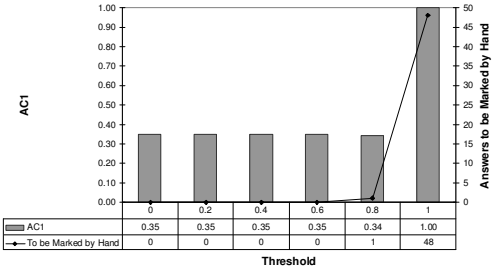


Figure 7-23 AC1 and # to mark by hand per threshold for Q21

The only result that can be generalised across all of the questions is that the higher the matching threshold, the more answers must be marked by hand because EMMA is unable to find any matching answers. The rest of the results of using thresholds are inconsistent. Eight questions (Q1-3, Q9, Q12-14, and Q20) show expected results - as the threshold increases, the AC1 increases or remains the same. Six more questions (Q10, Q15-18. and Q21) also show an increase in AC1 as the threshold increases if one assumes a couple of rounding errors. For example, in Q10, the AC1 is 0.43 for

thresholds 0 through 0.6 but becomes 0.42 at threshold 0.8. For these 14 questions, 0.6 would be an acceptable threshold (highest AC1 and fewer than 25 answers to be marked by hand) except for Q1 where a threshold of 0.6 causes 40 answers to require hand marking. Four questions (Q4, Q8, Q11, and Q19) show unexpected results in that the AC1 occasionally decreases instead of strictly increasing as the threshold increases. This result needs further investigation. I decided to continue using a threshold of 0.0 because of these inconclusive results.

7.10 Varying the number of similar answers whose marks are averaged to determine the mark awarded by EMMA

For the previous experiments, EMMA found the 5 closest marked answers to the answer-being-marked. It then averaged the marks previously given to these answers to determine the mark for the answer-being-marked. Five was chosen randomly as the literature provided little guidance. This experiment varied the number of answers: it used 2, 8, and 10. Table 7-10 summarises the results. Fifteen questions did not change; 3 questions showed an improvement in accuracy by altering the number of close answers to average (Q1 - 11%; Q14 - 4%; Q20 - 8%). Altering these numbers resulted in an overall improvement of 1%. However the t test showed that this result is not statistically significant ($t(17) = -1.641$, $p = .0595$). Since the improvement was modest and not significant, I decided to retain 5 as the number of close answers to average.

7.11 Using non-proportional averaging

The experiments so far used proportional averaging. i.e., the answers that were closer counted more in the averaging than the answers that were less close. Proportional averaging can be seen as fairer than non-proportional averaging because the latter gives each mark equal weight.

Recall that the cosine between the two vectors (previously-marked-answer and answer-to-be-marked) is the closeness metric. Table 7-11 shows the 5 closest marks and their cosines for a hypothetical case. It shows the calculations yielding a mark of 3.5 to be awarded to the answer-to-be-marked. Without using proportional averaging, the mark would be 16/5 = 3.2.

Table 7-10 Number of close answers to average

Items Assessed				Algorithm-specific Technical Details										Accuracy					
Type of Item	Granularity of Marking Scale	# of items assessed	Quest'n (See Table 7.1 for text of question)	preprocessing	# dimensions	weighting function	comparison measure	matching threshold	Number best training data	Size	Type	Number best training data	Size	Type	method used	average AC1 using best training data and no threshold	average AC1 using best number to average	best number to average	Difference between 5 and best	% Improvement
short answer	0-8	50	1	removing stop words; stemming	80	tfidf	cosine	none	13,983	1 word	text	56,006	1 paragraph	text	used AC1 to compare EMMA with and without stemming	0.37	0.41	2	0.04	11%
	0-12		2		80				15,041			55,995				0.63	0.63	5	0.00	0%
	0-4		3		40				13,611			55,873				0.70	0.70	5	0.00	0%
	0-4		4		400				13,030			54,873				0.43	0.43	5	0.00	0%
	0-4		8		300				12,431			54,009				0.81	0.81	5	0.00	0%
	0-4		9		100				12,481			54,009				0.65	0.65	5,8,10	0.00	0%
	0-4		10		40				12,517			54,007				0.43	0.43	2,5	0.00	0%
	0-4		11		60				12,424			53,473				0.48	0.48	5	0.00	0%
	0-4		12		40				12,658			53,998				0.40	0.40	5	0.00	0%
	0-2		13		100				12,374			54,824				0.97	0.97	5,8,10	0.00	0%
	0-2		14		400				12,365			54,005				0.83	0.86	8,10	0.03	4%
	0-2		15		400				12,572			54,000				0.35	0.35	5	0.00	0%
	0-2		16		90				12,380			54,001				0.94	0.94	5	0.00	0%
	0-2		17		10				12,398			53,990				0.24	0.24	5	0.00	0%
	0-2		18		10				12,422			53,995				0.48	0.48	5	0.00	0%
	0-2		19		900				12,372			54,000				0.30	0.30	5	0.00	0%
	0-3		20		70				12,335			53,273				0.12	0.13	10	0.01	8%
	0-3		21		30				12,447			53,473				0.35	0.35	2,5,10	0.00	0%
															mean	0.53	0.53		0.00	1%
															std dev	0.25	0.24		0.01	3%

Table 7-11 Proportional averaging

	Cosine	Mark	Proportion contributing to the average mark
	0.7	4	(.7/2) * 4 = 1.4
	0.6	3	(.6/2) * 3 = .9
	0.4	4	(.4/2) * 4 = .8
	0.1	2	(.1/2) * 2 = .1
	0.2	3	(.2/2) * 3 = .3
Total	2	16	3.5

The calculations to find the proportionally averaged mark are more complicated than those for non-proportional averaging. The purpose of this experiment was to determine whether the results of proportional averaging justify the added computational complexity.

Table 7-12 Results of using non-proportional averaging

Items Assessed				Algorithm-specific Technical Details												Accuracy				
Type of Item	Granularity of Marking Scale	# of items assessed	Quest'n (See Table 7.1 for text of question)	preprocessing	# dimensions	weighting function	comparison measure	matching threshold	Number best training data	Size	Type	Number best training data	Size	Type	method used	average AC1 using proportional averaging	average AC1 using non-proportional averaging	Difference between proportional and non-proportional	% Improvement
short answer	0-8	50	1	removing stop words; stemming	80	tfidf	cosine	none	13,983	1 word	text	56,006	1 paragraph	text	used AC1 to compare EMMA with and without stemming	0.37	0.36	-0.01	-3%
	0-12		2		80				15,041			55,995				0.63	0.64	0.01	2%
	0-4		3		40				13,611			55,873				0.70	0.70	0.00	0%
	0-4		4		400				13,030			54,873				0.43	0.36	-0.07	-16%
	0-4		8		300				12,431			54,009				0.81	0.80	-0.01	-1%
	0-4		9		100				12,481			54,009				0.65	0.65	0.00	0%
	0-4		10		40				12,517			54,007				0.43	0.45	0.02	5%
	0-4		11		60				12,424			53,473				0.48	0.49	0.01	2%
	0-4		12		40				12,658			53,998				0.40	0.40	0.00	0%
	0-2		13		100				12,374			54,824				0.97	0.97	0.00	0%
	0-2		14		400				12,365			54,005				0.83	0.83	0.00	0%
	0-2		15		400				12,572			54,000				0.35	0.34	-0.01	-3%
	0-2		16		90				12,380			54,001				0.94	0.93	-0.01	-1%
	0-2		17		10				12,398			53,990				0.24	0.23	-0.01	-4%
	0-2		18		10				12,422			53,995				0.48	0.51	0.03	6%
	0-2		19		900				12,372			54,000				0.30	0.21	-0.09	-30%
	0-3		20		70				12,335			53,273				0.12	0.11	-0.01	-8%
	0-3		21		30				12,447			53,473				0.35	0.32	-0.03	-9%
															mean	0.53	0.52	-0.01	-3%
															std dev	0.25	0.25	0.03	8%

Table 7-12 shows the results of using non-proportional averaging: 5 questions showed no change, 4 showed an improvement and 9 showed a decreased accuracy for an average decrease in accuracy of 3%. The t test showed that these results are not statistically significant ($t(17) = 1.468$, $p = .080$). I decided that, without evidence to the contrary, the increased computational complexity of using proportional averaging is justified by the possible increased accuracy, even if it is a modest increase.

7.12 Summary

This chapter has reported the results of calibrating EMMA. Table 7-12 shows the best results achieved. Table 7-13 shows the improvements of calibrating EMMA. Finding the right dimension accounts for a 26% improvement followed by a 24% improvement for using the optimum amount of training data. The standard pre-processing techniques, removing stop words and stemming, improved the results by 19% and 13% respectively. Minimal improvements were made by using proportional averaging (3%), tfidf weight function (2%) and choosing the best number of marked answers to average.

The next chapter will compare the best results obtained by EMMA with the accuracy of humans to draw conclusions about the overall acceptability of EMMA as a marking tool.

Table 7-13 Summary of results

Number	Study	Results	Improvement (average)	Statistically significant (level)	Effect Size
1	Baseline				
2	Weighting function	tfidf is slightly better than log-entropy	2%	no	
3	Number of dimensions	choosing the best dimension is better than using 300 - the standard	26%	yes (>99%)	large
4	Stop words	removing stop words is better	19%	yes (>99%)	large
5	Stemming	stemming is better	13%	yes (>99%)	large
6	Amount of training data	choosing the best amount of training data is better	24%	yes (95%)	medium
7	Matching threshold	inconclusive			
8	Number of answers to average	5 yields the best accuracy	1%	no	
9	Prop. vs non-proportional averaging	proport. slightly better than non	3%	no	

Chapter 8. Evaluation of EMMA, a Roadmap for Future Research and Conclusion

8.1 Introduction

In order for an automated marking tool to be accepted, it must produce marks at an acceptable level of accuracy. Automatic marks must correspond with human marks about as well as human marks correspond with other human marks. If humans agreed with each other all the time, EMMA would have to show perfect agreement with human makers. Chapter 4 provided evidence to support the widely held belief that human markers do not agree with other humans all the time. Thus, EMMA needs to be *good enough*, not perfect. EMMA is good enough if it matches or exceeds the agreement of humans.

This chapter describes the evaluation of EMMA, suggests future research to improve EMMA, and draws conclusions.

Subsection 8.2.1 evaluates EMMA by looking at the average IRR among the five human markers from the study described in Chapter 4, per question, and comparing it with the average IRR among EMMA and humans, per question. Subsection 8.2.2 looks in more detail at the IRR for each pair of raters, per question. Subsection 8.3 analyses the evaluations from the previous subsections. Subsection 8.4 gives implications for potential users of CAA systems. Subsection 8.5 discusses future research, subsection 8.6 summarises the results, subsection 8.7 gives advice for future researchers and subsections 8.8 and 8.9 close the dissertation with concluding remarks.

8.2 Evaluations

This section describes the results of two evaluation methods. The first looks at the average inter-rater reliability (IRR); the second is a more detailed comparison of all of the rater pairs to determine the worst marker.

8.2.1 Evaluation 1

The first evaluation compared the average IRR of human markers per question (Chapter 4) with the best results obtained from EMMA (Chapter 7). The average IRR for human markers is the average agreement of the ten pairs of human markers as measured by Gwet's AC1. Thus, it is a measurement, per question, of how well the humans agreed with each other. The average IRR for EMMA is the average IRR of EMMA and each of the five human markers, per question. Table 8-1 shows the results.

The table shows that, for 13 questions, the average agreement among EMMA and humans was worse than among humans alone. However, for the 5 shaded questions, the reverse was true; for these questions, the average agreement among EMMA and humans was actually better than among humans. Averaged over all of the 18 questions, there was a 7% drop in agreement by using a computer program to mark the questions. Therefore, on average, EMMA does not agree with human markers as well as they agree with each other although for 5 out of the 18 questions tested (28%), EMMA's marks were more consistent than the humans alone.

Table 8-1 Average IRR of Humans compared to EMMA

Question	Humans	EMMA	diff between EMMA and humans	%drop in IRR by using EMMA
1	0.52	0.37	0.15	29%
2	0.80	0.63	0.17	21%
3	0.58	0.70	-0.12	-21%
4	0.32	0.43	-0.11	-34%
8	0.82	0.81	0.01	1%
9	0.70	0.65	0.05	7%
10	0.46	0.43	0.03	7%
11	0.54	0.48	0.06	11%
12	0.39	0.40	-0.01	-3%
13	0.95	0.97	-0.02	-2%
14	0.91	0.83	0.08	9%
15	0.40	0.35	0.05	13%
16	0.96	0.94	0.02	2%
17	0.15	0.24	-0.09	-60%
18	0.66	0.48	0.18	27%
19	0.97	0.30	0.67	69%
20	0.20	0.12	0.08	40%
21	0.38	0.35	0.03	8%
			Average drop in IRR by using EMMA	7%
			Standard Deviation	28%

8.2.2 Evaluation 2

The second evaluation looks at the results in more detail. The idea of this evaluation is to find the worst of the six markers - five humans plus EMMA. If the worst marker is EMMA, it may not be an acceptable human substitute marker, depending on how far EMMA falls below the other markers. If, on the other hand, EMMA is not the worst marker, it is probably acceptable.

This evaluation compares the AC1 of 15 pairs of markers. Table 8-2 through Table 8-5 show the AC1 figures in decreasing order; the shaded areas are the five worse pairs. Marker 0 is EMMA; markers 1-5 are the 5 human markers. The bottom parts of the tables show the worst

pairs of markers for each question, i.e., the number of times a particular marker appeared in the 5 worst pairs. The exceptions are for Q17, which was marked by 4 markers and Q19 - 21 which were marked by 3 markers. Table 8-5 reports the results for Q17 and Q19 - 21; the darkest areas indicate that the given marker did not participate in the marking for the relevant question.

The charts can be interpreted by looking at the worst markers. For example, EMMA is clearly the worst of the 6 markers for Q1. It appeared in each of the five worst pairs. In other words, whenever EMMA was in one of the marker pairs, the results were one of the five worst. Q3 shows a different story - Marker 4 is clearly the worst marker as each time he was in one of the marker pairs, the results were one of the five worst.

Table 8-6 shows the overall figures: EMMA was in the worst five marker-pairs 24% of the time, Marker 4 comes in close at 22% of the time. Marker 3 is clearly the best marker as he came in among the worst five pairs only 7% of the time.

Table 8-2 Rater comparisons for Q1-4 and Q8

Question 1			Question 2			Question 3			Question 4			Question 8		
M1	M2	AC1	M1	M2	AC1	M1	M2	AC1	M1	M2	AC1	M1	M2	AC1
3	5	0.69	3	5	0.67	2	3	0.88	2	3	0.89	2	5	0.94
1	2	0.59	1	2	0.69	0	2	0.85	3	4	0.75	2	3	0.94
1	3	0.56	1	3	0.65	0	3	0.85	2	4	0.73	3	5	0.92
4	5	0.56	4	5	0.50	2	5	0.83	0	3	0.68	0	5	0.90
2	4	0.52	2	4	0.63	3	5	0.83	0	4	0.61	0	2	0.87
2	3	0.48	2	3	0.88	0	5	0.77	0	2	0.58	0	3	0.85
1	5	0.48	1	5	0.90	1	5	0.74	2	5	0.19	3	4	0.82
2	5	0.45	2	5	0.75	1	2	0.72	3	5	0.17	2	4	0.80
3	4	0.45	3	4	0.90	1	3	0.72	0	5	0.17	1	3	0.78
1	4	0.43	1	4	0.81	0	1	0.70	1	3	0.13	1	4	0.77
0	5	0.41	0	5	0.73	0	4	0.33	1	5	0.11	4	5	0.76
0	3	0.39	0	3	0.77	1	4	0.27	4	5	0.11	1	5	0.74
0	2	0.37	0	2	0.67	2	4	0.26	1	2	0.10	0	1	0.72
0	1	0.35	0	1	0.88	3	4	0.26	0	1	0.10	1	2	0.71
0	4	0.33	0	4	0.69	4	5	0.25	1	4	0.01	0	4	0.70
ave H		0.52	ave H		0.80	ave H		0.58	ave H		0.32	ave H		0.82
ave E		0.37	ave E		0.63	ave E		0.70	ave E		0.43	ave E		0.81
Worst Marker Count														
Marker		Count	Marker		Count	Marker		Count	Marker		Count	Marker		Count
0		5	0		5	0		1	0		1	0		2
1		1	1		1	1		1	1		4	1		3
2		1	2		1	2		1	2		1	2		1
3		1	3		1	3		1	3		0	3		0
4		1	4		1	4		5	4		2	4		2
5		1	5		1	5		1	5		2	5		2
		10			10			10			10			10

Table 8-3 Rater comparisons for Q9-13

Question 9			Question 10			Question 11			Question 12			Question 13		
M1	M2	AC1	M1	M2	AC1	M1	M2	AC1	M1	M2	AC1	M1	M2	AC1
0	2	0.88	3	5	0.77	4	5	0.76	2	5	0.68	2	3	1.00
3	5	0.80	3	4	0.77	2	4	0.63	2	3	0.66	0	1	0.98
3	4	0.79	4	5	0.60	2	5	0.62	0	5	0.59	0	2	0.98
4	5	0.79	0	2	0.58	3	4	0.60	3	5	0.59	0	3	0.98
1	4	0.72	0	5	0.55	2	3	0.58	0	3	0.57	0	5	0.96
2	3	0.72	2	5	0.53	0	3	0.54	1	3	0.51	1	2	0.96
0	3	0.70	2	3	0.50	3	5	0.53	0	2	0.50	1	3	0.96
2	5	0.70	0	3	0.48	0	5	0.51	1	5	0.43	1	4	0.96
0	5	0.67	2	4	0.43	0	2	0.50	1	2	0.36	0	4	0.94
1	3	0.67	1	4	0.37	0	4	0.49	0	1	0.30	1	5	0.94
1	5	0.67	0	4	0.37	1	5	0.46	1	4	0.27	2	5	0.94
2	4	0.65	1	2	0.30	1	4	0.44	4	5	0.18	3	5	0.94
0	4	0.58	1	3	0.23	1	3	0.42	2	4	0.13	4	5	0.94
1	2	0.48	0	1	0.19	0	1	0.35	3	4	0.11	2	4	0.92
0	1	0.43	1	5	0.16	1	2	0.35	0	4	0.02	3	4	0.92
ave H		0.70	ave H		0.46	ave H		0.54	ave H		0.39	ave H		0.95
ave E		0.65	ave E		0.43	ave E		0.48	ave E		0.40	ave E		0.97
Worst Marker Count														
Marker	Count		Marker	Count		Marker	Count		Marker	Count		Marker	Count	
0	2		0	2		0	1		0	1		0	0	
1	3		1	4		1	5		1	1		1	0	
2	2		2	1		2	1		2	1		2	2	
3	0		3	1		3	1		3	1		3	2	
4	2		4	1		4	1		4	5		4	3	
5	1		5	1		5	1		5	1		5	3	
	10			10			10			10			10	

Table 8-4 Rater comparisons for Q14 - 16 and Q18

Question 14			Question 15			Question 16			Question 18		
M1	M2	AC1	M1	M2	AC1	M1	M2	AC1	M1	M2	AC1
3	4	0.98	1	4	0.95	2	3	1.00	3	4	0.89
1	4	0.98	1	2	0.78	3	5	0.98	2	3	0.74
1	3	0.95	2	4	0.72	2	5	0.98	2	4	0.68
2	4	0.91	0	1	0.48	1	2	0.98	3	5	0.68
4	5	0.91	0	2	0.48	1	3	0.98	1	3	0.63
2	3	0.89	0	4	0.42	0	1	0.957	1	4	0.63
1	2	0.89	0	3	0.39	0	5	0.957	1	5	0.63
3	5	0.88	3	4	0.38	1	5	0.96	2	5	0.63
1	5	0.88	1	3	0.33	3	4	0.96	4	5	0.59
0	3	0.86	2	3	0.31	2	4	0.96	0	3	0.57
2	5	0.86	3	5	0.25	0	2	0.935	1	2	0.55
0	2	0.84	2	5	0.17	0	3	0.935	0	2	0.54
0	4	0.84	4	5	0.13	4	5	0.93	0	4	0.54
0	1	0.82	1	5	0.07	1	4	0.93	0	1	0.44
0	5	0.79	0	5	0.00	0	4	0.891	0	5	0.30
ave H		0.91	ave H		0.40	ave H		0.96	ave H		0.66
ave E		0.83	ave E		0.35	ave E		0.94	ave E		0.48

Worst Marker Count							
Marker	Count	Marker	Count	Marker	Count	Marker	Count
0	4	0	1	0	3	0	4
1	1	1	1	1	1	1	2
2	2	2	1	2	1	2	2
3	0	3	1	3	1	3	0
4	1	4	1	4	3	4	1
5	2	5	5	5	1	5	1
	10		10		10		10

Table 8-5 Rater comparisons for Q17 and Q19 - 21

Question 17			Question 19			Question 20			Question 21		
M1	M2	AC1	M1	M2	AC1	M1	M2	AC1	M1	M2	AC1
0	5	0.38									
0	1	0.36									
1	5	0.36									
3	5	0.23									
1	3	0.21	1	5	0.98	1	4	0.53	4	5	0.45
0	3	0.20	4	5	0.98	0	5	0.44	0	1	0.42
4	5	0.10	1	4	0.95	1	5	0.08	1	4	0.41
0	4	0.00	0	1	0.31	0	1	0.00	0	4	0.36
1	4	0.00	0	5	0.29	0	4	0.00	1	5	0.30
3	4	0.00	0	4	0.28	4	5	0.00	0	5	0.26
ave H		0.15	ave H		0.97	ave H		0.20	ave H		0.38
ave E		0.24	ave E		0.30	ave E		0.12	ave E		0.35
Worst Marker Count											
Marker		Count	Marker		Count	Marker		Count	Marker		Count
0		1	0		3	0		2	0		2
1		1	1		1	1		1	1		1
2			2			2			2		
3		1	3			3			3		
4		4	4		1	4		2	4		1
5		1	5		1	5		1	5		2
		8			6			6			6

Table 8-6 Rater comparisons showing EMMA is the worst marker overall averaged over all 18 questions

Marker ID	Times in Worst Marker Pair	%
0	40	24%
4	37	22%
1	32	19%
5	28	17%
2	18	11%
3	11	7%

8.3 Discussion of the evaluation results

EMMA is the worst marker by each of these evaluations. The first evaluation showed an overall drop of 7%; the second method showed that EMMA was 2% worse than the worst human.

However, the same data can be examined in a different way to provide more positive results. Another way to look at the data in Table 8-2 through Table 8-5 is shown in Table 8-7. It is a per question list of whether EMMA was the worst marker, tied for the worst marker, or not the worst marker. The table shows that EMMA was the worst marker for 6 questions, it tied for worst marker in 2 questions and was not the worst marker for 10 of the questions. Thus, EMMA was not the worst marker for over half of the questions. Therefore, if one examines the data, question by question, EMMA would be acceptable for 56% of the questions.

The two evaluation methods described in Section 8.2 show different results because they are based on different techniques. Evaluation 1 is based on averages; it reports that over the 18 questions, using EMMA results in a 7% drop in IRR. This method may be flawed, however, as Chapter 4 shows that the IRR of human markers varies widely and so the average is somewhat meaningless. Certainly, the average over all of the 18 questions overstates the IRR for some questions and understates it for others. For these reasons, Evaluation 2 may provide more useful results.

Evaluation 2 looks at the individual pairs of markers; using 6 markers yields 15 pairs per question. These 15 pairs are sorted by decreasing IRR. The lowest 5 pairs are the 5 worst markers. The marker who appears most frequently in the lowest 5 pairs is the worst marker. Using this technique, EMMA was not the worst marker for 10 out of the 18 questions. The idea behind this evaluation is that if EMMA is the worst marker, it should probably not be used. However, if it is not the worst, the results would improve by using EMMA leading to the conclusion that EMMA is probably acceptable.

Table 8-7 Questions for which EMMA is the worst / not worst / tied for worst marker

Question ID	Worst	Tied for worst	Not the worst
1	X		
2	X		
3			X
4			X
8			X
9			X
10			X
11			X
12			X
13			X
14	X		
15			X
16		X	
17			X
18	X		
19	X		
20	X		
21		X	
Total	6	2	10

8.4 Implications for CAA consumers

This dissertation provides evidence that the performance of human markers varies widely both over different questions and within a particular question. In addition, EMMA (the particular CAA system developed and studied for this dissertation) performed well for 56% of the questions; implicit in this percentage is that EMMA did not perform well for 44% of the questions. With this variability, it would be easy to give an incorrect picture of the accuracy of the CAA system by emphasizing the questions with better results. These findings suggest that consumers should ask CAA producers the following questions before adopting a particular CAA system.

- How many questions did you test your CAA system on?
- What is the text of the questions?
- How well did your CAA system perform for each question? On average?
- What success metric did you use?
- How well did human markers perform for your questions?
- How many humans did you use?
- How many answers did you mark?
- What kind of training data did you use?
- Does your system need human-marked answers for training data? If so, how many?
- What do I need to do before I can use your system?
- What calibrations do I need to make before I can use your system?
- Does your system need to be recalibrated for each question?

8.5 Future research

8.5.1 The corpus

LSA results depend on both corpus size and corpus content.

8.5.1.1 Corpus size

Existing LSA research stresses the need for a large corpus. For example, Summary Street, an LSA-based instructional software system, uses a general corpus of 11 million words (Wade-Stein & Kintsch, 2003). In contrast, I used a small general corpus of 829,519 words, a factor of ten smaller than that used by the Summary Street developers, who were part of the original team of LSA researchers. I need to identify and acquire a larger corpus for future work. A larger corpus to provide more general training data, according to the literature, is the single most important factor for improving results.

In addition to the general corpus discussed in the previous paragraph, it would be useful to increase the size of the specific corpus, i.e., the previously marked answers. Due to time and funding limitations, I was unable to verify 2/3 of these previously marked answers. Increasing the number of specific documents from 1000 (which were used for this work) to 3000 might result in a more accurate CAA system.

8.5.1.2 Corpus content

An earlier paper (Wiemer-Hastings, Wiemer-Hastings & Graesser, 1999) reports that size is not the only important characteristic of the corpus. Not surprisingly, the composition of the corpus affects the results of grading short answers by LSA. Wiemer-Hastings, Wiemer-Hastings & Graesser claim that two types of training data are necessary: general documents in the form of textbooks and other domain-specific text and specific documents comprising previously human-marked answers. They found the best composition to be about 40% general documents and 60% specific documents.

An ideal specific corpus would provide specific documents that give a spread of marks across the mark range and a variety of answers for each mark. While I believe that I have such a specific corpus, the results of the experiments described in subsections 7.8 (amount of training data) and 7.9 (threshold) suggest that the *quality* of the specific training data may be more important than the *quantity*. Not all of the questions showed improvement by increasing the amount of training data. This result suggests that something in the data itself changed the results. A similar conclusion can be drawn by looking at the threshold experiment. The expected result is that the agreement should rise as the threshold rises; some of the questions showed non-increasing IRR amounts as the threshold increased. Again, this implies that something in the training data itself changed the result. My hypothesis is that some of the training data was mismarked, which is likely since the training data has un-moderated marks and Chapter 4 shows a wide range of agreement among human markers. Further work could involve inspecting the thousands of human-marked answers and "cleaning" them, i.e., removing any answers that can be identified as incorrect.

8.5.2 Corpus pre-processing

Removing stop words and stemming are two types of pre-processing I have used. Stop words (e.g. a, the, my) are considered non-meaningful. Stemming involves conflating word forms to a common string, e.g., *write*, *writing*, *writes*, *written*, *writer* would be represented in the corpus as *writ*.

I plan one more form of pre-processing that has not yet been studied, to my knowledge: using compound nouns as LSA terms. Currently, only single nouns are used. I conjecture that, in the domain of computer science, such terms as *floppy disk* and *hard drive* are ubiquitous and could make a significant difference in my results.

8.5.3 Question analysis

The human IRR results obtained in Chapter 4 as well as the comparison of EMMA and human IRR suggest a wide range of IRR over different types of questions. Two immediate

questions arise: *what are the characteristics of questions that can be marked with high consistency by humans?* and *what are the characteristics of questions that can be marked well by EMMA?* Although I have started this investigation, I have no conclusive results as yet. I suspect that the investigation about characteristics of questions will be more fruitful after I perform the corpus cleaning described in subsection 8.5.1.2.

8.5.4 Increase the number of questions

The generalize-ability of the findings reported in this dissertation would be improved by increasing the number of questions that were investigated. I examined 18 questions in this work. I have raw data for another 22 questions, which I was unable to use because of funding limitations (see subsection 6.1.3 for details).

8.6 Summary of the dissertation

- The following experiments (discussed in detail in Chapter 7) were undertaken to calibrate EMMA:
    - weighting function
    - number of dimensions
    - stop words
    - stemming
    - amount of training data
    - matching threshold
    - number of answers to average
    - proportional weighting
- Two methods (discussed in this chapter) were used to evaluate EMMA. The results can be interpreted in various ways:

- EMMA drops overall agreement by 7% (averaging over all marker pairs and all questions).
- EMMA is worse than the humans for 13 out of 18 questions (72%) (averaging over marker pairs).
- EMMA is 2% worse than the worst human marker averaged over all the questions.
- EMMA is the worst marker for 1/3 of the questions (33%) (looking at each question and each marker pair individually).
- The last point stated positively is that EMMA was not the worst marker for 10/18 questions or 56% of the questions (looking at each question and each marker pair individually).

The apparent contradiction between the second and fourth points can be explained by recalling that the first evaluation used averages and the second evaluation looked at individual marker pairs.

- Future work to improve EMMA:
  - increase corpus size
  - clean specific training data, i.e., remove erroneously marked answers
  - use compound nouns
- Future work to understand human and computer IRR:
  - investigate the questions marked for this dissertation to discover the characteristics that make a question easier or harder to mark with high accuracy
  - use a larger number of questions to increase generalize-ability

8.7 Advice for future researchers

I believe that this dissertation sets the stage for future work that can produce results where EMMA is good enough to use as a second marker. Based on my experience creating EMMA, I offer the following advice to anyone contemplating future research in this area.

- Don't attempt to conduct LSA research without an adequate computer. The system described in Chapter 6 should be considered a minimum. LSA is both computationally and memory expensive. If your computer does not have enough RAM (16 G is what this work used), you won't be able to increase the corpus size.
- It might be worth the time to build an attractive interface for demos of EMMA. At present, the system can mark only the answers that are all ready in the database. It might be useful during a demo to allow an observer to enter an answer to a question and see, in real time, the mark calculated by EMMA.
- Some thought needs to be given towards how EMMA would actually be used in a production environment. This would necessitate a different front end for training. At present, EMMA is strictly a research prototype with none of the HCI niceties needed for a system with users other than the developer.
- Do not forget the idea of using EMMA as a second marker. In that scenario, a human mark and EMMA's mark would be compared. Only if they disagree by some percentage would a second human need to moderate the mark.
- Be very organized from the beginning. Keeping track of the masses of data generated by LSA research is necessary for legitimate results.
- Consider de-normalizing the database. At present, it is normalized to 3rd normal form, the standard practice when designing a database. The advantage of a normalized database is that it is easier to update the tables. The disadvantage is that the queries to examine the database are more difficult. Experience developing and testing EMMA has shown that the data is seldom updated. After the one-time task of

populating the database, the only table that is updated is the Marks table when EMMA marks an answer. A researcher will spend much more time examining the data in various database tables than she will spend writing the code to update the Marks table.

- Ensure that you have a good grasp of Excel, Java, MYSQL, and SPSS. You will use them extensively.
- Read and digest the papers in the taxonomy as well as the books by Gwet. You need to understand them.

8.8 Accomplishments

The following list shows the research questions from subsection 1.3 along with a description of how these questions were answered.

- What does the literature say about LSA?

provided a research taxonomy that identified gaps in the literature

- To what extent can LSA be used to assess short answers in the domain of computer science (CS)?

provided evidence that suggests that LSA can be used to mark 56% of the questions under consideration

- How can LSA results be reported to other researchers?

recommended a framework that could be used by both the LSA community and the CAA community for uniform, comprehensive reporting of research results

- What questions should be asked by those interested in adopting a CAA?

provided a comprehensive list of questions

- How hard is it to build an LSA based CAA?

showed that it can be done although it wasn't easy; it took almost 3 years of full-time work to understand LSA, create the database of answers to be marked, write and debug the Java program, and evaluate the results

- How inaccurate are human markers?

provided evidence supporting the widely held perception of a lack of agreement among human markers

- How accurate, compared to human markers, is LSA when used to assess short answers in the CS domain? How do you measure accuracy?

highlighted problems with existing metrics, located a new IRR metric, Gwet's AC1, and used it to compare humans and my CAA for 18 different questions.

- Given a suitable metric, how can you evaluate an LSA based CAA?

established a method of evaluating any CAA system illustrating this method throughout the experiments to calibrate EMMA

- What calibrations need to be made to an LSA-based marking system to assess short answers in the CS domain?
    - Corpus related questions
        - On what corpus should the LSA system be trained?

          provided results by using a general corpus of text books and a specific corpus of previously marked human answers
        - What is a good size for the corpus?

          provided results using a general corpus of just under a million words and varying sizes for the specific corpus
    - Pre-processing questions
        - Does it help to remove stop words?

yes, and quantified how much

- Does stemming help?

yes, and quantified how much

- Will using compound nouns improve the performance of LSA?

future work

- What number in the dimension reduction step gives the best results?

showed that it varies by question and quantified how much using the best dimension improved results

- Which weighting function gives optimum results - log-entropy or tfidf?

provided inconclusive results but used tfidf because it was slightly better

- Does proportional averaging improve the results?

provided inconclusive results but used proportional averaging for the slight improvement

- Does varying the amount of training data improve results?

demonstrated results that showed an improvement

8.9 Conclusion

This dissertation reports the results of EMMA, a Latent Semantic Analysis (LSA) based Computer Assisted Assessment (CAA) system. The work for the dissertation is complete, my research is not. Although I did not achieve my goal of creating a CAA that was conclusively good enough for all 18 questions under investigation in the course of the dissertation work, I believe that I can improve my results by following the roadmap laid out in the previous subsections.

References

Aberson, C. (2002). *Interpreting Null Results: Improving Presentation and Conclusions with Confidence Intervals*. **Journal of Articles in Support of the Null Hypothesis 1**(3): 36-42.

Alt (2003). *The Association for Learning Technology's response to the Future Of Higher Education White Paper.* www.alt.ac.uk/docs/he_wp_20030429_final.doc, 2003.

Anderson, Michael & McCartney, Robert (2003). *Diagram processing: Computing with Diagrams*. **Artificial Intelligence 145**(1-2): 181-226.

Benford, S. D., Burke, E. K., Foxley, E. & Higgins, C. A. (1996). *Ceilidh: A Courseware System for the Assessment and Administration of Computer Programming Courses in Higher Education.* Nottingham, UK, The University of Nottingham, http://cs.joensuu.fi/~mtuki/www_clce.270296/Burke.html, last accessed 24 October 2007.

Berglund, Anders (1999). *Changing Study Habits - a Study of the Effects of Non-traditional Assessment Methods. Work-in-Progress Report*. **6th Improving Student Learning Symposium**, Brighton, UK.

Berry, Michael W., Do, Theresa, O'Brien, G. W. & Krisna, Mijay (1993). *SVDPACKC (Version 1.0) User's Guide.* Technical Report UT-CS-93-194. Memphis, Tennessee, University of Tennessee.

Berry, Michael W., Dumais, S. T. & O'Brien, G. W. (1995). *Using linear algebra for intelligent information retrieval*. **SIAM Review 37 4**: 573-595.

Blood, Emily & Spratt, Kevin F. (2007). *Disagreement on Agreement: Two Alternative Agreement Coefficients*. **SAS Global Forum 2007**.

Bloom, Benjamin , Engelhart, M., Furst, E., Hill, W. & Krathwohl, D. (1956). **Taxonomy of educational objectives : the classification of educational goals. Handbook 1, Cognitive domain**. London, Longmans, Green.

Brown, G., Bull, J. & Pendlebury, M. (1997). **Assessing student learning in higher education**. London, Routledge.

Bull, Joanna (1999). *Update on the National TLTP3 Project 'The Implementation and Evaluation of Computer-assisted Assessment'*. **Proceedings of the 3rd International CAA Conference**, Loughborough, UK.

Bull, Joanna (2000). *Annual Report 2000 The Implementation and Evaluation of Computer-assisted Assessment.* TLTP85/AR2/2000. Loughborough, UK, University of Luton, http://www.caacentre.ac.uk/dldocs/Annual_Report_2K.pdf.

Bull, Joanna, Conole, Grainne, Davis, H. C., White, Su, Danson, Myles & Sclater, Niall (2002). *Rethinking Assessment through Learning Technologies*. **Proceedings of ASCILITE 2002**, Auckland, New Zealand.

Burgess, Curt, Livesay, Kay & Lund, Kevin (1998). *Explorations in context space: Words, sentences, discourse*. **Discourse Processes 25**: 211-257.

Burstein, Jill, Chodorow, Martin & Leacock, Claudia (2003). *Criterion Online Essay Evaluation: An Application for Automated Evaluation of Student Essays*. **Proc. of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence**, Acapulco, Mexico.

Carneson, John, Delpierre, Georges & Masters, Ken (1996). *Designing and Managing Multiple Choice Questions*. Cape Town, South Africa, University of Cape Town, http://web.uct.ac.za/projects/cbe/mcqman/mcqman01.html, last accessed 1 November 2007.

Carter, Janet, Ala-Mutka, Kirsti, Fuller, Ursula, Dick, Martin, English, John, Fone, William & Sheard, Judy (2003). *How Shall We Assess This?* **Proceedings of the ITiCSE 2003 working group reports**, Thessaloniki, Greece, ACM Press.

Chan, Y. H. (2003). *Biostatistics 104: Correlational Analysis*. **Singapore Medical Journal 44**(12): 614-619.

Christie, James R. (2003). *Automated Essay Marking*. **Proceedings of the 4th Annual LTSN-ICS Conference**, NUI Galway, LTSN Centre for Information and Computer Sciences.

Cohen, J. (1960). *A coefficient of agreement for nominal scales*. **Educational and Psychological Measurement 20**: 37-46.

Conole, Grainne & Bull, Joanna (2002). *Pebbles in the Pond: Evaluation of the CAA Centre*. **Proceedings of the 6th International CAA Conference**, Loughborough, UK.

Conole, Grainne & Warburton, Bill (2005). *A review of computer-assisted assessment*. **ALT-J, Research in Learning Technology 13**(1): 17-31.

Croft, A., Danson, M., Dawson, B. R. & Ward, J. P. (2001). *Experience of using computer assisted assessment in engineering mathematics*. **Computers and Education 37**(1): 53-66.

Dancey, Christine P. & Reidy, John (2002). **Statistics Without Maths for Psychology: Using SPSS for Windows**. Essex, England, Prentice Hall.

Daniel, Wayne W. (1977). **Introductory Statistics with Applications**. Boston, Houghton Mifflin Company.

Daniels, Mats, Berglund, Anders, Pears, Arnold & Fincher, Sally (2004). *Five Myths of Assessment*. **6th Australasian Computing Education Conference (ACE2004)**, Dunedin, New Zealand.

Davies, Phil (2001). *CAA must be more than multiple-choice tests for it to be academically credible?* **Proceedings of the 5th International CAA Conference**, Loughborough, UK.

Davies, Phil (2002). *"There's No Confidence in Multiple-choice Testing, ..."* **Proceedings of the 6th International CAA Conference**, Loughborough, UK.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, Thomas K. & Harshman, R. (1990). *Indexing by Latent Semantic Analysis*. **Journal of the American Society for Information Science 41**(6): 391-407.

Dennis, Simon, Landauer, Thomas K., Kintsch, Walter & Quesada, Jose (2003). *Latent Semantic Analysis: Theory, Use and Applications.* Tutorial given at CogSci2003, http://lsa.colorado.edu/~quesadaj/tutorial - last accessed 25/11/2007.

DiEugenio, Barbara & Glass, Michael (2004). *The kappa statistic: a second look.* **Computational Linguistics 30**(1).

Duke-Williams, Emma & King, Terry (2001). *Using Computer-Aided Assessment to Test Higher Level Learning Outcomes*. **Proceedings of the 5th International CAA Conference**, Loughborough, UK.

Dumais, S. T. (1991). *Improving the retrieval of information from external sources*. **Behavioral Research Methods, Instruments & Computers 23**(2): 229-236.

Dumais, Susan. T. (2003). *Data-driven approaches to information access*. **Cognitive Science 27**: 491-524.

Dumais, Susan. T. (2007). LSA and Information Retrieval: Getting Back to Basics. **Handbook of Latent Semantic Analysis**. Landauer, T. K., McNamara, D. S., Dennis, S. and Kintsch, W. Eds, New Jersey, US, Lawrence Erlbaum Associates**:** 293 - 321.

Farthing, Dave W. & McPhee, Duncan (1999). *Multiple choice for honours-level students? A statistical evaluation*. **Proceedings of the 3rd International CAA Conference**, Loughborough, UK.

Feinstein, Alvan R. & Cicchetti, Domenic V. (1990). *High agreement but low kappa: I. The problems of two paradoxes*. **Journal of Clinical Epidemiology 43**(6): 543-549.

Field, Andy (2005). **Discovering Statistics Using SPSS, Second Edition**. London, Sage Publications Ltd.

Foltz, Peter W. (1990). *Using latent semantic indexing for information filtering*. **Conference on Office Information Systems**, Cambridge, MA.

Foltz, Peter W. (1996). *Latent semantic analysis for text-based research*. **Behavior Research Methods, Instruments and Computers 28**(2): 197-202.

Foltz, Peter W., Gilliam, Sara & Kendall, Scott A. (2000). *Supporting content-based feedback in online writing evaluation with LSA*. **Interactive Learning Environments 8**(2): 111-129.

Foltz, Peter W., Kintsch, Walter & Landauer, Thomas K. (1998). *The Measurement of Textural Coherence with Latent Semantic Analysis*. **Discourse Process 25**(2&3): 285-307.

Foltz, Peter W., Laham, D. & Landauer, Thomas K. (1999). *Automated Essay Scoring: Applications to Educational Technology*. **Proceedings of ED-MEDIA '99**, Seattle.

Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, Thomas K., Harshman, Richard A., Streeter, Lynn A. & Lochbaum, Karen E. (1988). *Information retrieval using a singular value decomposition model of latent semantic structure*. **Proceedings of 11th annual international ACM SIGIR conference on research and development in information retrieval**, ACM.

Furnas, G. W., Gomez, L. M., Landauer, Thomas K. & Dumais, S. T. (1982). *Statistical semantics: How can a computer use what people name things to guess what things people mean when they name things?* **Proceedings of the SIGCHI Conference on Human Factors in Computing Systems**, ACM.

Gerald & Wheatley (1970). **Applied Numerical Analysis**. Addison-Wesley.

Graesser, Arthur C., Wiemer-Hastings, Peter, Wiemer-Hastings, Katja, Harter, Derek & The Tutoring Research Group (2000). *Using latent semantic analysis to evaluate the contributions of students in AutoTutor*. **Interactive Learning Environments. [Special Issue, J. Psotka, guest editor] 8**(2): 129-147.

Gwet, Kilem (2001a). **Handbook of Inter-Rater Reliability: How to Estimate the Level of Agreement Between Two or Multiple Raters**. Gaithersburg, MD, STATAXIS Publishing Company.

Gwet, Kilem (2001b). **Statistical Tables for Inter-Rater Agreement; Tables Providing Exact t Critical Values for Testing Statistical Significance**. Gaithersburg, MD, STATAXIS Publishing Company.

Gwet, Kilem (2002a). *Kappa Statistic is not Satisfactory for Assessing the Extent of Agreement Between Raters*. **Statistical Methods for Inter-Rater Reliability Assessment 1**.

Gwet, Kilem (2002b). *Inter-Rater Reliability: Dependency on Trait Prevalence and Marginal Homogeneity*. **Statistical Methods for Inter-Rater Reliability Assessment 2**.

Haley, Debra, Thomas, Pete, De Roeck, Anne & Petre, Marian (2007). *Measuring Improvement in Latent Semantic Analysis-Based Marking Systems: Using a Computer to Mark Questions about HTML*. **Proceedings of the Ninth Australasian Computing Education Conference (ACE2007)**, Ballarat, Victoria, Australia, Australian Computer Society Inc.

Haley, Debra Trusso, Thomas, Pete, De Roeck, Anne & Petre, Marian (2005). *A Research Taxonomy for Latent Semantic Analysis-Based Educational Applications*. **International Conference on Recent Advances in Natural Language Processing'05**, Borovets, Bulgaria.

Hopkins, Kenneth (1998). **Educational and Psychological Measurement and Evaluation 8th Edition - Cased**. Allyn & Bacon.

Hu, Xiangen, Cai, Zhiqiang, Wiemer-Hastings, Peter, Graesser, Art C. & McNamara, Danielle S. (2007). Strengths, Limitations, and Extensions of LSA. **Handbook of Latent Semantic Analysis**. Landauer, T. K., McNamara, D. S., Dennis, S. and Kintsch, W. Eds, New Jersey, US, Lawrence Erlbaum Associates**:** 401-425.

Huo, Ming, Zhang, He & Jeffrey, Ross (2006). *An Exploratory Study of Process Enactment as Input to Software Process Improvement*. **WoSQ'06**, Shanghai, China.

Joy, Mike & Luck, Michael (1998). *Effective Electronic Marking for On-line Assessment*. **Proceedings of ITiCSE'98**, Dublin, Ireland.

Kintsch, Walter & Bowles, Anita R. (2002). *Metaphor comprehension: What makes a metaphor difficult to understand?* **Metaphor and Symbol 17**: 249-262.

Landauer, Thomas K. (2007). LSA as a Theory of Meaning. **Handbook of Latent Semantic Analysis**. Landauer, T. K., McNamara, D. S., Dennis, S. and Kintsch, W. Eds, New Jersey, US, Lawrence Erlbaum Associates**:** 3-35.

Landauer, Thomas K. & Dumais, S. T. (1997). *A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge*. **Psychological Review 104**(2): 211-240.

Landauer, Thomas K., Foltz, Peter W. & Laham, D. (1998). *An introduction to Latent Semantic Analysis*. **Discourse Processes 25**: 259-284.

Landauer, Thomas K., Laham, Darrel & Foltz, Peter W. (2003). *Automatic essay assessment*. **Assessment in Education: Principles, Policy & Practice 10**(3): 295-308.

Landauer, Thomas K., Laham, Darrell, Rehder, Bob & Schreiner, M. E. (1997). *How Well Can Passage Meaning be Derived without Using Word Order? A Comparison of Latent Semantic Analysis and Humans*. **Proceedings of the 19th Annual Meeting of the Cognitive Science Society**.

Lemaire, Benoit & Bianco, Maryse (2003). *Contextual effects on metaphor comprehension: Experiment and simulation*. **Proceedings of the 5th Int'l Conference on Cognitive Modeling (ICCM'2003)**, Bamberg, Germany.

Lemaire, Benoit & Dessus, Philippe (2001). *A system to assess the semantic content of student essays*. **Journal of Educational Computing Research 24**(3): 305-320.

Manning, C. D. & Schütze, H. (1999). **Foundations of Statistical Natural Language Processing**. Cambridge, Massachusetts, MIT Press.

Marcus, Andrian & Maletic, Jonathan I. (2003). *Recovering documentation-to-source-code traceability links using latent semantic indexing*. **Proceedings of the 25th International Conference on Software Engineering (ICSE'2003)**, Portland, Oregon, IEEE.

Marcus, Andrian, Sergeyev, A., Rajlich, V. & Maletic, Jonathan I. (2004). *An Information Retrieval Approach to Concept Location in Source Code*. **Proceedings of the 11th IEEE Working Conference on Reverse Engineering**, Delft, The Netherlands.

Martin, Dian I. & Berry, Michael W. (2007). Mathematical Foundations Behind Latent Semantic Analysis. **Handbook of Latent Semantic Analysis**. Landauer, T. K., McNamara, D. S., Dennis, S. and Kintsch, W. Eds, New Jersey, US, Lawrence Erlbaum**:** 35-55.

Mason, Oliver & Grove-Stephensen, Ian (2002). *Automated free text marking with paperless school*. **Proceedings of the 6th International CAA Conference**, Loughborough, UK.

McAlpine, Mhairi (2002). *Principles of Assessment.* CAA Centre, University of Luton, www.caacentre.ac.uk/resources/bluepapers/index.shtml, last accessed 28 October 2007.

McKenna, Colleen (2001). *Academic Approaches and Attitudes Towards CAA: A Qualitative Study*. **Proceedings of the 5th International CAA Conference**, Loughborough.

Miller, Tristan (2003). *Essay assessment with Latent Semantic Analysis.* **Journal of Educational Computing Research 28**.

Mitchell, Tom, Aldridge, Nicola & Broomhead, Peter (2003). *Computerised Marking of Short-Answer Free-Text Responses*. **Proceedings of IAEA**, Manchester, UK.

Mulligan, Brian (1999). *Pilot Study on the Impact of Frequent Computerized Assessment on Student Work Rates*. **Proceedings of the 3rd International CAA Conference**, Loughborough, UK.

Nakov, Preslav (2000). Latent Semantic Analysis of Textual Data. **Proceedings of the International Conference on Computer Systems and Technologies**. Eds, Sofia, Bulgaria.

Nakov, Preslav, Valchanova, Elena & Angelova, Galia (2003). *Towards Deeper Understanding of the LSA Performance*. **Proceedings of Recent Advances in Natural Language Processing '03**, Borovets, Bulgaria.

NCFOT, The National Center for Fair & Open Testing (1998). *Multiple Choice Tests.* Cambridge, Massachusetts, http://www.fairtest.org/facts/mctfcat.html, last accessed 1 November 2007.

Newstead, Stephen (2002). *Examining the examiners: Why are we so bad at assessing students?* **Psychology Learning and Teaching 2**(2): 70-75.

Newstead, Stephen & Dennis, I. (1994). *Examiners examined: The reliability of exam marking in psychology*. **The Psychologist 7**: 216-219.

Onix (2008). *Text Retrieval Toolkit*.

Osborne, C. & Winkley, J. (2006). *Developments in On-Screen Assessment Design for Examinations*. **Proceedings of the 10th International CAA Conference**, Loughborough, UK.

Perez, Diana, Gliozzo, Alfio, Strapparava, Carlo, Alfonseca, Enrique, Rodriquez, Pilar & Magnini, Bernardo (2005). *Automatic Assessment of Students' free-text Answers underpinned by the combination of a Bleu-inspired algorithm and LSA*. **Proceedings of the 18th International FLAIRS Conference**, Clearwater Beach, Florida.

Porter, Martin (2006). *The Porter Stemming Algorithm.*

Preston, Jon Anderson & Shackelford, Russell (1999). *Improving On-line Assessment: an Investigation of Existing Marking Methodologies*. **Proceedings of ITiCSE'99**, Cracow, Poland.

Price, Blaine A., Baecker, Ronald M. & Small, Ian S. (1993). *A Principled Taxonomy of Software Visualization*. **Journal of Visual Languages and Computing 4**(3): 211-266.

Purpura, Stephen & Hillard, Dustin (2006). *Automated Classification of Congressional Legislation*. **The 7th Annual International Conference of Digital Government Research '06**, San Diego, CA, ACM.

Quesada, Jose, Kintsch, Walter & Gomez, Emilio (2001). *A computational theory of complex problem solving using the vector space model (part 1): Latent Semantic Analysis, through the path of thousands of ants*. **Cognitive Research with Microworlds 43-84**: 117-131.

Race, Phil (1995). *The Art of Assessing*. **New Academic 5**(3).

Ricketts, Chris, Filmore, Paul, Lowry, Roy & Wilks, Sally (2003). *How Should We Measure the Costs of Computer Aided Assessment?* **Proceedings of the 7th International CAA Conference**, Loughborough, UK.

Ricketts, Chris & Wilks, Sally (2002a). *Improving Student Performance Through Computer-based Assessment: insights from recent research*. **Assessment & Evaluation in Higher Education 27**(5): 475-479.

Ricketts, Chris & Wilks, Sally (2002b). *What Factors affect Student Opinions of Computer-Assisted Assessment*. **Proceedings of the 6th CAA Conference**, Loughborough, UK.

Rowntree, Derek (2004). **Statistics Without Tears: A Primer for Non-Mathematicians**. Boston, Pearson Education, Inc.

Sabar, Naama (2002). *Towards principle practice in evaluation: learning from instructors' dilemmas in evaluating graduate students*. **Studies in Educational Evaluation 28**(4): 329-345.

Salton, G., Wong, A. & Yang, C. S. (1975). *A Vector Space Model for Automatic Indexing*. **Communications of the ACM 18**(11): 613-620.

Sclater, Niall & Howie, Karen (2003). *User requirements of the 'ultimate' online assessment engine*. **Computers and Education 40**(3): 285-306.

Sclater, Niall & MacDonald, Mary (2004). *Developing a National Item Bank*. **Proceedings of the 8th International CAA Conference**, Loughborough, UK.

Sebastiani, Fabrizio (2002). *Machine Learning in Automated Text Categorization*. **ACM Computing Surveys 34**(1): 1-47.

Sim, Gavin & Holifield, Phil (2004). *Piloting CAA: All Aboard*. **Proceedings of the 8th International CAA Conference**, Loughborough, UK.

Sim, Gavin, Holifield, Phil & Brown, Martin (2004). *Implementation of computer assisted assessment: lessons from the literature*. **ALT-J, Research in Learning Technology 12**(3).

Srihari, Sargur, Collins, Jim, Srihari, Rohini, Srinivasan, Harish, Shetty, Shravya & Brutt-Griffler, Janina (2008). *Automatic scoring of short handwritten essays in reading comprehension tests*. **Artificial Intelligence 172**: 300-324.

Stegmann, Jens & Lucking, Andy (2005). *Assessing Reliability on Annotations (1): Theoretical Considerations*, University of Beilefeld.

Steinhart, David J. 2001. *Summary Street: An intelligent tutoring system for improving student writing through the use of Latent Semantic Analysis*. PhD thesis. Department of Psychology, University of Colorado, Boulder, Colorado.

Stephens, D. (1994). *Using computer-assisted assessment: time saver or sophisticated distractor?* **Active Learner 1**: 445 - 464.

Stephens, Derek & Mascia, Janine (1997). *Results of a (1995) Survey into the use of Computer-Assisted Assessment in Institutions of Higher Education in the UK*. Loughborough, UK, Loughborough University, (formerly found at:)http://www.lboro.ac.uk/service/ltd/flicaa/downloads/survey.pdf.

Summons, Peter, Coldwell, Jo, Henskens, Frans & Bruff, Christine (1997). *Automating Assessment and Marking of Spreadsheet Concepts*. **Proceedings of the 2nd Australian Conference on Computer Science Education, SIGCSE**, Melbourne, Australia, ACM.

Thomas, Pete, Haley, Debra, De Roeck, Anne & Petre, Marian (2004). *E-Assessment using Latent Semantic Analysis in the Computer Science Domain: A Pilot Study*. **Proceedings of the eLearning for Computational Linguistics and Computational Linguistics for eLearning Workshop at COLING 2004.**, Geneva.

Thomas, Pete, Smith, Neil & Waugh, Kevin (2008). *Automatic Assessment of Sequence Diagrams*. **Proceedings of the 12th International Conference on Computer Assisted Assessment**, Loughborough, UK.

Thomas, Pete, Waugh, Kevin & Smith, Neil (2005). *Experiments in the automatic marking of ER-Diagrams*. **Proceedings of ITiCSE 05**, Lisbon, Portugal, ACM.

Thompson, Bruce (2002). *What Future Quantitative Social Science Research Could Look Like: Confidence Intervals for Effect Sizes*. **Educational Researcher 31**(3): 25-32.

Tsintsifas, Athanasios. 2002. *A Framework for the Computer Based Assessment of Diagram Based Coursework*. unpublished PhD thesis. School of Computer Science and Information Technology, University of Nottingham, Nottingham. 235 pp.

Venables, Anne & Haywood, Liz (2003). *Programming students NEED instant feedback!* **Proceedings of the 5th Australasian conference on Computing education, ACE03**.

Wade-Stein, Dave & Kintsch, Eileen (2003). *Summary Street: Interactive computer support for writing.* Technical Report from the Institute for Cognitive Science. University of Colorado, USA.

Warburton, Bill & Conole, Grainne (2003a). *CAA in UK HEIs: the state of the art.* **Proceedings of the 7th International CAA Conference**, Loughborough, UK.

Warburton, Bill & Conole, Grainne (2003b). *Key Findings from recent literature on Computer-aided Assessment*. **Proceedings of ALT-C 2003**, Sheffield, UK.

Whittington, Dave & Hunt, Helen (1999). *Approaches to the Computerized Assessment of Free Text Responses*. **Proceedings of the 3rd Annual CAA Conference**, Loughborough, UK.

Wiemer-Hastings, Peter (1999). *How Latent is Latent Semantic Analysis*. **Proceedings of the 16th International Joint Conference on Artificial Intelligence**, Stockholm, Sweden.

Wiemer-Hastings, Peter, Graesser, Arthur & Harter, D. (1998). *The foundations and architecture of Autotutor*. **Proceedings of the 4th International Conference on Intelligent Tutoring Systems**, San Antonio, Texas.

Wiemer-Hastings, Peter, Wiemer-Hastings, Katja & Graesser, Arthur C. (1999). Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. **Artificial Intelligence in Education**. Lajoie, S. P. and Vivet, M. Eds, Amsterdam, IOS Press.

Woit, Denise & Mason, David (2003). *Effectiveness of online assessment*. **Proceedings of SIGCSE'03**, Reno, Nevada, ACM.

Wood, Joseph & Burrow, Michael (2002). *Formative assessment in Engineering using "TRIADS" software*. **Proceedings of the 6th International CAA Conference**, Loughborough, UK.

Yang, Hui, Callan, Jamie & Shulman, Stuart (2006). *Next Steps in Near-Duplicate Detection for eRulemaking*. **Proceedings of the Sixth National Conference on Digital Government Research**.

Appendix A The Latent Semantic Analysis Research Taxonomy

Appendix A
The Latent Semantic Analysis Research Taxonomy

System Name	Refer-ence	Who	Where	What / Why	Stage of Development/ Type of work	Purpose	Innovation	Major Result / Key points
indexing not assessing essays	DDF90	Deerwester, Dumais, Furnas, Landauer, Harshman	U of Chicago, Bellcore, U of W. Ontario	explain new theory that overcomes the deficiencies of term-matching	LSI research	information retrieval	LSI: explains SVD and dimension reduction steps	for Med: for all but the two lowest levels of recall, precision of the LSI method lies well above that obtained with straight-forward term matching; no difference for CISI
	Dum91	Dumais	Bellcore	attempt better LSI results	LSI research	information retrieval	compared different weighting functions	log entropy best weighting function; stemming and phrases showed only 1-5% improvement; 40% better than raw frequency weighting
	BD095	Berry, Dumais, O'Brien	U of Tenn, Bellcore	explain new theory	LSI research	information retrieval	LSI	LSI - completely automatic indexing method using SVD, shows how to do SVD updating of new terms
	FBP94	Foltz, Britt, Perfetti	New Mexico State University, Slippery Rock U, U of Pittsburgh	matching summaries to text read, determine if LSA can work as well as coding propositions	LSA research	text comprehension to evaluate a reader's situation model	matching summaries to text read, analyses knowledge structures of subjects and compares them to those generated by LSA	representation generated by LSA is sufficiently similar to the readers' situation model to be able to characterize the quality of their essays
	FKL98	Foltz, Kintsch, Landauer		measure text coherence	LSA research		using LSA to measure text coherence	LSA needs a corpus of at least 200 documents; online encyclopedia articles can be added
	LD97	Landauer, Dumais	U of Colorado, BellCore	explain new theory	LSA research			LSA could be a model of human knowledge acquisition
	LLR97	Landauer, Laham, Rehder, Schreiner	U of Colorado	compared essays scores given by readers and LSA, to determine importance of word order	LSA theory	grading essays	investigating the importance of word order; combined quality (cosine) and quantity (vector length)	LSA predicted scores as well as human graders; separating tech and non-technical words made no improvement
	RSW98	Rehder, Shreiner, Wolfe, Laham, Landauer, Kintsch	U of Colorado	explore certain technical issues	LSA research	grading essays	investigated technical vocabulary, essay length, optimal measure of semantic relatedness, and directionality of knowledge in the high dimensional	nothing to be gained by separating essay into tech and non tech terms cosine and length of essay vector are best predictors of mark
	WSR98	Wolfe,Shreiner, Rehder, Laham, Foltz, Kintsch, Landauer	U of Colorado, New Mexico State Univ	compared essay scores after reading one of 4 texts	LSA research	select appropriate text	using LSA to select appropriate text	LSA can measure prior knowledge to select appropriate texts
Intelligent Essay Assessor (IEA) http://psych.nmsu.edu.essay	FLL99	Foltz, Landauer, Laham	New Mexico State University, Knowledge Analysis Technologies, U of Colorado	reports on various studies using LSA for automated essay scoring	deployed application for formative assessment	practice essay writing		Over many diverse topics, the IEA scores agreed with human experts as accurately as expert scores agreed with each other.

Appendix A
The Latent Semantic Analysis Research Taxonomy Page 2 of 9

System Name	Refer-ence	Who	Where	What / Why	Stage of Development/ Type of work	Purpose	Innovation	Major Result / Key points
Summary Street	KSS00	Kintsch, Steinhart, Stahl, LSA Research Group, Matthews, Lamb	U of Colorado, Platt Middle School http://www.k-a-t.com /cu.shtml	helps students summarize essays to improve reading comprehension skills	deployed application for formative assessment	provide feedback on length, topics covered, redundancy, relevance	graphical interface, optimal sequencing of feedback	students produced better summaries and spent more time on task with Summary Street
Summary Street	Ste01	Steinhart	U of Colorado http://www.k-a-t.com /cu.shtml	helps students summarize essays to improve reading comprehension skills	deployed application for formative assessment	provide feedback on length, topics covered, redundancy, relevance	graphical interface, optimal sequencing of feedback	the more difficult the text, the better was the result of using Summary Street, feedback doubled time on task
	Lan02b	Landauer	U of Colorado	explaining LSA		LSA general research		LSA works by solving a system of simultaneous equations
AutoTutor	WWG99	Wiemer-Hastings, P., Wiemer-Hastings, K, Graesser, A.	U of Memphis	test theory that LSA can facilitate more natural tutorial dialogue in an intelligent tutoring system (ITS)	deployed application for formative assessment	assess short answers given to Intelligent Tutoring System	tested size and composition of corpus for best LSA results	LSA works best when specific texts comprise at least 1/2 of the corpus and the rest is subject related; works best on essays > 200 words
	Wie00	Wiemer-Hastings	U of Memphis	determine effectiveness of adding syntactic info to LSA	LSA research	assess short answers given to ITS	added syntactic info to LSA	adding syntax decreased the effectiveness of LSA - as compared to Wie99 study
Select-a-Kibitzer	WG00	Wiemer-Hastings, Graesser	U of Memphis	give meaningful feed back on essays using agents	deployed application for formative assessment	assess short answers given to ITS	investigated types of corpora for best results	best corpus is specific enough to allow subtle semantic distinctions within the domain, but general enough that moderate variations in terminology won't be lost
SLSA - Structured LSA	WZ01	Wiemer-Hastings, Zipitria	U of Edinburgh	evaluate student answers for use in ITS	LSA research	assess short answers given to ITS	combines rule-based syntactic processing with LSA - adds part of speech	adding structure-derived information improves performance of LSA; LSA does worse on texts < 200 words
	Nak00b	Nakov	Sofia University	explore uses of LSA in textual research	LSA research		uses correlation matrix to display results; analysis of C programs	
	NPM01	Nakov, Popova, Mateev	Sofia University	evaluate weighting function for text categorisation	LSA research	analyse English literature texts	compared 2 local weighting times 6 global weighting methods	log entropy works better than classical entropy

Appendix A
The Latent Semantic Analysis Research Taxonomy Page 3 of 9

System Name	Refer-ence	Who	Where	What / Why	Stage of Development/ Type of work	Purpose	Innovation	Major Result / Key points
	FKM01	Franceschetti, Karnavat, Marineau, et al	U of Memphis	constructing different types of physics corpora to evaluate best type for an ITS	LSA research for formative assessment	intelligent tutoring	used 5 different corpora to compare vector lengths of words	carefully constructed smaller corpus may provide more accurate representation of fundamental physical concepts than much larger one
	OFK02	Olde, Franceschetti, Karnavat, et al	U of Memphis, CHI Systems	evaluate corpora with different specificities for use in ITS	LSA research for formative assessment	intelligent tutoring	used 5 different corpora to compare essay grades	sanitizing the corpus provides no advantage
Apex	LD01	Lemaire, Dessus	U of Grenoble-II	web-based learning system, automatic marking with feedback	deployed application for formative assessment	provide feedback on topic, outline and coherence		LSA is a promising method to grade essays
	QKG01a	Quesada, Kintsch, Gomez	U of Colorado, U of Grenada	investigate complex problem solving using LSA	CPS and LSA research		represent actions taken in a Microworld as tuples for LSA	LSA is a promising tool for representing actions in Microworlds.
Distributed LSI	BB03	Bassu, Behrens	Telcordia	improve LSI by addressing scalability problem	LSI research	information retrieval	subdivide corpus into several homogeneous subcollections	a divide-and-conquer approach to IR not only tackles its scalability problems but actually increases the quality of returned documents
SELSA	KKP03	Kanejiya, Kumar, Prasad	Indian Institute of Technology	evaluate student answers in an ITS	LSA research	intelligent tutoring	augment each word with POS tag of preceding word, used 2 unusual measures for evaluation: MAD and Correct vs False evaluation	SELSA has limited improvement over LSA
Indexing not assessing essays	NVA03	Nakov, Valchanova, Angelova	U of Cal, Berkeley, Bulgarian Academy of Sciences	investigating the most effective meaning of "word"	LSA research	text categorisation	compared various methods of term weighting with NLP pre-processing	linguistic pre-processing (stemming, POS annotation, etc) does not substantially improve LSA; proper term weighting makes more difference
	THD04	Thomas, Haley, DeRoeck, Petre	The Open University	assess computer science essays	LSA research for summative assessment	assess essays	used a very small, very specific corpus necessitating a small # of dimensions	LSA works ok when the granularity is coarse; need to try a larger corpus
Atenea	PGS05	Perez, Gliozzo, Strapparava, Alfonseca, Rodriquez, Magnini	U de Madrid; Istituto per la Ricerca Scientifica e Technologica	web-based system to assess free-text answers	LSA + ERB research		combine LSA with a BLEU-inspired algorithm; ie combines syntax and semantics	achieves state-of-the-art correlations to the teachers' scores while keeping the language-independence and without requiring any domain specific knowledge

Appendix A
The Latent Semantic Analysis Research Taxonomy Page 4 of 9

	Options				Training Data										
								Terms			Documents				
Refer-ence	Pre-processing	# dimensions	Weighting function	Comparison measure	Size	Composition	Subject	Number	Size	Type	Number	Size	Type	Human Effort	
DDF90	remove 439 stop words (from SMART)	100		cosine		MED	medical abstracts	5,823	words		1,033	average 50	title and abstract		
		100		cosine		CISI	information science abstracts	5,135	words		1,460	avg 45 words			
Dum91	remove 439 stop words (from SMART)	60, 100	log entropy			MED, CISI, CRAN, TIME, ADI	various (described in paper)	374 - 5831	words		82 - 1460		title and abstract		
BD095	none	70-100	log entropy	cosine										none	
FBP94		100		cosine	27.8 K	21 articles about the Panama Canal; 8 encyclopedia articles, excerpts from 2 books	Panama Canal	4829	word	prose text	607				
		100		cosine		21 articles on the the heart	heart	2,781	words	prose text					
LD97		300	ln(1+freq)/entropy	cosine	4.6M	Grolier's Academic American Encyclopedia		60.7k	word	prose	30.4k	average 151	words		
LLR97	remove 439 stop words	94		cosine vector length		27 articles from Grolier's Academic Amer. Encyclopedia	heart anatomy	3034	word	prose	830	sentence	words		
	no stop words	1500				textbook	psychology	19,153	words	prose	4,904	paragraphs	words		
RSW98				cosine		27 articles fromGrolier's Academic Amer. Encyclopedia	heart anatomy							separated essays into technical and non technical	
														created subsections of essays	
WSR98		100		cosine	17,880	36 encyclopedia articles	heart	3,034	word	prose					
FLL99						a portion of the textbook	psycholinguistics								
							standardised test - opinion essays								
							standardised test - argument essays								
							diverse								

Refer-ence	Options				Training Data										
								Terms			Documents			Human Effort
	Pre-processing	# dimensions	Weighting function	Comparison measure	Size	Composition	Subject	Number	Size	Type	Number	Size	Type	
KSS00	correct spelling			cosine		specialized texts	heart and lung	17,688	1 word	prose text	830		prose text	no pregraded summaries but mark up text into topics to appear in summaries
							Meso-American history	46,951	1 word	prose text	530		prose text	
Ste01	correct spelling			cosine		general knowledge space	sources of energy		1 word	prose text				
				cosine		specialized texts	heart and lung							
				cosine		specialized texts	Meso-American history							
Lan02		300		cosine										
WWG99		200	log entropy	cosine	2.3 MB	2 complete computer literacy textbooks, ten articles on each of the tutoring topics, entire curriculum script including expected good answers	computer literacy							collect good and bad answers
Wie00	yes , see human effort			cosine			computer literacy		1 tuple	subject - verb - object		1 tuple	subject - verb - object	segmented sentences into subject, verb, object tuples; resolved anaphora; resolved ambiguities with "and" and "or"
WG00														researcher's task to find or create appropriate texts to serve as the corpus and comparison texts
WZ01	removed 440 stop words			cosine	2.3 MB	same as WWG99	computer literacy							segmented sentences into subject, verb, object tuples; resolved anaphora and ambiguities with "and" and "or"
Nak00b	removed 938 stop words	30	log and or entropy			religious texts	religion	20,433			196			
						C programs				C code				
NPM01	removed stop words and those occuring only once	15	6 different	dot product / cosine	974 K	Huckleberry Finn and Adventures of Sherlock Holmes		5534	words	prose	487	2 KB	prose	

Appendix A
The Latent Semantic Analysis Research Taxonomy Page 6 of 9

Refer-ence	Options				Training Data									Human Effort
								Terms			Documents			
	Pre-processing	# dimensions	Weighting function	Comparison measure	Size	Composition	Subject	Number	Size	Type	Number	Size	Type	
FKM01		300		cosine		physics text book and other science text books	physics					paragraph	paragraph	prepare specialised corpora
OFK02		300		cosine		physics text book + related to curriculum script	physics	from 1,564 to 6,536	word	prose	from 416 to 3,778	paragraph	prose	sanitize corpora; write "expectations" for each answer
LD01					290K + size of course text	3 French novels plus course text	sociology of education							no pre-graded essays; mark up text into topics and notions
QKG01a						tuples representing actions in a Microworld	complex problem solving	75565	1	tuple	3441	1 trial		
BB03							various							create a classification scheme for LSI vector spaces
KKP03	removed stop words		log entropy	cosine	2.3M	used Auto tutor corpus	computer literacy	9,194	word	word - part of speech tags	5,596	paragraph	prose	part of speech tagging
NVA03	removed 442 stop words, stemming; POS	0, 10, 220, 40	various			Bulgarian	various - see paper for details							
THD04	none	10	log (no global weighting)	cosine	< 2,000	human marked answers to the essays	computer literacy				17	1 paragraph	prose	none
PGS05			tf-idf			10 different corpora: student answers + text from popular computer magazines								

Appendix A
The Latent Semantic Analysis Research Taxonomy Page 7 of 9

Refer-ence	Accuracy						Effectiveness	Usability
	Method used	Granularity of marks	Item of Interest	Number items assessed	Results			
					Human to LSA correlation	Human to Human correlation		
DDF90	evaluate using recall and precision		queries	30				
			queries	35				
Dum91	evaluate using recall and precision							
BD095	evaluate using recall and precision							
FBP94	compare against human graders	100	essay	24	0.68	.367 to .768		
	compared sentences with cosine measure							
LD97			TOEFL - multiple choice test	80	LSA: 64.4%; students: 64.5%			
LLR97	compare against human graders	5	short essay - 250 words	94	0.77	0.77		
	gold standard - a short text written by an expert			94	0.72			
	compare against human graders			273	0.64	0.65		
RSW98	compare with 1 or more target texts		short essay - 250 words	106				
WSR98	compared with 4 texts of increasing difficulty and specificity	5 point scale	essay of about 250 words	106	0.63	0.77		
FLL99	holistic - compare with graded essays		essays		0.8	0.73	average grade 85; after revisions, average grade 92	survey showed 98% of students would definitely or probably use system
				695	0.86	0.86		
				668	0.86	0.87		
				1,205	0.701	0.707		

Appendix A
The Latent Semantic Analysis Research Taxonomy Page 8 of 9

Refer-ence	Accuracy						Effectiveness	Usability
	Method used	Granularity of marks	Item of Interest	Number items assessed	Results: Human to LSA correlation	Results: Human to Human correlation		
KSS00	compare with teacher - provided topic list	10	summary of essay				no sig difference	in classroom 1997-1999; students like immediate feed back
		10	summary of essay	50	0.64	0.69	scores of those using SS for difficult texts significantly higher than those not using SS	in classroom 1997-1999; students like immediate feed back
Ste01		5	summary of essay	108			scores of those using SS for difficult texts significantly higher than those not using SS	
		5		52				
		10		52				
Lan02	holistic, Pearson product-moment correlation coefficient	5 or 10 points	essay	3,500	0.81	0.83		
WWG99	compare against pre-graded answers for completeness and compatibility	2: threshold of .55	short answers average length is 16 words	192	0.49	0.51		
Wie00	compared tuples in student answer with tuples in expected answer				.18, .24, and .4			
WG00								
WZ01	evaluate two texts using cosine							
Nak00b	created correlation matrices							
NPM01	defined precision as ration of chunks from same text to num of chunks at a level							

Appendix A
The Latent Semantic Analysis Research Taxonomy Page 8 of 9

Refer-ence	Accuracy						Effectiveness	Usability
	Method used	Granularity of marks	Item of Interest	Number items assessed	Results: Human to LSA correlation	Results: Human to Human correlation		
KSS00	compare with teacher - provided topic list	10	summary of essay				no sig difference	in classroom 1997-1999; students like immediate feed back
		10	summary of essay	50	0.64	0.69	scores of those using SS for difficult texts significantly higher than those not using SS	in classroom 1997-1999; students like immediate feed back
Ste01		5	summary of essay	108			scores of those using SS for difficult texts significantly higher than those not using SS	
		5		52				
		10		52				
Lan02	holistic, Pearson product-moment correlation coefficient	5 or 10 points	essay	3,500	0.81	0.83		
WWG99	compare against pre-graded answers for completeness and compatibility	2: threshold of .55	short answers average length is 16 words	192	0.49	0.51		
Wie00	compared tuples in student answer with tuples in expected answer				.18, .24, and .4			
WG00								
WZ01	evaluate two texts using cosine							
Nak00b	created correlation matrices							
NPM01	defined precision as ration of chunks from same text to num of chunks at a level							

Appendix B Database Tables

The EMMA database has thirteen tables; the primary keys in each table are non-meaningful, i.e., they are computer-generated sequential numbers. The name of each table in the list below is in **bold** and is followed by its attributes and short descriptions. The primary key is underlined; foreign keys are in *italics*.

Tables and Attributes

1. **MegaChunk** – the original source document
    a. megaChunkID – links this table to the Chunk table, which contains text of "meaningful" units
    b. megaChunkContents – the text
2. **Chunk –** a unit of text, can be any size
    a. chunkID - links this table to the LearningingResource Passage, EssayQuestion, EssayAnswers or MarkerComments table; a chunk can be a general document or an answer to an essay
    b. *megaChunkID* - links this table to the MegaChunk table
    c. chunkType – LP for LearningResourcePassage, EQ for EssayQuestion, EA for essay answer, or MC for marker comments
    d. chunkContents – the text
3. **LearningResourcePassage –** contains chunks from general documents, e.g. textbooks, web sites used as resources for the learners
    a. LRPID – primary key
    b. *chunkID* – links this table to the Chunk table
    c. *learningResourceID* – links this table to the LearningResource table
    d. location – where the chunk can be found in the LearningResource, e.g. page 14, paragraph 4
4. **LearningResource –** books or web sites where the chunk comes from
    a. learningResourceID – links this table to the LearningResourcePassage table
    b. title – the title of the book or website
    c. author – author or authors
    d. pubDate – publication date
    e. publisher – company that published the book or web site
    f. url – if applicable, else null
5. **EssayQuestion –** the question (or subpart if applicable) on an assessment

   a. questionID – links this table to the EssayAnswers table
   b. *chunkID* – links this table to the Chunk table
   c. *assessmentID* – links this table to the Assessment table; the assessment on which this question appears
   d. questionName – short name, e.g., Q2b
   e. pointsAvailable – the maximum points that can be awarded for this question

6. **Assessment –** the TMA (tutor marked assessment) or ECA (end of course assignment) or other exam
   a. assessmentID – links this table to the Question table
   b. *courseID* – links this table to the Course table; the course for which this assessment is given
   c. assessmentName – short name, e.g. TMA01

7. **Course –** the course that provided the essay answers
   a. courseID – links this table to the Assessment table
   b. courseNameShort – e.g. M150, U500
   c. courseNameLong – e.g. Data, Computing and Information
   d. courseStartDate – mm/dd/yy; needed to distinguish among different offerings of the same course

8. **EssayAnswer -** contains chunks that consist of answers to essay questions, can be either student or tutor generated
   a. answerID – primary key
   b. *chunkID* – links this table to the Chunk and Mark tables
   c. *questionID* – links this table to the question that this essay answers
   d. *answerSetID* – links this answer to the AnswerSet table

9. **AnswerSet -** all of the answers for a particular assessment by a particular student
   a. answerSetID – primary key
   b. *assessmentID* – links this table to the Assessment table
   c. *megaChunkID* – links this table to the MegaChunk table, which contains the original text of the answers
   d. sourceID – the anonymized ID of the person who provided the answers
   e. sourceType – S for student, T for tutor

10. **MarkerComments -** contains chunks that consist of feedback written by the marker for the learners
   a. commentID – primary key

b. *chunkID* – links this table to the Chunk table

c. *markID* – links this table to the mark for which it applies; if null, then it is a general comment, i.e., it applies to the entire answer set

d. *answerSetID* – links this table to the AnswerSetTable

11. Mark – the mark given to an essay

a. markID – primary key

b. *answerID* – links this table to the EssayAnswer table

c. *markerID* – unique identifier of marker (must be anonymous), e.g., 0 for LSA, 1 for Cathy, 2 for Robert

d. pointsAwarded – the total points for this answer

12. Marker – the person (or EMMA) who marked the answer

a. markerID – primary key

b. codedName – anonymized name of marker

13. **Essays50** – list of essays especially marked for the human inter-rater reliability study done for this dissertation

a. essays50ID – primary key

b. *courseID* – link to Course table

c. *assessmentID* – link to Assessment table

d. *questionID* – link to EssayQuestion table

e. *answerID* – link to EssayAnswer table

Entity Relationships

1. **MegaChunk** 1 ----- ∞ **Chunk**

   Ǝ (there exists) exactly 1 entry in the **MegaChunk** table for every entry in the **Chunk** table; Ǝ many entries in the **Chunk** table for every entry in the MegaChunk table

2. **Chunk** 1 ----- 0,1 **LearningResourcePassage**

   Ǝ 0 or at most 1 entry in the **LearningResourcePassage** table for every entry in the **Chunk** table; Ǝ exactly 1 entry in the **Chunk** table for every entry in the **LearningResourcePassage** table

3. **LearningResourcePassage** ∞ -----1 **LearningResource**

   Ǝ exactly 1 **LearningResource** entry for every **LearningResourcePassage**; each **LearningResource** can have many entries in the **LearningResourcePassage** table; i.e., a LearningResource contains many passages

4. **Chunk** 1 ----- 0,1 **EssayQuestion**

   Ǝ 0 or at most 1 entry in the **EssayQuestion** table for every entry in the **Chunk** table; Ǝ exactly 1 **Chunk** for every entry in the **EssayQuestion** table

5. **EssayQuestion** ∞ ----- 1 **Assessment**

   each question can occur on only 1 assessment; each assessment may have many Questions

6. **Assessment** ∞ ----- 1 **Course**

   1 **Course** may have many A**ssessment**s; each **Assessment** relates to one **Course**

7. **EssayAnswer** ∞ ----- 1 **EssayQuestion**

   every **EssayAnswer** pertains to 1 **Question**; 1 **Question** may have many **EssayAnswers**

8. **Chunk** 1 ----- 0,1 **EssayAnswer**

   Ǝ 0 or at most 1 entry in the **EssayAnswer** table for every entry in the **Chunk** table; Ǝ exactly 1 **Chunk** for every **EssayAnswer**

9. **AnswerSet** ∞ ----- 1 **Assessment**

   every **AnswerSet** comes from exactly 1 **Assessment**; 1 **Assessment** may have 0 or many **AnswerSets**

10. **EssayAnswer** ∞ ----- 1 **AnswerSet**

1 **AnswerSet** contains many **EssayAnswers**; every **EssayAnswer** pertains to exactly one **AnswerSet**

11. **EssayAnswer** 1 ----- ∞ **Mark**

    every **Mark** relates to exactly 1 **EssayAnswer**; 1 **EssayAnswer** may have many **Marks**

12. **Chunk** 1 -----0,1 **MarkerComments**

    every **Chunk** relates to at most one entry in **MarkerComments**; Э 1 **Chunk** for every entry in **MarkerComments**

13. **MarkerComments** 1 -----0,1 **Mark**

    1 **Mark** can have no associated **MarkerComments**, or at most 1 **MarkerComments**; 1 **MarkerComments** can pertain to 0 **Marks** i.e. they are general comments, or at most 1 **Mark** i.e. they are comments specific to 1 **Mark**;

14. **MarkerComments** ∞ -----1 **AnswerSet**

    each **AnswerSet** can have many **MarkerComments** (from different markers and both general and specific comments); each entry in **MarkerComments** pertains to exactly 1 **AnswerSet**

15. **Mark** ∞ -----1 **Marker**

    each **Mark** can have 1 **Marker**; each **Marker** can give many **Marks**

16. **Essays50** 0,1 -----1 **EssayQuestion**

    each entry in the **Essays50** table corresponds to 1 entry in **EssayQuestion**; each **EssayQuestion** can have 0 or at most 1 entry in **Essays50**

17. **Essays50** ∞ -----1 **Assessment**

    each entry in the **Essays50** table corresponds to 1 entry in **Assessment**; each entry in **Assessment** can have many entries in **Essays50**

18. **Essays50** ∞ -----1 **Course**

    each entry in the **Essays50** table corresponds to 1 entry in **Course**; each entry in **Course** can have many entries in **Essays50**

19. **Essays50** 1 -----1 **EssayAnswer**

    each entry in the Essays50 table corresponds to 1 entry in EssayAnswer; each EssayAnswer can have 1 entry in Essays50

Appendix C Sample Answers

Question 1

79 a. The tutor marked assignments and the end of course assessment will be delivered via the course website. b. The role of the study calendar is to provide a schedule of studies and activities for the course. The cut off date for TMA02 is the 20^th April 2004. c. Of the learning outcomes, one of particular interest is 'analyse a simple problem in terms of the necessary operations that are required to develop a program'. I feel that I will be able to apply the concepts of this skill in many areas of my profession, not only in relation to writing programs. The ability to dissect a given problem will be invaluable in pursuing my career interests. d. ETMA stands for Electronic Tutor Marked Assignment. The document 'Using the Electronic TMA System(a guide to ETMA's for students)' should be read to prepare for using the system. e. LTS, the Open Universities Learning Teaching Solutions HelpDesk, should be contacted with any queries regarding course software.

87 a. * ECA - End of Course Assessment * TMA - Tutor Marked Assessment b. * The Study Calendar is a document which informs the student where they should be in the course at any particular time. It also informs the students of deadlines for the various set tasks - CME / TMA / ECA etc * April 20^th 2004 c. Practical and/or professional skills As a long term user of a PC both at work and home, plus having passed various City & Guilds courses on PC maintenance and installation, I would like to think I am fairly proficient at an application software level and an internal workings level. As a manager at a local company using proprietary software and having to maintain personal details on my PC I would like to learn something about the legality, privacy aspect. Plus cryptography issues of software as this will also have a direct impact on my work life - I use an item called Deslock at work to encrypt files and it would be interesting to have an idea how it works. d. Electronic Tutor Marked Assessment Using the Electronics TMAs System: A Students Guide to eTMAs booklet e. LTS Student Helpdesk - various contact methods are displayed on the course website these being phone, fax, post and email. *

243 (a) The two elements of the course materials that will be distributed to us via the M150 course website are CMEs or computer marked exercises and TMAs or tutor marked assignments. (b) The Study Calendar provides an organisational aid. It gives an outline of what units should be studied at what time, gives cut off dates for assignments and shows when CMEs should be done. The cut off date for TMA 02 is 20^th April 2004. c. I feel I am most interested in achieving the following learning outcome, listed in the M150 Course Companion: `modify part of a computer program to incorporate specific operations on given data by choosing appropriate program structures.' At present, the concept of programming to me is complex and enveloped in mystery. It would be like learning a different language and therefore very satisfying. All I know about programming to date, is that it is about writing instructions for software in a language that the computer understands and to take the leap from this to actually being able to change part of a computer program would be a huge achievement for me - a powerful skill I would like to acquire and develop further. d. eTMA stands for electronic tutor marked assignment. The document `Using the Electronic TMA System' should be read in order to prepare oneself for submitting an eTMA. e. If one has any queries about course software one should contact the Student Computing Helpdesk staff (Learning and Teaching Solutions or LTS) by e-mail, First Class Helpdesk conference or telephone.

273 a. TMA's & CME's are the two elements of the course materials that will be distributed to you via the M150 course website. b. The Cut-off date for the TMA02 is the 20^th April 2004. c. The learning outcome that I am most interested in achieving is to be able to describe some of the common uses of data and how they influence the way data is stored. I feel this knowledge would be extremely useful in the current world of information systems, as data is so diverse and can be used in many formats. The sole reason the computer has evolved at the rate it has in the last ten years, is the fact that the human race needs data in order to create useful information. d. eTMA stands for Electronic Tutor Marked Assignments. The name of the document you should read in order to prepare yourself is :- Using the Electronic TMA System - A Guide to eTMA's for Students. e. LTS Helpdesk (Learning and Teaching Solutions) should be contacted if you have any queries about course software.

329 Question 1 Part (i) (a) Two of the course materials distributed on the M150 course site are: - TMA's (tutor marked assignments), and - CME's (computer marked exercises) (b) The role of the study calendar is to give the students and tutors an overview of the timescale of the course. The cut-off date for TMA02 is March 2nd 2004. (c) The area I am most interested in improving on is in the Key Skills. Having been out of education for several years, this area is the one I'm finding the most challenging, and wish to improve on. (d) eTMA stands for electronic tutor marked assignment. The document to be read before submitting an eTMA is Using the Electronic TMA System, code SUP-72860-8. (e) I would contact LTS - Learning and teaching solutions.

345 (a) TMA'S and CME'S (b) The study calendar gives dates of assignments and helps you plan when to study. The cut off date for TMA 02 is 20^th Apr. (c) I would be most interested in practical and professional skills because I one day hope to work in computing and the course will help me study at a higher level. (d) ETMA stands for electronic computer marked assignments and the document you should read is called using the electronic TMA system. (e) You should contact the LTS Student helpdesk.

999 (a) Study Calendar and TMAs (b) The purpose of the Study Calendar is to help organise your studies throughout the year. The cut-off date for TMA 02 is 20^th April 2004. (c) list the fundamental principles of information design (including principles of human-computer interaction) and apply them in simple situations; I am interested in the "human - computer interaction". Maybe this will help in developing an application for my project course TM421 which I'm also studying. (d) eTMA stands for Electronic Tutor Marked Assignment The document to use to prepare for submitting an eTMA is "Using the Electronic TMA System. A guide to eTMAs for Students" (e) LTS Computing Helpdesk.

1025 TMA M150 01 XXXXXXX 1 (i) (a) - Study Calendar and Assessments (eTMA's and CME's) (b) - The Role of the Study Calendar is to guide students to plan when to study specific subjects and gives details of deadlines for assignments. - The Cut Off Date for TMA 02 is 20th April 2004 (c) ?Demonstrate study skills at a level appropriate to higher education, such as timetabling study; read critically for meaning and take effective notes; and use study aids such as dictionaries or glossaries;? (1) I have chosen the above as the learning outcome I am most interested in achieving as I feel that studying the M150 with the Open University will assist me to develop organisational and study skills which will be invaluable in my future study and in my work environment by way of effective note taking (in training sessions and meetings), interaction with colleagues (via various media) and planning (for projects and day to day tasks). (d) - eTMA stands for Electronic Tutor Marked Assignments. - Using the Electronic TMA System A Guide to eTMA's for Students (e) The LTS (Learning Teaching Solutions) Student Helpdesk will be able to assist with queries about course software.

1175 Question 1 a. Two elements of course materials distributed via the web are: TMA's Computer Marked Assignments. ECA End of Course Assessment. b. The role of the study calendar, to show everyone study weeks and start dates, course text and activities available, cut of dates for TMA's. The cut of date for TMA02 is the 29^th of April 2004. c. I cannot pick anyone thing I have an interest in them all. The reason I have started the OU courses is to try for a degree and qualified at the end to help people to set up and use their computer and software. Teaching them in their homes, therefore I want to move later on more toward the software side on how to use and set up programs/applications. d. ETMA stand for Electronic Tutor Marked Assignments. You should read the Electronic TMA system A guide to eTMAs for students to prepare yourself. e. For software help you should ring the LTS Universities Learning Teaching Solutions telephone helpdesk

1187 a. The two elements that will be distributed via the M150 website are the CME's & TMA's and the ECA b. The role of the study calendar is to enable you to set your own study time around the relevant parts of the course and give you important dates such as when assignments must be submitted. The cut off date for TMA 02 is 20th April. c. The outcome that I am most interested in achieving is the ability to read and understand a simple computer program. I feel this is the best place to start to learn about creating my own software. d. ?eTMA? stands for ?Electronic Tutor Marked Assignments? and the document that you should read is ?Using the Electronic TMA System. A Guide to eTMA's for Students? e. For software queries you must contact Learning and Teaching Solutions (LTS) Helpdesk.

1231 Question 1 a. The two elements that are distributed via the course website are TMA's (tutor marked assignments) and CME's (computer marked exercises). b. The role of the study calendar is to organise your study during the year and advise you of cut-off dates. The cut-off date for TMA 02 is April 20^th. c. The learning outcome I am most interested in achieving is read and understand a simple computer program. I want to learn this as it will help me with other courses I am studying at the moment. d. eTma stands for Electronic TMA System. The document I should read in order to prepare myself for submitting an eTma is Using the Electronic TMA system. e. You should contact LTS Student Helpdesk if you have problems with course software.

1287 Answer 1 i. (a) The 2 elements of the course materials that will be distributed via M150 course website are * The Course News * Conferencing Facilities b. The study calendar gives guidance and direction as to when various sections of the course should be studied and also the final date of submission of assignments. Cut off date for TMA02 is April 20, 2004 c. Analyse a small computer program in terms of it inputs, programming structures and outputs (see M150 course guide P35) I choose the learning outcome because I am fascinated and enthusiastic about programming. d. Electronic Tutor Mark Assignment Using the Electronic TMA System - A guide to eTMAs for students e. The Course Tutor

1417 a) First element would be the Course Calendar and the second element is the TMA assignments. b) The study calendar is a document that helps us to organise our studies throughout the year. It gives the dates by which we should have submitted our assignments, a guide to pacing our studies and the date by which we should have submitted our ECA. The cut off date for TMA 02 is 20th April 2004 c) I think that I am particularly interested in the cognitive skills particularly with regard to programming. The analysis of problems and the use of computer programs to solve these problems together with the writing of such programs. I have chosen this particular learning outcome because I have experienced, both at work and in my hobbies, a number of situations when such skills would have enhanced my enjoyment of a particular activity. d) eTMA stands for electronic Tutor Marked Assignment. The use of this system is explained in a booklet entitled Using the Electronic TMAs System: A Student Guide to eTMAs e) The LTS Student Computing Help desk

1549 (a) Two elements of the course materials that will be distributed via the course website are the Electronic Tutor Marked Assignments (eTMAs) and the End of Course Assessment (ECA). a. The Study Calendar is designed to be a guide for organising a students studying throughout the duration of the course. It provides a guide as to when each section in the course should be started and it will give a student the relevant dates for completing the TMAs and CMEs as well as the date for the ECA. The study calendar tells us that the cut-off date for TMA 02 is the 20^th of April. b. The Course Companion lists the learning outcomes for M150. The first outcome that I would like to achieve would be to be able to "modify part of a computer program to incorporate specified operations on given data by choosing appropriate programming structures". The reason I have chosen this outcome is that I have always been interested in learning about computer programming since my time at secondary school. I would like to go on after completing M150 to do more advanced courses including those involving programming. Also I would like to be able to "demonstrate basic skills" to allow me to "progress to more advanced level studies at the OU or any other University". As I mentioned previously I am interested in continuing my studies after completing M150. I believe that the skills I will develop in this course will provide me with a good basis for future studies. c. eTMA stands for Electronic Tutor Marked Assignment. The document that should be read in order to prepare for submitting an eTMA is "Using the Electronic TMA System, a guide to eTMAs for students". d. Any student with a query about course software should firstly check the Computing Helpdesk FAQ site to see if the problem can be resolved without having to contact the Helpdesk. If it cannot be resolved then contact details for the LTS Computing Helpdesk are also provided on the website. Links to the LTS helpdesk can be found on the M150 course website.

1567 a. Two of the elements of the course materials distributed to me by the M150 Website are: i. Tutor Marked Assessments (TMAs) ii. The Study calendar b. The role of the study calendar is: i. To help with the organisation of my studies, by providing a reference as to when I need to submit my assignments, TMAs or ECAs. ii. To provide a valuable help to pacing my studies. iii. The cut off date for TMA02 is 20April 2004 c. The learning outcome I have chosen as my primary objective and am most interested in achieving is: i. Knowledge and Understanding ii. For me, if I am able to broaden my knowledge and achieve a better understanding of how computers operate, then I should in the future be able to use them as a more affective tool and source of information. a. eTMA i. eTMA stands for Electronic Tutor Marked Assignment ii. The document that should be read in order to prepare and before submitting an eTMA is: Using the Electronic TMA system A Guide to eTMAs for Students b. Regarding course software Students should contact: i. LTS Student Helpdesk; e-mail lts-student-helpdesk@open.ac.uk

Question 2

32 (a) http://uk.altavista.com/web/default (b) (1) the aquarium is called The Deep. (2) I entered "Hull aquarium" in the query where I used the search result [3]http://www.hullcc.gov.uk/news/01_june/131rd01.php to find the web site for The Deep. (c) The URI for The Deep is [4]http://195.44.57.244/ (d)(1) The minimum pages between the main site and the information of the Ballan Wrasse is 3 these are the addresses: (2)[5]http://195.44.57.244/discovery/library.php [6]http://195.44.57.244/discovery/lib_atoz.php [7]http://195.44.57.244/discovery/lib_indexb.php (3) The Ballan Wrasse can grow to 60 cm. (4) No the Ballan Wrasse page does not tell you anything about the age it can reach. (E)(1) The Ballan Wrasse can live up to 20 years. (2) The URI of the web page I used for the above answer is [8]http://www.tolgus.com/marinelife/ballan.htm (3) I used the freeserve.com search engine and put the words Ballan Wrasse in the query.

74 (a) The Uniform Resource Indicator (URI) of the UK AltaVista site is [1]http://uk.altavista.com (b) (1) The Large Aquarium in Hull is called "The Deep" 2. In order to find out the information required, I entered the search term as "hull aquarium" (minus the quotes). c. The URI of the website dedicated to "The Deep" aquarium in Hull is [2]http://195.44.57.244 This is unusual in that the URI contains only the IP address rather than a more descriptive name. d. (1) From the main site it is necessary to visit a total of three intervening web pages before accessing the page containing information regarding the Ballan Wrasse. (2) The URIs of the intervening web pages (in order visited) are: [3]http://195.44.57.244/discovery/library.php [4]http://195.44.57.244/discovery/lib_atoz.php [5]http://195.44.57.244/discovery/lib_indexb.php (3) The Ballan Wrasse can grow up to a maximum size of 60cm^(2) (4) The Ballan Wrasse page on this particular website does not contain any information regarding the age that a Ballan Wrasse can reach. (e) (1) The Ballan Wrasse can reach an age of approximately 20 years^(3) (2) The URI of the web page where this information was sourced is: http://www.marlin.ac.uk/learningzone/species/LZ_Labber.htm (3) To find this information I used the Google Search engine (www.google.co.uk) and entered the search terms "ballan wrasse age" (minus the quotes)

252 a. The URL for Altavista's UK website is: [1]http://uk.altavista.com/ b. 1. The name of the large aquarium in Hull is The Deep. 2. By entering the terms `hull aquarium' into the Altavista search engine c. The URL of the site is [2]www.thedeep.co.uk, although Altavista provided only the IP of the website, which is 195.44.57.244, rather than the URL as requested. d. 1. You have to view 4 pages to get to the page on the ballan wrasse, although the fourth is the actual page with the information, so technically the answer is 3. 2. [3]http://www.thedeep.co.uk/discovery/library.php [4]http://www.thedeep.co.uk/discovery/lib_atoz.php [5]http://www.thedeep.co.uk/discovery/lib_indexb.php 3. The ballan wrasse can grow up to 60cm in length. 4. No, the page holds information about most other aspects of the fish, but no information on what age the fish can reach. e. 1. The ballan wrasse can reach a maximum age of 29 years. 2. [6]http://www.fishbase.org/Summary/SpeciesSummary.cfm?genusname=Labrus&speciesname=bergylta 3. To get this information, I used Yahoo's search engine, with the query `max age ballan wrasse', and the above site came up as the first site on the results page.

434 ii) (a) I don't normally use the Altavista search engine. I assumed the UK Altavista search engine would be [2]www.altavista.co.uk. However, when I typed this in the address box, it returned the search engine [3]http://uk.altavista.com/, which still gives you an option to search worldwide. Its Uniform Resource Indicator (URI) is [4]http://uk.altavista.com/. (b) I formulated the "name + large + aquarium + hull" query which returned 197 hits. 1. The name of the large aquarium in Hull is "The Deep". 2. The first hit, out of the 197 hits, was [5]www.yorkshire-coast.com/hull.html. I entered into the site and found the following explanation "Famous attractions and landmarks include "The Deep" (a large aquarium complex), and the Humber Bridge, one of the world's longest single span suspension bridges" (c) At this stage I assumed that the dedicated website to the aquarium could be [6]www.thedeep.co.uk. However, using the Altavista search engine, I searched this name and no hits were returned. Probably the site was not registered with the Altavista search engine. I therefore used Google and Yahoo search engines which returned the [7]www.thedeep.co.uk website. When I typed in the URI [8]www.thedeep.co.uk in the address bar originally, it returned the URI [9]http://195.44.57.244/. This is the Internet Protocol (IP) address for the website. The IP address is rather like the address of a house whereas the website address is the house name. Two weeks later, I typed in the website and it did not revert to the IP address. (d) Find information on a fish called `Ballan Wrasse' in the fish library. 1. There are three web pages between the main site and the page that contains the information on the Ballan Wrasse. 2. List the URI of each intervening web page. i. [10]http://195.44.57.244/ or [11]www.thedeep.co.uk the main page. Clicked on the quick dive drop down menu to find the fish library. ii. [12]http://195.44.57.244/discovery/library.php - the fish library iii. [13]http://195.44.57.244/discovery/lib_atoz.php - The A-Z page. iv. [14]http://195.44.57.244/discovery/lib_indexb.php The fish library for names beginning with B and clicked on the Ballan Wrasse hyperlink. v. [15]http://195.44.57.244/discovery/popups/lib_ballanwrasse.php the `Ballan Wrasse' page. This page opens in a different window. It does not show the URI of the page. I right hand mouse clicked and went into properties to find the URI. 3. The Ballan Wrasse can grow up to 60 cm. 4. The Ballan Wrasse page does not say the age a Ballan Wrasse can reach. (e) Choose a different search engine and use it to answer the questions below. 1 The Ballan Wrasse can reach 20 years. * I found the information at the [16]http://www.tolgus.com/marinelife/ballan.htm URI. * I used the Google search engine and used the query Ballan Wrasse.

458 The URI for the Uk altavista is " Question 1 B1 The Name of the Large Aquarium in Hull is The Deep Question 1 b2 I found this by typing in sealife centres hull uk I then went into " which gave me the web site for the aquarium in Hull. XXXXXXX XXXXXXX XXXXXXX TMA M150 01 Question 1c The website URI is " Question1 d 1 and 2 There are 3 web pages you have to visit between the main site and the page that contains the information on ballan wrasse. These are: " " " Question1 d3 The ballan Wrasse can grow up to 60cm Question 1 d4 On this website it doesn't tell you the age it can grow to. Question 1 e1 Wrasse are slow growing and long lived (up to 20 years). Their longevity is also helped by being considered inedible by the British Question 1 e2 The URI of the web page is " Question 1 e3 I used the search engine URI " I asked for an advanced search inputting Ballan Wrasse Fish

584 The URI of the Altavista UK search engine is: - http://uk.altavista.com/ (b) (1) `The Deep' (2) The query, which led to my answer, was `aquarium and kingston upon hull'. (c) The URL of `The Deep' is: - http://195.44.57.244/ (d) (1) The feature on `Ballan Wrasse' is four pages from the Home Page therefore there are three intervening web pages. (2) The URIs of the intervening pages are: - (i) http://195.44.57.244/discovery/library.php (ii) http://195.44.57.244/discovery/lib_atoz.php (iii) http://195.44.57.244/discovery/lib_indexb.php (3) The `Ballan Wrasse' has a maximum size of 60 cms. (4) The `Ballan Wrasse' page does not say anything about the age they can reach. (e) (1) A `Ballan Wrasse' can live up to 20 years of age. (2) The URI of the web site leading to the information regarding the `Ballan Wrasse's age is: - http://www.marlin.ac.uk/learningzone/species/LZ_Labber.htm (3) I used `Google' as my search engine with the query "ballan wrasse"+ age.

626 a. The URI of the AltaVista site is: [1]http://uk.altavista.com/ b. 1. The name of the large aquarium in Hull is "The Deep" The query I used was "The Deep" aquarium, Hull c. The URI of "The Deep" aquarium in Hull is: [2]http://195.44.57.244/ d. 1. The minimum number of intervening web pages you have to visit between the main site and the page containing information on the Ballan wrasse is: 4 2. The URI of each intervening page are: [3]http://195.44.57.244/discovery/library.php [4]http://195.44.57.244/discovery/lib_atoz.php [5]http://195.44.57.244/discovery/lib_indexb.php The final page with the picture and facts on the Ballan Wrasse does not show a URI 3. The maximum size a Ballan Wrasse can grow is up to 60cm 4. No, the page does not tell you what age a Ballan Wrasse can reach e. 1. A Ballan Wrasse can reach up to the age of 20 2.The URI of the web page [6]http://www.tolgus.com/marinelife/ballan.htm 3. Using the search engine "Google" I typed in the query: Ballan Wrasse
954 a) http://uk.altavista.com/ b) 1. The name of the aquarium is The Deep. 2. The query The Deep, Hull UK fish aquarium led me to the answer. c) The URI of the site is http://195.44.57.244/ d) 1. Minimum pages between main site and the page displaying the Ballan Wrasse is 3. 2. http://195.44.57.244/discovery/library.php http://195.44.57.244/discovery/lib_atoz.php http://195.44.57.244/discovery/lib_indexb.php 3. A Ballan Wrasse can grow up to 60cm. 4. The page tells nothing about the age this fish can reach. e) 1. Up to 20 years. 2. http://www.tolgus.com/marinelife/ballan.htm 3. Google.com, using the query Ballan wrasse
1066 NULL
1078 (a) Using the UK Altavista search engine, find the Altavista site that allows you to enter queries. What is its URI? http://uk.altavista.com/ (b) Using the Altavista UK search engine, find the name of the large aquarium in Hull. This will require you to formulate a query (possibly several). 1. What is the name of the large aquarium in Hull? The Deep 2. Which query led you to the answer? At first I entered a search paramater of: fish + aquarium + in Hull This then lead me to the fourth search result which was from the Hull corporate press office advertising a new fish aquarium called The Deep, at the bottom of their website was a link to The Deep. I then refined my search adding The Deep this led to more results but not a direct link to the website only various press releases and information pages with links to the Deep. fish + aquarium + in Hull + thedeep (c) Still using the same search engine, find the website dedicated to the attraction. Click on the site. Hint Here is how you know you have the right site. We cannot tell you the name of the aquarium, of course, but if you assume that **** stands for its name, the site you need will appear as `****: Welcome to ****' on the UK Altavista search results page. What is the URI of the site? http://www.thedeep.co.uk/index.php (d) The site in question holds information about all sorts of fish, in a browsable fish library. Find information on a fish called `Ballan Wrasse' in the fish library. Hint Start by finding the `Browse the Fish Library' button on the main web page of the site and take it from there. Remember also that some browsers bring up windows behind the one you clicked on, and not necessarily in front, so if nothing appears to happen, remember to check on your desktop, underneath the window you started from! 1 What is the minimum number of intervening web pages you have to visit between the main site and the page that contains the information on the ballan wrasse? The minimum number of web pages is 3 2 List the URI of each intervening web page. [2]http://www.thedeep.co.uk/discovery/library.php [3]http://www.thedeep.co.uk/discovery/lib_atoz.php [4]http://www.thedeep.co.uk/discovery/lib_indexb.php 3 How big can a ballan wrasse grow? A ballan wrasse can grow Up to 60 cm 4 Does the ballan wrasse page tell you anything about the age a ballan wrasse can reach? No, there is no information on what age the ballan wrasse can reach on this particular page (e) Choose a different search engine and use it to answer the questions below. * What age can a ballan wrasse reach? A ballan wrasse fish mature around 6 to 9 years, and are hermaphroditic, with females (16 to 18 cm), changing into males (28cm). * What is the URI of the web page where you found the information? http://fp.kevthefish.f9.co.uk/wrasse%20species.htm 3 Which search engine, and which query got you to the page that contained your answer? The search engine I used was [5]http://www.ask.co.uk/ (Ask Jeeves) The query I used was ballan wrasse + age
1208 (a) The URI for the UK Altavista site is: http://uk.altavista.com/ (b) The name of the large aquarium in Hull is called The Deep. The search results from "hull aquarium tourist attraction" did not give the actual site for The Deep but did give the correct answer. However, doing a search for ""the deep" +hull" did give the website. (c) The URI for The Deep is: http://195.44.57.244/ (d) The minimum number of intervening pages is three (The Fish Library, Browse A-Z and The Fish Library - B). The search function on the home page was not working when I tried the site. The URIs for the intervening web pages are: http://195.44.57.244/discovery/library.php http://195.44.57.244/discovery/lib_atoz.php http://195.44.57.244/discovery/lib_indexb.php A ballan wrasse grow up to 60cm. The web page does not mention anything about the age a ballan wrasse can reach. (e) A ballan wrasse can reach up to 20 years old. This information was taken from: http://www.marlin.ac.uk/ learningzone/species/LZ_Labber.htm The search engine was Google and the query was: ""ballan wrasse" age".
1300 (ii) (a) http://uk.altavista.com/ (b) 1. "The Deep" 2. aquarium hull (c) [1]http://195.44.57.244/ (d) 1. At a minimum by using the Browse A-Z, there are 3 intervening web pages. 2. [2]http://195.44.57.244/discovery/library.php [3]http://195.44.57.244/discovery/lib_atoz.php [4]http://195.44.57.244/discovery/lib_indexb.php 3. The Ballan Wrasse can grow up to a maximum size of 60cm. 4. The Ballan Wrasse page on "The Deep" web site does not tell you the age a Ballan Wrasse can reach. (e) 1. A Ballan Wrasse can reach the age of twenty. 2. [5]http://www.tolgus.com/marinelife/ballan.htm 3. Search Engine: AltaVista UK Query: "Ballan Wrasse"
1372 a) UK Altavista URI = http://uk.altavista.com/ b) 1. The name of the aquarium is "The Deep". 2. query = hull aquarium (1st hit had an article on the architecture of the aquarium) c) http://195.44.57.244/ d) 1. 3 intervening web pages 2. browse fish =1 http://195.44.57.244/discovery/library.php search a-z =2, http://195.44.57.244/discovery/lib_atoz.php letter B = 3, http://195.44.57.244/discovery/lib_indexb.php 3. 'up to 60cm' (though another site about marlins which gives the age for question e)1. below says "up to 50cm") 4. No - there is no information on the age it can reach. e) 1. up to 20 years 2. http://www.marlin.ac.uk/learningzone/species/LZ_Labber.htm 3. I used 'Google' and the query "Labrus Bergylta" age

Question 3

8291 An analogue quantity is something that does not have an exact value. For example the volume mechanism on a radio. Using a knob you can control the volume simply by rotating it. The volume will increase gradually and not in a series of jumps. However a discrete quantity will increase in a series of clear steps. Such as the volume on a digital radio which increases with a button. Where on each press the volume increases by one step.

8671 An analogue system has an infinite number of possible states, where as a discrete system has a set number of possible states. As an example of an analogue system, the wiper blades on a car's window screen have an infinite number of possible positions on the window screen. The switch that controls the wiper blades is an example of a discrete system, with off, slow, medium and fast settings.

9241 An analogue quantity is one that changes continuously like sound, light, temperature etc that our senses deal with every day. The change happens at infinite smoothness. A discrete quantity changes in clear steps instead of continuously, an example of this would be if you use an for an electrician who uses a needle meter to measure voltages you can see the movement continually happen, but if you used a digital meter the voltage would be shown in numbered steps no mater how many decimal points it goes to.

10400 Explain, with examples, the difference between an analogue and a discrete quantity. We would class most things in our human world as having an analogue quantity, what this means is that if we take something like the colour spectrum, and try to analyse the different colours to see if there is a break between them, we would be unable to do this, as one colour merges into the next colour, with no definitive break between the colours except that one colour will at some point take over from the next. In our own minds we would know that at the point where the next colour takes over, we would expect a dividing line between the two, however with analogue quantities, they have an infinite number of possibilities about then. So for the example of colours we would never be able to say that there is a stop and start due to the fact that the different and merging colours would have infinite possibilities of shades of each colour. Another example would be a form of measurement of an object, again the measurement could be infinite, as we can now get down to molecular measurement of objects, and even beyond this, it may be possible that measurement may never stop. A discrete quantity has an altogether different quality about it, yet in some ways we could be applying the quantity to a similar object that we may use for describing the analogue quantity. A discrete quantity, is something we may alter from one state to another, yet is done in definitive steps, and each step has a boundary line, break or could even mean switching form one state to another. The example that could describe this is one of temperature control of an oven with a digital readout for temperature. The control would possibly be a dial, which one would turn to a temperature required, in turning the dial, we would feel the dial, action clicking through pre set parameters. The parameters would be taking the control up or down from one parameter to another with a definite end of one state to a definite beginning of another. The discrete quantity of measurement for the digital readout is where the temperature is measured in degrees, and does not have other measurement after a decimal point, therefore the measurement would go from one decimal digit to another, to another. There is no infinite possibilities in between as there are definite steps from one digit to another.

11350 An analogue quantity is something that changes continuously and has an infinite number of possible changes examples are an analogue thermometer and an analogue volume control. A discreet quantity is something that changes in clear steps for example the number of people at a concert or the price of a book. (i)

13136 The difference between an analogue quantity and a discrete quantity is that an analogue quantity is one that is continuously changing whereas a discrete quantity changes in a series of steps that are very clear. An example of an analogue quantity is a thermometer or temperature gage that has an infinite number of temperature differences between each degree. The thermometer moves up continuously and smoothly through all the possible degrees. An example of a discrete quantity is one that has a fixed number of values between the two points on any scale. For example a volume knob that you turn in stages that clicks into place as the volume gets louder.

15644 Explain, with examples, the difference between an analogue and a discrete quantity. Analogue quantities have continuous (and infinite) values and generally involve a measurement of sorts, e.g. weight, volume, height, speed, temperature, time, etc. In any measurement of an analogous quantity there will always be an infinite number of values between any two given points. For example values between 1 and 2 include all of the following values: 1.5, 1.55, 1.555, 1.5555 and so on. Examples of analogue quantities might include: * The length of a branch on an apple tree * The weight of an apple from the tree * The circumference of any one apple Discrete quantities are characterised by having a finite number of values (often whole numbers) and are generally measured by counting. Examples of discrete quantities might include: * The number of branches on an apple tree * The number of apples that the tree bears in a season * The number of pips contained in any or all of the tree's apples. (i)

17430 Analogue quantities change smoothly and continuously, discrete quantities change in noticeable steps. A thermometer using mercury that expands when it heats up moves up the tube smoothly without any sharp jumps, this is analogue. Where as a digital thermometer will not move slowly through the temperatures but will jump from 12.5^0C to 13^0C being a discrete method. Another example is measuring a gap between numbers. For example 29 - 30, an analogue view would be there is an infinite amount of numbers between these two numbers. A discrete method would round these numbers up because the numbers would get so small it just wouldn't be needed.

17544 An analogue quantity is one that changes smoothly and continuously such as when the volume on a radio is turned up or down. A discrete quantity is one that goes up or down in a series of steps such as the price of petrol which goes in a series of steps of per pence per litre.

18000 A good example of an analogue device is a clock with hands as they change continuously caused by a small motor. On many marine projects I work on, we have temperature probes that are made from material which changes its resistance as its temperature changes. This provides a signal to a computer or a moving coil meter. An example of discrete quantity is a roll of ciné film, which is made up from a number of frames. Even when the film is running it gives the appearance of an analogue quantity, when it is slowed down it shows a series of clear steps which is discrete. Another example is your mortgage payments which increase and decrease in fixed ranges such as 0.4%, 1% etc. Therefore analogue quantities move in an infinite movement, and discrete quantise move in a stepped movement.

19577 Analogue quantities are quantities that are continuous, meaning that they can be measured at any point and the points at which they can be measured are infinite in number. An example of this is sound which is a continuous wave. Discrete quantities are quantities that change in distinct steps with a fixed number of values between two points, such as a ruler. This then becomes confusing because

people would say that there are an infinite number of measurements on this, but we assign it with certain marks and it goes in defined steps this give it an appearance of being discrete and we treat it as such.

19824 Analogue is a constantly changing quantity/measurement i.e. volume, length, width, pressure, heat. Discrete is a definite quantity that is a countable set of values i.e. number of people in a car, price of a newspaper, number of pages in a book.

20546 The term analogue is used to describe a continuously changing source of information. An example of an analogue data source would be the sound made by a musical instrument. Although the individual notes in the music are discernible, the instruments sound wave if viewed on an oscilloscope would show a continuous wavy line. A discrete quantity is a set of distinct data values. For example, I have 13 wooden file boxes on the shelves in our study. Each file box is a separate entity and can be counted individually, unlike the sound wave produced by the musical instrument.

20869 An analogue quantity can be defined as one that may be measured to represent an infinite number of values or a value that varies continuously with time. An example of an analogue quantity could be the altitude of an aircraft, travelling between ground level and 1000 feet the aircraft must pass through each of an infinite number of intermediate altitudes (999 ft, 999.1ft, 999.11ft etc). Conversely, a discreet quantity is one that is measured in defined and finite steps, for example, the altimeter on our aircraft could give a discrete measure of the aircrafts altitude, perhaps giving a value for every 10^th of a foot and ignoring the infinite number of possible values (analogue values of altitude) that lie in the space between each 10^th of a foot.

21154 Analogue quantities are not restricted within a defined set of values and are said to change continuously. An example of an analogue quantity is shown with a dimmer switch on a light, as you turn it towards high the light gets brighter but the transition between low and high is continuously. Discrete quantities can be measured to one of a set of exact values i.e. it is made up of a defined set of values. An example of a discrete quantity is shown by the dial on safe, as you turn it you can hear it click as it moves to the next/previous value between a defined set ranging from 1 to 50.

21458 Analogue quantities are ones that change continuously. (That is to say they change smoothly not constantly). There are no distinct `steps' as the quantity changes. For example turning up the volume on a radio - the volume increases smoothly and constantly with an infinite number of volumes between low and high. Another good visual example is the column of mercury in a thermometer - when the temperature increases the mercury rises in one continuous movement and does not jump from one degree to the next. In contrast discrete quantities do change in a series of clear defined steps. Take the example of the thermometer - a digital thermometer has a reading that changes in distinct values (either 1 or 0.5 degrees at a time) ignoring all the possible values in between. Anything we can count is fundamentally a discrete quantity - for example the price of an item in a shop or the number of people at a cricket match.

21648 An analogue quantity is something that changes continuously without noticing , an example of this being the volume control on a Hi-fi which increases steadily through the ranges when turned from low to high. A discrete quantity unlike an analogue quantity increases in a series of clear steps, for example in the case of the price of petrol which increases by a penny at a time so we see clear increases in the cost.

22940 Analogue quantities are not bound by precision. Within a range can they represent an infinite number of positions. Discrete quantities however are bound by precision and can only represent a limited amount of values. If a person were to be measured using a measure that only measured in feet a discrete value would be produced whereas in reality the persons height could be somewhere between two discrete values. For example if a person were 6.5 feet tall using the discrete system they would be measured as either 6 foot or 7 foot tall depending on how the system deals with halves.

24745 Analogue quantities change continuously and will always have an infinite number of steps between measurable changes. An example of an analogue quantity is the temperature of an oven. A discrete quantity changes in a series of clear steps with no intervening steps being available. An example of a discrete quantity is the number of people in a football crowd. (b) An example of a computer standard is the GIF image compression standard. It is used to reduce the number of bits used to store each pixel and therefore the overall file size of the image. There is a general need for standards in computing because it enables exchange of information between computers so that any computer is capable of receiving and displaying information created or forwarded by another computer with exactly the same output displayed on each machine

25144 he differences between an analogue and a discrete quantity are: 1). Analogue quantities are continuous i.e. when they change, they change at a steady rate without jumps in the quantity. For example a dimmer switch can be used to adjust the brightness of a bulb steadily. Discrete quantities increase in jumps, e.g. when a light switch is pressed to the on setting, it jumps from off to lit in one step. 2). Discrete quantities have a particular value whereas analogue quantities are changeable. For example, the number of tea bags in a box is a discrete quantity as there is a set amount in the box.

27519 An analogue quantity is one that can vary to an infinite number of values. The quantity measured will depend on the precision with which it is measured. The speed of a car is an example of an analogue quantity. A discrete quantity is one that changes in a series of steps and as such can be measured exactly. An example of a discrete quantity is the number of people travelling in a car.

28621 In order to explain the difference between analogue and discrete quantities, it is necessary to give definitions of both terms. Analogue quantities change continuously. In order to demonstrate this term, examples help. For instance, temperature is an analogue quantity as when it is measured, temperature rises continuously and smoothly through two points. It moves through all the temperatures in-between. Another example is volume. When increasing or decreasing volume using a control, it moves smoothly through the different levels of volume. Analogue quantities do not jump through the different levels; rather it is a progressive movement. In contrast, discrete quantities are those that change is a series of clear steps. Temperature and volume are also discrete quantities depending on how it is measured. For instance a clinical thermometer shows temperature using a given scale which misses out certain measurements because of the instrument used, as does some volume controls who have specified or fixed values on it.

28792 An analogue quantity is continuously changing with an infinite amount of measurements between two points on a quantities scale, an example could be a persons walking pace. A discrete quantity rises and falls in clearly defined steps within a scale and can never be between two points. Currency, in everyday use by the public, is a good example, it's smallest units cannot be split.

29134 An analogue quantity is a natural continuous smooth movement passing through all the infinite possible points between the 2 extremes. A discrete quantity progresses through all the possible points between 2 extremes in a series of tiny clear steps. An analogue quantity is what humans experience in the real world, whereas the discrete quantity is only experienced in the digital world of man made instruments. The weight of a mound of earth is an analogue quantity whilst the number of clods of earth used to build the mound would be a discrete quantity as their will be a finite number of clods used.

Question 4

9090 A `computer standard' allows for a common practice to be maintained within the computer industry. An example of a standard may be the American Standard for Code Information Interchange (ASCII) system. ASCII allows for any user to understand information written by either another user or for 2 machines to talk to one another. The need for `computer standards' allows for a consistent and uniform method of storing information.

9489 An example of a computer standard would be the design of the PC floppy disk drive and the 3.5-inch disks that use them. The standard construction of the drives and the way that information is encoded on the disk, allows computer users to store information with the knowledge that it can be accessed and used by other compatible computers. Standards have become important in computing to ensure that software and hardware work reliably and as expected by computer users. The agreeing of standards, or their setting by a de facto system, allow computer users to easily share information and increase productivity. Manufacturers of hardware and software gain by having access to a much larger market than they would have for a single platform product.

9622 There are numerous computer standards; one example is JPEG, popular for displaying photographs. This is a standard format for images, in the way an image is compressed for storage on a computer and in turn decompressed in order to display the image on screen. There is a general need for standards in computing so as to ensure compatibility between different computers. If standards are followed it makes it easy to share data. If, for example, your office and home computers did not use the same standards it would be impossible for you to work on a document at home that had been created at work, using different standards, as your home computer would not understand how to read the data, or vice versa.

10021 An example of a computer standard is the TCP/IP protocols that provide the ability for computers to transmit and receive data from other computers over a network. Without this standard computers would be unable to interpret communication from other computers. Standards are required not only between computers but also between different computer programs.

10230 There is a general need for standards in computing because there are so many people using computers with different programs, that they must be compatible with each other, otherwise a lot of software / hardware wouldn't work. For example, take the MPEG2 compression standard for movies. If there were many different standards for movie compression, all of the hardware which writes such data to disc, the hardware that reads the data, and the software that plays the data, would have difficulty in reading such files if everyone decided to save them in a different format. I have downloaded music from msn.co.uk in Microsoft's new format - .wma. This standard is not compatible with my software to burn music files as this relies on the MP3 standard. This would also be an example of introducing a new standard - unless more people use it, it will be difficult to achieve convergence over different platforms.

11655 A good example of a computer standard is ASCII (American Standard for Computer Information Interchange) which uses numbers that computers interpret to represent text characters (i.e. 65=A, 66=B etc.). Standards are created by large groups of people who agree on the meaning of the representations involved. There is a general need for standards in the computing industry because it enables all computer users to share and use data and information without having to convert the data into a format that they can work with. For example, I created this document using Microsoft Word that is an Industry Standard (or more precisely - a de facto standard which means that this application has become so popular that it is widely accepted as a standard format for text documents) and in order for you to read this document, you too must be using Microsoft Word. If there were no standards and everybody made up their own, then nobody would not be able to exchange information in this way. Question 1 (Continued)

12776 An example of a computer standard is the UNICODE standard for character representation. Computers require standards so that they can communicate with each other and with peripheral equipment. For example, if this standard did not exist the computer of the tuter marking this assignment would not be able to read this.

14087 A computer standard is an agreed compatibility between components used in computing both software and hardware, this is required to allow the movement of data either between two computers or a computer and an add-on device such as a printer. A common example of a computer standard is a word processing text format, when writing this document I need to ensure it can be read by my tutor on his personal computer.

15588 NULL

15683 An example of a computer standard is the http protocol (hypertext transfer protocol) used for transmitting data over the Internet. This standard method enables the Internet to be global and accessed by all regardless of geographical position, computer hardware, software or network used. Standards in computing are necessary, otherwise the Internet would not be successful as a means of sharing and sending information as either the file format or the form of compression would not be able to be read or decompressed by the recipient.

18780 There is a computer standard for the representation of keyboard characters, called ASCII (American Standard Code for Information Interchange). It assigns all characters a specific number to be used by computers to facilitate the conversion from keyboard input to computer instructions and back. This demonstrates the need for computing standards in general; in the above example, if different computers allocated different numbers to characters, documents would not be able to be transferred between computers and make sense or mean the same.

21459 An example of a computer standard is the JPEG (Joint Photographic Experts Group) standard. This a technique used to compress a photograph prior to sending it to another computer. In order for the recipient to decompress and view it there must be an agreement, or standard, between the sender and recipient as to the technique used. I.e. The recipient would also have to use JPEG to view the photograph. There are many computer standards covering every aspect of computing - without them people would not be able to share information, read each others documents, view each others images and so on.

21972 Computer standards are rules that govern programs in the way that they treat binary information. For example JPEG (Joint Photographic Experts Group) is a standard that is used to compress/decompress image files. The standard JPEG allows many different programs to use the binary information for the purpose that is required for example word processing, desktop publishing, web design, picture editing to mention a few, if the standard was not widely distributed then the file would only be readable by a few or even only one program. For certain uses a standard that can only be read by a few limited computers may be an advantage. In conclusion a standard allows binary information to be shared by many or a few depending on the use of the information.

Question 8
8030 The <B> was missing from the end, it should read <B>Always look left and right before crossing the road,</B>
9759 <B> Always look left and right before crossing the road. </B> As this line is to appear emboldened the bold tag <B> needs a corresponding closing tag which in this case is </B> to indicate where the emboldening is to end. As the original text did not contain the closing tag the result would have been that any following text would have appeared emboldened.
9968 Original HTML <B>Always look left and right before crossing the road. Intended appearance. Always look left and right before crossing the road. Corrected HTML <B>Always look left and right before crossing the road.</B> Although the original HTML will in fact display the text as intended, the command to turn off the bold setting </B> is required to prevent anything that follows from also appearing as bold.
10006 <B>Always look left and right before crossing the road.</B> No closing HTML tag was placed at the end of the sentence in order to indicate the end of the bold tag. Potentially the rest of any further text would be in bold.
10405 <B>Always look left and right before crossing the road - should appear as: <B> Always look left and right before crossing the road</B> There was no closing tag at the end of the text, telling the computer how to style the text.
10462 <B>Always look left and right before crossing the road </B> The stop bold tag </B> is missing from the end of the sentence.
10899 <B>Always look left and right before crossing the road.</B> The tag <B> which produces Bold text requires a closing tag. A closing tag is the opening tag with the addition of / thus forming </B>
11431 Although the example would appear as it should, there is | | | no closing bold tag so any further text would also be in | | 2) i) | bold. | | | The fragment should be written like this: | | | Always look left and right before crossing the | | | road.
11583 <P> <CENTER> <B> Always look left and right before crossing the road.</B> </CENTER> </P> * The original HTML did not indicate that the statement should stand alone, therefore to achieve this tags of <P> AND </P> have been inserted. * No end tags for bold after the word `road'. * No tag to indicate that the statement is centred on the page.
11697 <B>Always look left and right before crossing the road.</B> The end tag is missing from the original fragment, so the bad command would be ignored.
11868 <B>Always look left and right before crossing the road</B> The HTML Bold tag </B> was missing, this is required to delimit the area of bold text.
11925 In question (i) the writer wishes the sentence to be in a bold format, to do this correctly, you must indicate were you wish the bold formatting to finish using the </B> tag. The HTML should read: <B>Always look left and right before crossing the road.</B>
12305 The problem with the original it has no closing tags. The computer would not recognize where to end the BOLD tag and would carry on reading the rest of the paragraph as bold. <B>Always look left and right before crossing the road.</B>
12324 For the text to appear as: Always look left and right before crossing the road. the correct HTML to use is: <B>Always look left and right before crossing the road.</B> The HTML as stated in the question: <B>Always look left and right before crossing the road. does not have an end tag `</B>' which is not following best practice. It is considered best practice that when writing HTML there is both a start and end tag, however, there a few exceptions to this rule. If the code were to be left like this, the sentence would appear as required, in bold, however, if more text was added to this line, it to would appear in bold.
12400 The HTML fragment should appear as follows: <B> Always look left and right before crossing the road. </B> The browser will then read this as Always look left and right before crossing the road. The problem with the fragment before was that the </B> end tag telling the document that bold type is no longer needed was missing. Without this closing tag, everything else in the document appearing after this line of text would be in bold type.
12571 <B> Always look left and right before crossing the road.</B> The bold tag was not closed. This would make any further sentences bold as well.
12970 Always look left and right before crossing the road. This should appear as follows: Answer: Always look left and right before crossing the road The first line of text above has not got the end tag
13331 <B>Always look left and right before crossing the road.</B> This code was missing the closing bold tag.
14357 <HTML> <HEAD> <BODY><B>Always look left and right before crossing the road.</B> <P> </BODY> <HTML> The original HTML does not have a correct structure and the bold print has not been closed off.
14680 <B>Always look left and right before crossing the road.</B> The original HTML text did not have a closing tag, which would probably be considered bad HTML, and may not create the desired effect in some browsers. Any text after the sentence that is not supposed to be in bold would be displayed in bold until a closing tag was used.
14946 To convert decimal 1183 to hexadecimal I would use the same principle as used for converting to binary. Each value increases by the value of x16 as shown in the table. +-------------------------+ | 4096 | 256 | 16 | 1 | |------+-----+----+--------| | 0 | 4 | 9 | 15 (F) | +-------------------------+ 1183 divided by 4096 won't go so 0 goes in the 4096 column. 1183 divided by 256 equals 4 with 159 remaining. So 4 goes in the 256 column. 159 divided by 16 equals 9 with 15 remaining. So 9 goes in the 16 column. 15 divided by 1 equals 15 with none remaining. So 15 goes in the 1 column. Because hexadecimal uses the letters A to F for the numbers 10 to 15 then the letter F is used in the 1 column. Therefore the decimal number 1183 is Hex 0[x] 4 9 F.
14984 <P><B>Always look left and right before crossing the road.</B> There was no bold closing tag, which is necessary (assuming the bold styling was to end there). The paragraph tag is needed to produce the blank line separation formatting.
16333 The original says: <B>Always look left and right before crossing the road. And should appear: Always look left and right before crossing the road. The problem with the HTML is that it doesn't have the closing tag </B> after the statement. Should read: <B>Always look left and right before crossing the road.</B>
16732 . Always look left and right before crossing the road The author requires the text to be in bold letters but has omitted the end tag which always has a slash within the angled brackets i.e. Without it the hypertext will not recognize the command and the text will not be displayed in bold.
16789 <B>Always look left and right before crossing the road.</B>
17359 <B>Always look left and right before crossing the road</B>. The highlighted closing tag was missing from the original. The sentence would not appear in bold as this type of tag must always be closed
17378 The correct HTML for this should be, <B>Always look left and right before crossing the road.</B> The problem in the example given was that there was no closing BOLD tag.
17796 I have corrected the HTML (shown below) so that the tag begins and ends with the HTML tag. I have also closed the bold tag. <HTML><B>Always look left and right before crossing the road.</B></HTML>

Question 9

7594 Important </B> Do <B> not </B> place metal items in the microwave First edit was to correct the closing tag at the end "Important" by adding the / Second edit was to add a closing tag after "not", otherwise the whole rest of the sentence would be in bold a not just "not".

7746 Correct HTML : <B>Important!</B> Do<B> not</B> place metal items in the microwave. With only the use of the <B> tag, the original HTML would make all the text bold. The <B/> tag is inserted to turn off bold as required, ensuring that the <B> and <B/> are used to turn on and off the bold type in the correct places.

8772 <B>Important!</B> Do <B>not</B> place metal items in the microwave. The bold tag after the word `important!' needs to be an end type, otherwise the following text will also appear bold. An end type tag is also required after the word `not' for the same reason.

9513 NULL

9817 <B>Important!</B> Do <B>not</B> place metal items in the microwave. This is similar to the error in (i) above in respect of the failure to use closing tags. Failure to use the correct tag after the first word suggests confusion with the Bold toggle employed in programs such as WinWord.

10330 <B>Important!</B> Do <B>not</B> place metal items in the microwave. No use of closing tags, 'Important' and 'not' may not be seen in bold.

10615 <B>Important!</B> Do <B>not</B> place metal items in the microwave. No closing tags had been used which meant that the whole sentence would have been displayed in bold text. If you want only certain parts to be in bold they must be enclosed in bold tags.

11242 <B> Important! </B> Do <B>not </B>place metal items in the microwave. In this case everything was appearing in bold due to missing tag indicating ending of bold writing after ?Important!?, which means that the text will continue to appear in bold, also after ?not? there wasn't a bold-ending tag either.

11261 <B> Important! /B> Do <B> not </B> place metal items in the microwave. The problem with the original HTML in this instance is that what should be an end tag after the word `important!' is in fact another start tag. Also, an end bold tag </B> is missing from after the word `not' and again, if the browser can not recognise a full pair of tags it can not style the text as intended.

11774 <B>Important!</B> Do <B>not</B> place metal items in the microwave.

12002 The correct HTML is: <P><B>Important!</B> Do <B>not</B> place metal items in the microwave.</P> The original code had no closing bold tags and no opening or closing paragraph tags. The failure to use closing bold tags would have meant that all the text would have been in bold type. The attempt to apply bold type to the word `not' failed as there was no closing of the bold tag after the word `not' and even if there had been, the failure to close the first bold tag would still have resulted in all the text being bold.

12610 The correct HTML is:- Important! Do not place metal items in the microwave. Unlike the first example this sentence would not have appeared correctly as it would have been all in bold and again so would the rest of the document.

12914 <B>Important!</B> Do <B>not</B> place metal items in the microwave. Is the correct format to display the proper line in html as before you must close tags like <B> with a </B> unless you want other text to be bold.

13009 <P> <B>Important!</B> Do <B>not</B> place metal items in the microwave. </P> The original fragment lacks the starting Paragraph tag <P> that would provide a blank line between this text and any preceding text. In addition the original turns Bold on with the <B> tag at the start of the text and then tries to turn it on again after the word `Important' rather than turning bold off with an Close Bold tag. Bold is once more turned on after the word `Do' but not turned off after the word `not' the result would be the entire sentence appearing in Bold.

13066 The corrected HTML is: <B>Important!</B> Do <B>not</B> place metal items in the microwave. The original HTML had an opening bold tag after the word Important instead of a closing bold tag which nested the bold tag rather than terminate it. This caused the word Do to be bold. The HTML was also missing a closing tag for the bold tag after the word not so the whole line of text ended up bold.

13180 The corrected fragment should be: <B>Important! </B>Do <B>not</B> place metal items in the microwave. This fragment is missing the closing part of the first bold tag for the word `Important!' and also the closing bold tag for the word `not'.

13408 <B>Important!<B> Do <B>not place metal items in the microwave. This should appear as follows. Important! Do not place metal items in the microwave. The Bold tag before the first word should have a matching /B tag to finish highlighting in bold (the Slash is missing from the tag). Again the Bold tag to highlight the word "not" has no matching finishing tag, which would mean that the rest of the sentence would be bold not just the word intended. The correct HTML is : <B>Important!</B> Do <B>not</B> place metal items in the microwave

13427 <B> Important! </B> Do <B> not </B> place metal items in the microwave HTML tag must be in pairs, an opening tag, and a closing tag preceded with /, in this case the closing tag is missing the /, and the 2^nd closing tag is missing.

13446 <B>Important!</B> Do <B>not</B> place metal items in the microwave. The original HTML did not have a bold closing tag after `Important!' therefore the entire text would have been in bold type.

13465 <B> Important! </B> Do <B> not </B> place metal items in the microwave. The opening tag was used instead of the closing tag after Important! Also there is no closing tag after not.

13712 correct HTML- <B>Important!</B> Do <B>not </B>place metal items in the microwave. In the example given, there were two mistakes firstly, the word ``Important!" had been preceded by a correct start tag for `bold' <B>, but terminated by another start tag, rather than by an end tag </B>, which applies the bold style to that word alone. Secondly, the word ``not" required an end tag </B> immediately after, which was absent - each application of this particular text markup must be applied and concluded each time it is required.

15118 <B>Important!</B> do <B>not</B> place metal items in the microwave The second bold tag did not have a / to make it a closing tag. The third bold tag did not have a closing tag.

15232 <B>Important!</B> Do <B>not</B> place metal items in the microwave. The text styling closing tag after the word `important' had the / missing. This made it into another opening tag, which would be ignored. Thus whole fragment would be in bold. After the word `not' the closing text styling tag had been omitted. Thus the remainder of the fragment would be in bold.

15688 <P><B>Important!</B> Do <B>not</B> place metal items in the microwave.</P> The original HTML had a forward slash missing from the closing bold tag after the word `Important!'. Also there was no closing bold tag after the word `not'. As in (i), I have also put paragraph tags to isolate the line from any surrounding text.

Question 10

7481 The HTML fragment should be: It is <I><B>very</B></I> important to read this text carefully. There were several errors made in the original fragment. Firstly, the italics tag <I> was inserted at the beginning of the sentence; assuming the rest of the HTML was correct, the words `It is' would have been displayed in italics when it was not required. Secondly, the two sets of tags (<I></I> and <B></B>) overlapped, and were not nested. Where a tag starts inside another pair of tags, it needs to close within the original pair in order to work correctly. If this were not corrected the text would not be displayed as required.

7614 It is <B><I>very </I></B> important to read this text carefully. The original fragment starts with an italic tag, resulting in the first two words being italic, with the third being italic and bold. The end tags are okay.

8146 It is <B><I>very</I></B> important to read this text carefully. The original fragment uses the italicise tag <I> at the beginning of the sentence, within the HTML structure. In this case the browser will italicise all the text prior to the word very, then embolden and italicise very, before displaying the rest of the sentence with no styling. Eg: It is very important to read this text carefully. It is important to note that when using nested tags, that they be opened and closed in the correct order as in; <tr><td>Table</td></tr>

8374 To make this ~ <I>It is <B>very </I></B>important to read this text carefully. Read this ~ It is very important to read this text carefully. The italic writing should not start until the start of "very", just now it starts at the word "It" Also, the start itialic (<I>) should be after the start bold as after the word "very" it closes in the opposite order that it starts. It should read this ~ It is <B><I> very</I></B> important to read this text carefully.

9001 It is <I><B>very</B></I> important to read this text carefully. The text `This is very' would have all been formatted in italics in this example because the opening tag, <I> appears at the beginning of the sentence rather than just before the word to be formatted in italics. The bold tag would not have worked because HTML requires you to close the inner tag before closing the outer one. </B> should therefore have appeared before </I>.

9799 It is <B><I>very</I></B> important to read this text carefully. This text uses the bold and italic tags to stylise the word "very". For this to work correctly the <I> tag needs to be next to the word and nested inside the bold tags.

10711 <I>It is<B>very</I></B>important to read this text carefully. The problem with this fragment is that, by placing the italic tag before the words `It is would result in these words being italics, it should be placed before the word `very'. It is <I><B>very</I></B>important to read this text carefully.

11110 It is veryimportant to read this text carefully. These are nested text styling tags, and the closure tags need to be ordered according to the position of the opening tags. Here the italic tag is nested inside the bold tag, and is used in the body of the html.

11528 It is <B><I>very</I></B> important to read this text carefully. The original fragment has the opening italic tag in the wrong place. It is at the beginning of the sentence. It should be nested inside the bold tags before the word `very'. The original fragment would put the words `It is very' in italics which is incorrect.

11699 It is<I><B>very</B></I>important to read this text carefully. In this example the closing tag for bold had been put in the wrong place. . inner tags must be closed before outer tags. Also the start tag for italic is in the wrong place.

12193 It is <I><B> very</B></I> important to read this text carefully. The first italic tag should have been before the bold tag and not at the beginning of the sentence. You also should close the inner tag first and then the outer tag.

12573 It is <B><I>Very</I></B> important to read this text carefully The Italic tag before? It is? should be placed before ?very? .If not, ?It is? would be in italics as well.

12725 It is <B><I> very </I></B>important to read this text carefully. As originally shown the words displayed between the <I> tags would have been displayed in the requested fashion, "italics" but there would have been no "bold" text.

12858 It is <I><B>very</I></B> important to read this text carefully. Problem: The first <I> tag should be before the word `very' not `It is' (in front of the <B> tag).

12953 It is <I><B>very</I></B> important to read this text carefully. The opening italic tag was in the wrong place, it should be as above.

13276 It is <B><I>very </B></I> important to read this text carefully. To only display the word `very' in bold & italics, the original fragment of text had used the italic start tag in the incorrect place. Using it at the start of the sentence would result in `It is very' all being displayed in italics. The bold tags were actually placed correctly. In summary, the bold & italic tags should have been nested around the word `very'.

13371 It is<I><B> very </B></I> important to read this text carefully The first italic tag was placed to soon in the sentence it should have been directly in front of the word to be changed. When using nested tags you should always close the inner ones first before closing the outer tags.

13732 It is <B><I>very</I></B> important to read text carefully The <I> tag was in the wrong place. It needs to be nested within the <B> tag, so as not to affect the `it is' part of the sentence.

14093 HTML fragment: <I> It is <B> very </I></B> important to read this text carefully. Appearance required: It is very important to read this text carefully. The italic tag <I> is placed too early with this fragment, "It is" would also appear in italic, the tag should be placed after the word "is". Correct version: It is <B><I>very</I></B>important to read this text carefully.

14207 It is <I><B>very</B></I>important to read this text carefully. The word very needs to be in bold and italic so therefore either the B tag or I tag needs to be nested and both tags complted with the /.

15613 The original HTML has the beginning Italic tag in the wrong place, it should come before `very' not before `it is'. You must also close the inner tag before closing the outer tag and this was round the wrong way. It should have read: - It is <I><B>very</B></I> important to read this text carefully.

15670 It is very important to read this text carefully. In the original, the opening italics tag is positioned in the wrong location. Also, the closing bold and italics tags have been reversed. In this case it would not prevent the correct display of the text, but it is advisable to use correct embedding of tags by closing embedded tags before containing tags. In other HTML tags this is essential for correct function. I have also used the emphasis tag to produce italics as the italics tag is deprecated in the latest HTML standards.

16031 Correct HTML It is <I><B>very</I></B> important to read this text carefully. In this case, the italics tag <I> was in the wrong place.

16240 It is <B><I>very</I></B> important to read this text carefully. In the original HTML, the <I> tag had been placed at the start of the sentence. This meant that "it", "is" and "very" would all have appeared italicised with the word very in bold.

Question 11

8280 Things to do:

 Pack suitcase,
 Book taxi. To the fourth example I added two break tags in order to start a new line and to leave a blank line. I then removed the end break tag as these tags do not need a closing tag.

8755 Things to do:

 Pack suitcase,
 Book taxi.

9363 There are two approaches that would produce the required result: Option 1. Things to do:

 Pack suitcase,
 Book taxi Option 2. Things to do: Pack suitcase,
Book taxi The original HTML fails to provide a break at the end of the initial line. The results required could be produced by using either two breaks (as per option 1) or a paragraph tag (as per option 2). The
</br> tags in the original code will also produce too many lines between the "Pack suitcase," and "book taxi". This can be resolved by using only one
 tag. (I have used the
 notation to ensure that the tag is XHTML compliant.)

9515 NULL

9648 +----------+ | | Things to do:<P></P> | | | | Correct HTML: | Pack suitcase,
 | | | | | Book Taxi. | |----------+----------| | | The original HTML had the paragraph tags | | | missing after "Things to do:" meaning that | | | "Pack suitcase" followed "Things to do:" on | | | the same line. Although the </P> tag is not | | | essential, I have included it to aim for `good | | Problem with original | HTML'. | | HTML: | | | | Furthermore, after "Pack suitcase" in the | | | original HTML there was a break tag
 with | | | a closing break tag </BR>. A closing break tag | | | is not required and in my browser this | | | combination produced the same effect as | | | <P></P>. | +----------+

10123 Things to do: <P> Pack suitcase,
</BR> Book taxi. I had to insert <P> on the first line to inset a blank line between "Things to do: and Pack suitcase,"

10446 Answer= <P>Things to do:</P> Pack suitcase,
 Book Taxi. In the original text, the tag
 was used incorrectly. When using HTML to mark up our text, we need to remember that browsers ignore blank spaces, so to create a blank line we use the paragraph tag <P></P> (This is like hitting the enter key twice on your keyboard). Although the <P> tag doesn't need to have a closing tag, it is seen as good coding to do so. To create a line break we use the break tag
. This does not need a closing tag.

10902 Things to do: Pack suitcase,
 Pack Suitcase Book Taxi,
 is a tag that requires no closing tag. Its purpose is to create a line break. This forces the text to start on a new line, without leaving the normal blank line.

11016 Corrected html should read: <P>Things to do:

 Pack Suitcase,
Book Taxi</P> Reason for this is It is usual that text is encapsulated between <P> (indicating the beginning of text / paragraph) & </P> (indicating the end of text / paragraph) The original had no `<BR> ` after `things to do', without any tag no line space would appear. (The coding ignores the `layout', simply reading the coding, text and tags, ignoring all `typed' spaces and carriage returns). It does however require a second <BR> to return a new line, the first ends the text and goes to next line of display, the second will `break' and also start a new line (leaving a blank line). The original had ` <BR> </BR> `, after `Pack Suitcase'. This is incorrect; a break tag does not require a closing tag. A single `break' tag is sufficient to create new line / carriage return.

11073 The paragraph tag has not been used around the "Things to do:" text, this will allow a blank line to appear under this text. The break tag has been used but this is one of the few tags that do not require an end tag. <P>Things to do:</P> Pack suitcase,
Book taxi. This is the corrected text.

11396 Things to do: Pack suitcase,
</BR> Book taxi. This text would have to be re-written as follows to appear Things to do: Pack suitcase, Book taxi. Things to do: <P> pack suite case,
Book Taxi. In the original version the tags are in the wrong place. A paragraph tag should be placed after "Thing to do" statement to make the text skip a line. After pack a suite case a break tag should be used to make the text go to the next line. Break tags are single tags so </BR> is illegal and the browser will disregard the html instructions.

11491 Things to do: <P>Pack suitcase,
 Book taxi</P> The
 tag was in the wrong place also it does not have a closing tag. A <P> tag is required before "Pack suitcase" to ensure an empty line between it and "Things to do". The
 tag is needed to make "Book taxi" appear on a separate line and the closing </P> tag is not essential but is considered good practice.

11586 <H6> Things to do:</H6> <P> Pack suitcase
 Book taxi </P> * `Things to do' is not indicated as a heading. Therefore heading tags inserted <H6></H6> the number indicates the prominence of the heading * No paragraph break between `things to do' and ` pack suitcase' * After `Pack suitcase'
 an end break tag is inserted </BR> - there is no need for this end tag.

11757 The correct HTML should be Things to do:
 <P> Pack suitcase,
 Book taxi. </P> The break tag,
, is used to break off the text and to continue it on the next line. It doesn't have a closing tag. There is also white space between the colon after the word `do' and the word `pack' which must be marked by the tag <P>. A closing tag is not required for <P> although it is considered good HTML to include one.

12118 Things to do: Pack suitcase,
</BR> Book taxi. Things to do: Pack suitcase, Book taxi.

12802 Things to do:

 Pack suitcase,
 Book taxi.

13125 The HTML fragment, things to do: Pack Suitcase,
</BR>Book taxi.' is incorrect as the paragraph 'Things to do:' and following line of text, 'Pack suitcase,' would appear as one line of text as line breaks are ignored by browsers. A break tag,
 also does not require a closing tag (4). The corrected HTML fragment is as follows: <P>Things to do:</P>Pack suitcase,
Book taxi.

13581 Things to do: Pack suitcase,
</BR> Book taxi. This should appear as follows. Things to do: Pack suitcase, Book taxi. ANSWER: Things to do:
 Pack Suitcase,
 Book taxi. The
 (break) tag should always appear at the end of the line of text to be broken. The
 tag does not require closing.

14227 Things to do:

Pack suitcase,
Book taxi. In the original HTML there were no break tags between ?Things to do:? and ?Pack suitcase,? so they would have appeared on the same line. There is no closing break tag; the use of this would result in two breaks instead of one.

14284 Things to do: <P></P> Pack suitcase,
 Book taxi.
 is a tag that does not need an end tag. It will put the next word on the line below. To get a white line between Things to do: and Pack suitcase, we have to use <P></P>, which will give us the white line between the words.

14626 . Things to do:

 Pack suitcase,
 Book taxi. The original statement has some omissions and invalid tag </BR>. For break tag or new line tag (
) does not have a closing tag. Two new
 tags have been added after first line of text `Things to do:' to have blank a line in the output and invalid tag </BR> is removed.

14683 <P>Things to do:</P> Pack suitcase,
 Book taxi. The first thing is that the break tag does not have a closing tag. The second is that it requires the paragraph tags so that there is a blank line left between the lines.

Question 12

8604 More information can be found <A HREF="help.htm">here</A>. Name is for where the link is need to use HREF.

9250 More information can be found <a HREF="help.htm">here. The HTML in question actually set the destination for a link of that name. The revision above sets up a link, as was asked for, to go to another web page of the said name.

9725 More information can be found <A HREF = "help.htm">here</A> The <A name> tag is used to create the name of an anchor tag however in this example we need to use the <A HREF> tag to create a hyperlink around the word "here" and the name/location of this is identified by "help.htm".

9991 More information can be found <A HREF="help.htm">here</A> In the original HTML fragment the anchor tag has been used incorrectly. This should be changed to the correct hyperlink tag to achieve the desired result.

10200 More information can be found<A HREF="help.htm">here</A> The word "here" in the original HTML is to be link to another web page so the anchor tag should have been followed by the attribute HREF and not NAME as was done.

10257 More information can be found <A HREF="help.htm"> here </A> This should be appears as , More information can be found here.

10713 More information can be found here. The problem with this fragment, is that `<a name=... indicates that it is an anchor tag, making the word `here' the target. To link it with another web page the tag should contain <A HREF=... indicating that it is a link to another web page. More information can be found here.

10903 More information can be found <a href name= " help.htm " >here The word "here' is made into a hyperlink and points to a file named help.htm with the introduction of href. The letter `a' relates to Anchor and href is the attribute for the anchor

11055 More information can be found here. The required effect is to create a link to another web page. The <a name> is an anchor tag and normally would create a link with the name held between the speech marks. However, the <a name> as it stands in the line of code will not work as an anchor because it is, in its own right, malformed. To achieve the intended link, the code should read: More information can be found here.

11663 Original HTML: More information can be found here. Corrected HTML: More information can be found <a HREF = "help.htm">here. The <a> tag defines a hyperlink and is used to indicate that the following text is either a target or a link to a target. An example of a target might be the URI of a website and clicking on the link that is associated with the target would direct your browser to that URI. In the above HTML code, the <a> tag is being used to define a target, where instead it should actually be a link. Links are defined by inserting HREF = "link" into the <a> tag and targets are defined by using the inserting name = "target" into the <a> tag.

11948 The corrected HTML: <A HREF= "help.htm">here</A> The wrong kind of link tag was used on the original, as it was a tag to produce an `anchor' with a name, rather than a hyperlink to another web page. Also, the tag had been written in lowercase instead of uppercase, therefore the browser software did not recognise the tag, thus it took no action.

12119 More information can be found here. This should appear as follows. More information can be found here. More information can be found here. (Note that this contains a link to another web page.)

12594 More information can be found <A HREF= "URI Address">here</A> The wrong anchor has been used to insert a link to a web page. To link a web page you need a hyperlink indicated by HREF in the tag. In the example the incorrect name has been chosen for a web page as the full URI (uniform resource indicator) would be necessary to make the link work. Eg "http.www.open.ac.uk/"

13696 Original HTML: More information can be found here The attribute `name' only takes you where a link is not to the actual link the attribute `HREF' takes to the link at the different web page.

14304 More information can be found <a HREF="help.htm">here The writer became confused between the hyperlink tag and the anchor tag, using both syntax. The word `name' in the original text is used in an anchor tag, whereas the text `HREF' is used in a hyperlink tag which is what the writer is trying to do, by using the word `here' to create a link to the html page titled `help.htm'.

14627 . More information can be found <a HREF = "help.htm"> here . The original statement <a name = "help.htm"> is an anchor tag which serve as destinations for links. The statement need to link to a different page "help.htm" and need to use a hyperlink tag as per the corrected statement above. When you click on the word here the help.htm page will be open.

14760 More information can be found here. This should appear as follows. More information can be found here. (Note that this contains a link to another web page.) Answer: <P> More information can be found <A HREF="help.htm">here</A> As you are linking to another page the hyperlink tag should by used to link to hekp.htm. The <A NAME= tag is a hyperlink to an AnchorName on the same page.

15406 More information can be found here. This should appear as follows. More information can be found here. (Note that this contains a link to another web page.) The "<A NAME" tag is used to set an anchor point in a document. The tag required here is the hyperlink to another page. The correct HTML is: More information can be found <A HREF="help.htm">here</A> HTML tags are not case sensitive so I didn't need to change them but I find upper case is clearer.

15729 The correct HTML for this sentence should be: More information can be found <A HREF = "help.htm">here</A> The original HTML is actually to create an anchor named `help.htm', which is nonsensical because an anchor would only need to be called `help' without the .htm file extension. To create an external link a hypertext reference tag is required. The HTML written on the assignment booklet in this question is in lower-case; where as in the rest of question 2 it is written in upper-case. Call me a pedant but this is worth pointing out because it goes against the convention set out on the M150 Essential HTML Guide!

15862 More information can be foundhere The href is needed as a attribute to the ,a. anchor tag to make this a hyperlink.

15938 More in formation can be found<A HREF="help.html">here</A> Would read: More information can be found here As in the original code the name attribute used with the anchor tag <A> is used to give the anchor a name but is invisible in the browser. To indicate a link tag then HREF is used to specify an element (word, image, etc) in a link to another page, in this case a web page in the same folder as the original linking page

15957 The correct HTML would be: More information can be found here. The attribute href= in the <a> tag indicates that this is a link to another document; in the original HTML the attribute name= was used, but this should be used in the target page, to create the link from the current page. The use of a file name only, with no more detailed path name or full URI specified, implies that the document help.htm is in the same folder as the current document.

Question 13

7731 The data becomes persistent as soon as it is saved to the hard disk or any other persistent storage media.
7845 The data becomes persistent the first time Victoria saves her document to the hard disk.
8586 The data became persistent when Victoria initially saved the report to the hard disk of her computer.
9612 The data becomes persistent data once it has been saved and is easily retrieved.
10410 The data becomes persistent when the data is written on to some form of media, such as a hard drive, floppy disk, Zip drive etc. Once the data is written it is then saved to the media, at this point it becomes persistent, because the data now persists.
11303 Victoria's document becomes persistent data the moment it is saved to the hard disk. It is persistent as it will be retained even after the computer is powered down.
12348 The point the report became persistent data was when it was saved to the hard drive.
17117 The data becomes persistent when it is saved.
17839 The data becomes persistent the moment Victoria saves her document to disk.
18333 Persistent data is data that exists after turning off the power to the computer or closing the program that was used to create it, therefore Victoria's report becomes persistent data when she saves the report to her hard disk. Part
18409 The data in Victoria's report becomes persistent as soon as she saves it to her computer's hard disk.
18561 ata becomes persistent once it has been saved, either to the hard drive of a computer, or to a portable storage device, such as CD-ROM or a floppy disk.
19169 NULL
19720 The point at which the data becomes persistent is when a copy of it is saved to the hard drive. Additionally, a new address is placed in the VTOC (Volume table of contents) which contains information regarding the cylinder, sector, and surface values.
19891 For data to be termed persistent it needs to be present after the computer has been switched off. The only way for this to be achieved is to save the data onto either the hard disk or another disk. The data in question therefore becomes persistent once Victoria has saved it to her hard disk.
20176 the data becomes persistent when the data is saved to the hard disk, since the data will exist after the application has been closed or the computer has been switched off.
20689 The data became persistent when the report had become fully saved.
21487 The data becomes persistent when she instructs the software to save the document to her hard disk
21981 The data becomes persistent the first time that Victoria saves her report.
22912 It becomes persistent when it has been saved to the hard drive.
24584 The data becomes persistent when the document is first saved to the hard disk.
24812 Victoria's report will become persistent data as soon as she saves it to the hard disk on her computer. From then on it will continue to exist as a document even when the creating application is closed and the computer is turned off.
25116 Immediately after saving the data to a suitable storage medium such as a hard disk.
25705 The data becomes persistent when the data is saved to a suitable storage medium i.e. the computer's hard disk.
26237 The contents of the data becomes persistent the first time it is saved.
26731 The data becomes persistent when it is saved on a hard disk or other storage mediums and the application is closed and the computer is turned off. It could also be considered persistent when printed out (hardcopy).
27187 Data becomes persistent as soon as it is saved, either on to the hard drive of the computer or on to a removable storage medium such as a CD. Persistent data exists after the application used to create it has been exited or after the computer has been switched off.
28042 The data becomes persistent when it has been saved on a suitable storage medium such as the hard drive.
28175 The contents of Victoria's document become persistent the first time she saves it.
29410 The data becomes persistent as soon as she saves the document to the hard disk.
30341 The data becomes persistent when saved to the hard disk on her computer.
30379 The point at which Victoria's report becomes persistent data is when she saves it to her hard disk on her computer. This means that after she closes down the application she created the report on and has shut down the computer, she can then retrieve the data at a later date because it has been stored on a suitable storage medium, i.e. her hard disk.
30531 The data will become persistent when the document is saved to the hard disk.
31310 The data becomes persistent once it has been saved to a means of persistent storage media i.e. the hard drive of a computer, a floppy disc or CD Rom.
31880 The report becomes persistent as soon as it has been saved.
32241 The data becomes persistent when it is written to permanent media which will hold the data when power is removed.
34445 Persistent data is data which continues to exist after the application that created or modified it finishes and / or after the computer that stores it is switched off. In this case the first time Victoria saved a copy of her report onto her hard disk it became persistent data. When she switched off the computer and switched back on later it would still be possible to see the file.
34502 It becomes persistent once it has been saved to the hard disk.
35034 The data becomes persistent when Victoria has saved it to the hard disk. At this point it will be available to access even if the application and computer have been closed and re-opened
35395 The data in her report became persistent data when she first saved it to her computers hard disk.
35547 The data becomes persistent when it is saved to the hard disk for the first time.
35661 Persistent data is that which is kept even `after closing down the application that created them or after switching off your computer' (Unit 5 p6). Once the document is save to the hard disk, the data contained is persistent.
36687 The data (a report in this case) needs to be saved on the computer's hard disc drive, or perhaps a floppy disc in order to be `persistent'.
36763 The data becomes persistent data when she saves the final version to the hard disk. The initial report will be replaced on the hard disk by the revised version when she saves.
37276 Considering the contents of the report as data, this data becomes persistent when it is `saved` to the storage medium known as the hard disk.
37865 The data becomes persistent at the point that the report is first saved on the hard disk.

Question 14

7561 The vtoc entry for the file is erased, though the file still exists until the data on the hard-drive is physically written over. Once the data is written over, which may happen when the second version is saved, the file is lost beyond recovery. Some word-processors, however, have a feature called version-tracking, which allows earlier versions of documents to be recovered by saving newer versions alongside the old versions transparently. In this case, the old version may still exist.

7637 Victoria's first saved file. When a file is saved onto the computer, the operating system ensures that the file does not overwrite any existing data that is already stored in the same volume. The OS then searches for suitable free space in which to house the document. If the document can fit on one ?block? i.e. if it is usually less than 0.5kb then it is stored in the easiest to find space available. If it is larger than 0.5kb it will search for a space big enough to house the ?blocks?. Victoria's first saved file will be replaced with the recently saved file if the recently saved file fits into the space occupied by the older saved version. This will then overwrite the original saved document. The Volume table of contents will show the new date for file size, file modified and date. If the recently saved document does not fit into the space saved for the original document, the system will place it in a space suitable and the starting address in the hierarchy will change accordingly. Victoria's saved document will then go into a hierarchy (absolute address) of files e.g. if she saves it in C:AssignmentsReports?name of doc? She can then access it by double clicking on her C drive then the folder Assignments-then the folder Reports and inside that will be her document art the lowest level of the hierarchy. Of course short cuts to finding this document are there but that is the path of the hierarchy to where the document lies.

7713 The first saved version of the file will be lost as the new updated version will either overwrite or be relocated, depending on the size of the new updated document (the criteria for this decision is, if it fits in the same volume space). Once the document has been updated or moved the VTOC (Volume table of contents) is updated, if required. Information that needs updating may be Date time stamp, File size and location.

8074 The first saved version will be a named file and this named file will be stored in a folder named by Victoria.

8511 The first saved version will be deleted or overwritten.

8625 Because Victoria saved the second version of her document under the same name as her first version, the first version will be overwritten and therefore cease to exist on the hard disk leaving only the second version.

9727 The first saved version of the document is overwritten by the second and subsequent versions of the document providing the new version of the document fits into the same space, otherwise the operating system finds a larger space.

10012 The first saved version no longer exists as such as it was revised and then saved with the same document name. It does not therefore exist in its original format.

10164 Assuming the file path of the final version is also the same as the first version the second version will replace the first. It is likely that the first version will have been overwritten on the hard disk although it is possible (but unlikely) that the first version is still recoverable using specialised utilities and depends on the actions of the operating system and the filing system it uses.

10696 If the second version of the document fits onto the same disk space as the first, then the first document will be overwritten by the second version. However, if the second version is larger, the first version will remain on the disk until overwritten in the future, when the space is needed.

10905 When the second version is stored it effectively overwrites the original version.

11114 The first saved version of the document will be overwritten by any subsequent versions given the same name and saved to the same folder.

11874 It was overwritten by the final saved version.

12577 The first version of the report will have been marked for deletion and will no longer have a directory entry. The space it was taking up on the hard drive will be overwritten with the new data.

13109 . The first saved version is stored to a storage medium (hard disk). She will have to name the report and store it to the root directory in a folder. The VTOC will check that it does not overwrite any saved data on the hard disk.

13603 The revised document is written over the top of the original document, replacing it with a new version.

13983 The first saved version of the document is overwritten when Victoria saves her final version.

14211 The first saved version of the document will be overwritten by the new version.

14800 This depends on the operating system. It is likely that the VTOC for the original version will be changed to point to the new version saved elsewhere on the hard drive. The original version would then remain on the hard drive but be inaccessible by normal means. The other outcome is that the VTOC is changed and the new version saved exactly where the original version was thus destroying it.

15902 The first saved document will be automatically overwritten by a subsequent document with the same name, unless the updated document is too large to fit in the space occupied by the original - in the latter case the original document will be inaccessible as it's file-name has been taken by the replacement.

15978 When the second version of the report is saved, it is saved over the top of the first version and so the first version is lost.

16928 The first version gets overwritten with the revised one.

17270 After the revisions have been made and the document is saved again, the first saved version will be replaced by the final version. If there is sufficient space where the original version was saved then the new version will replace the old and the Volume Table of Contents (VTOC) that records the storage details of the document will be updated with the new details i.e. the new size, date and time last modified. If the new document occupies a larger space than the original then there may not be sufficient space at the original location. In that case, the operating system will try and find a suitably sized location to store the new version. The report's starting address in the VTOC may need updating in this case^2.

17612 The first saved version of the document is overwritten the next time the document is saved, since it has the same name.

17783 The first saved document is overwritten when she edits and saves the final version with the same document name.

18087 When a document is edited and then resaved the new version will either overwrite the old version on the storage media if the size of the new document still fits into the space (number of disk sectors) where the original was saved. Or if the document has increased in size the new version will be saved to a new space on the media, at which point the `Volume Table Of Contents'/`VTOC' which keeps a list of the addresses of files on the hard disk (hard disk cylinder number, surface number and sector number) will be updated for the entry detailing the new starting address of the file.

Question 15

7467 Until a document is saved to the hard drive, it only exists in the PCs volatile RAM (Random Access Memory that is, it only exists when power is on. Should the power/PC fail all the work still in the memory - Victoria's report for instance 17 will disappear forever. However to the hard drive as for example Victorias_report.doc then should the power[on the fail, it will upon restarting open the report using the thus saving Victoria a lot of heartache

7695 One improvement to Victoria's work technique would be save the documents that she was working on more frequently, while they were being worked on. If the computer were to crash while the document was being worked on, all of the work would be lost, if the incomplete document had been saved, then only the work done since the last saved version was created would have been lost.

7961 A suggested improvement for Victoria's working practice would be for her to make back-up copies of her work onto other persistent storage mediums, other than the hard disk of her computer. This is because, if the computer disk crashes, the magnetic pattern will be destroyed so the data stored can no longer be retrieved and will be permanently lost. She will therefore have a second exact copy, which can be transferred to another computer, which has a suitable reading device, for retrieval. I would also suggest that Victoria needs to be aware of the location she is storing her document in. If she saves a document with an identical name, but in a different folder, the data will be saved successfully for retrieval, however, if she is not aware of where she is saving her document, an original document with the same name, in the same folder, could be overwritten accidentally.

8227 It would be good practice to save the report at regular intervals, not wait until it has been finished. This will prevent the loss of all but small amounts of data should the computer fail, power fail or the application fail.

8531 To improve Victoria's work practice she should save her report as she is working on it, this could be every 10 minutes or so. It may be possible for her to have the application that she is working on auto-save the work. In this way should she have a power failure, or her computer crashes, she will only lose the last 10 minutes of work. She should also consider having a second back up version of her work on another storage medium.

8645 Victoria's work practice could be improved by saving the document at regular intervals. This will ensure the document is not lost on the system in the event of the program or operating system crashing.

9158 Victoria should have saved either her first or second version of the document with a different name. This would have prevented the loss of the data of the first version and would have made the first version data persistent.

10013 An improvement in Victoria's work practice would be to save the original report with "V1" (to signify version 1) at the end of the document name. When Victoria makes amendments to the first version she should then save the report with "V2" (to signify version 2). In this way she would have a historic record for each version saved in a logical order which would allow her to quickly and easily find any version of the report when necessary.

10070 Victoria waited until she had finished the document to save it, it is good working practice to save a document at regular intervals so that if the computer crashes the whole document does not have to be re-written. She should also name the second document the same as the first but denote it with a date, For example: first document is called, document180404 Second document is called, document200404 this way both copies of the document are kept.

10184 One improvement to Victoria's work practice would be to save her work at regular intervals. The first time she saved the report was when she had finished it, this could be a problem if her computer had crashed or there was a power cut. Because she hadn't saved her work she would have lost it all.

10792 By using version control (entering v0.1 etc) the document name changes slightly allowing all versions to be saved in the same place. This way she will have a running amendment history of all her work to date

11058 After making the changes to her original document, Victoria would be best advised to save the new version with a new name, modifying it slightly, possibly by simply adding a version number to it. E.G. if the original document was called `Employment Report March 2004' it could become `Employment Report March 2004 V1'. Employing such a system would mean that should it be necessary to revert back to a previous version it is available for retrieval, on the storage medium without any further work or effort.

11229 Victoria should give her final copy a different name, this would enable her to double check her first draft

11666 It would be wise to save the updated document with a new name each time. The reason I would do this is because once the document has been updated and overwritten (and the application closes) there is no way to get the original back. I would personally save each version of the document with a new name in a temporary folder, appending the date or version number to the filename of each one until the document is completed. When the document is completed I would back up the final document to CD-R or Floppy Disk and delete the temporary folder to clean up the hard disk.

11970 Victoria's work practice could be improved by also saving the document to a removable data storage system e.g. a floppy disc. This would be a more full proof practice just in case something such as the document on the hard drive becoming corrupted or the computers hard drive fails. All would not be lost if a back up copy of the documents was created on removable storage media the document would still be available.

12350 Victoria could have named her final report a different name when saving so she could have had the original as a hard copy for future reference.

12901 Whenever Victoria saves she should save with a different document name. That way if there is a problem, she accidentally saves a bad copy of the document or simply wishes to have a look at a slightly older version of the document she has it.

12939 Victoria should save her revised document(s) with another name, e.g. "report_document_2". This would ensure she still has the original report were she to perhaps make a mistake with the revised version, or decide the original were better etc.

13015 She should set her word-processor to automatically produce a backup copy of the original document. A backup copy would allow her to go back to the original copy of the report if necessary. The document would typically have the same file name but with the file extension .bak

13300) An improvement to Victoria's practice would be to have a sub-folder. A subfolder can have for example: Draft report, Revised report, this then will give Victoria different category's within a folder. It would then make it easier for Victoria to remember where to find specific files. It would help a great deal if she uses different file names . For example : other names, adding a digit on the end, etc.

13414 Suggest an improvement in Victoria's work practice, giving a reason for your answer. [2] If Victoria saved her data to another filename (Ver X on the end for example), she could revert back to an earlier version if her machine crashes or she makes a mistake in her work.

Question 16

7563 The Zip-drive, the CD-Rom.
7753 Two examples of persistent storage medium are 1.44MB Floppy drives and Magnetic Tape drives.
8095 Two other forms of persistent storage media are the CD -ROM and printed material.
8475 Two examples of persistent storage media other than the hard disk are: 1. Compact disk Floppy disk
8741 CD, DVD (Zip disk, floppy disk etc.)
8855 Floppy disks and Zip disks are two examples of persistent storage media.
9577 1- floppy disk. 2- optical disc
10014 Two other examples of persistent storage media other than hard disk are: 1. CD's (compact disks)
10470 Zip drive tapes and Read/Write CD's
11819 Two examples of persistent storage media other than hard disk are Zip Drive or CD-ROM.
12807 Other examples of persistent storage media other than the hard disc include: floppy disks and optical discs, such as CDs (compact disc) or DVDs (digital versatile disc).
12978 Two further examples of persistent storage media other than a hard disk could be a CD RW or a magnetic tape.
13073 Two examples of persistent storage media are: * Other magnetic media such as floppy disks. * Optical media such as CD-ROMs.
13985 Two other examples of persistent data storage are Optical Disks and magnetic Tape Drives.
15296 Examples of persistent storage media include recordable optical discs, i.e. CD-RW and DVD-RW, and Zip disks.
15714 NULL
15885 two examples of persistent storage media are magnetic tape and floppy drive.
15961 Two other examples of persistent storage media are: optical disks (CDs and DVDs) and magnetic tape.
16303 CD-R, Databases.
16740 . CD Roms (Compact Disk Read only memory) Floppy disks (1.44 Mb)
17215 Two examples of persistent storage medium are DVD and floppy disk.
17899 * CD ROM. * Floppy Disk.
18127 Two types of persistent storage media (other than the hard disk) would be a 3.5" Floppy disk, or optical disks (CD's or DVD's) for small documents or small amounts of data. For high capacity storage such as system back-up, a magnetic tape would be possibly more suitable, especially if the data is not necessarily needed to be accessed immediately or easily.
18222 Persistent Storage Media is one that will store data even when the power has been turned off to the computer. Other than the computer's Hard Disk examples are ZIP Drives, these are removable drives that can hold up to 250 MB of data. Optical Discs such as DVD/CD ROM are another example of removable persistent storage media, they use lasers to read the contents, capacity of a CD ROM is 650 MB while a DVD can hold up to 4.7 GB of data. There is also a double-sided DVD that can hold 9 GB but must be turned over half way through, a standard DVD is capable of holding two hours of full motion video.
18469 Give two examples of persistent storage media other than the hard disk. Zip Drive, CD R/W
19476 A 1.44M floppy disc or a 650M CD-ROM.
19875 NULL
20027 Other media which allow persistent Data to be maintained writable CD/DVD's and data tape media. (Dat Tapes).
20312 Two examples of persistent media storage would be a floppy disk and a memory card.
20369 Two examples of persistent media would be a floppy disk, or an optical disk such as a CD-Rom or DVD-Rom.
20559 Magnetic tapes used for backup and CD-ROM (Compact Disc - Read Only Memory) are two other forms of persistent storage media.
20825 Two examples of persistent storage other than the hard drive would be 1, Tape Drive (Magnetic Tape) 2, CD/DVD rewritable (Optical Disk)
21015 Other examples of persistent storage media are : - Compact Disc - Magnetic tape
21034 Floppy Disk 2. Rewritable Compact Disc
21300 Two examples of persistent media storage are: * CD ROM * Floppy disk
22079 Other then the hard disk, a zip disk or CD could be used for storage of media.
22307 Two examples of persistent storage media other than the hard disk are * Zip drive for removable hard disks. * Floppy disks.
22421 One example of a persistent storage medium is a Zip drive, another is a CD.
22782 Two examples of persistent storage media other than the hard disk are Magnetic tape and Optical Disc (CD or DVD).
23371 You could back up really important information from your hard drive on to CD ROMs. Magnetic tape could also be used as they hold more information.
23504 Two alternatives to the hard drive storage are magnetic floppy disks (either a three and a quarter inch disks, which holds 1.44 megabytes of data or a zip disk which can hold a lot more information, up to 250 megabytes. i) d) continued Secondly there are optical disks, technology which uses the optical properties of the surface of the disk. These can hold a lot more data in the region of 800 megabytes for cd's. These disks are called compact disks and DVD disks.
23580 Answer: Two examples of persistent storage media apart from a hard disk are:- CDR/CDRW media the file could have been burned to such a media in which case it is persistent. Another example would be something like a USB Minidrive. Its small enough you can carry it around on a key ring; and it retains data even when disconnected from the computer.
24169 Magnetic Tape & Optical discs.
24321 CD/R, ZIP
24492 Persistent storage Media other than a hard disc includes Compact Disc/ DVD and floppy discs.
24796 A floppy disk and a compact disc.
25594 Two examples of persistent storage media are a Writable Compact Disc and a Zip Drive.
25898 Another two examples of persistent storage media are, * CD Rom * DVD
25955 The zip drive and CD-ROM would be examples of persistent storage media.
26506 The CD-ROM and the computer's ROM(Read Only Memory).

Question 17

7944 The data being transmitted is personal as it may contain data that covers both facts and opinions about a person. Therefore, anyone transmitting such data under the Data Protection Act has to have a good and justifiable cause, such as the person or persons being involved in illegal activities. The transmitted data about this individual or individuals would also have to be accurate.

8001 Within UK law Victoria may send the unedited report if she takes steps to make sure the data is protected from being processed by unauthorised persons. The data should only be sent if the recipient is inside the European Economic Area, unless the country it is sent to ensures a level of protection for the data subject's rights and freedoms, to an adequate level.

8951 She may send the unedited report but she must gain permission from the Data Protection Commissioner for using data and must not pass off some of the data which she must have collected from the company's data (plagiarise). She must adhere to the Data Protection Act, so she should have a good, justifiable reason to send the unedited report with personal details of the data's subjects and it should be accurate and not malicious.

9844 Personal data is protected under UK law. Therefore if the report were to contain personal information about an individual it should not be shared without the owners prior consent.

9920 Victoria would need to comply with the eight principles of Data Protection under UK Law. Assuming she has justifiable reason to transmit the report via email then it is perfectly legal for her to do so.

10053 Under UK law, Victoria can send the unedited report to John so long as she is legally entitled to disclose any personal data it contains to John, and providing the report does not infringe someone else's copyright. It must also be legal for the contents of the report to be transmitted over a public network.

10737 If Victoria sends the unedited report to John, within the guidelines of the UK law she should have good cause to do so and should abide by the 8 principles of good practise which state how the data should be used and treated. Some of the main guidelines are the data should be accurate, relevant, protected, secure and not kept longer than necessary. (ii)(

10870 She must have justifiable cause for transmitting the report to John, ensuring that it is accurate and that all steps must be made to protect the data from unauthorised access. Any such transmissions may only be kept for as long as is necessary and for no other purpose other than that which it was originally intended. (ii)

11136 Victoria may only send the report if the data on the individuals contained in it, is both true and is not of an offensive nature or could be distressing to the individuals concerned. Also, she must get the individuals consent to be able to disclose the personal data contained in the report.

11193 To send on the report containing data on identifiable individuals, one should ensure that the information is to limited purposes, fairly and lawfully processed, adequate relevant and not excessive, accurate, not kept longer then than necessary, secure and non transferable to countries without adequate protection.

11573 The unedited report contains no identifiable individuals so no personal date will be contained within the Email so there is no notion of privacy and as long as the report Email complies with the eight enforceable principles of data protection under UK law. Emphasis on one main principle for the report processed in accordance with the data subject's rights.

11706 Within UK law she may send the details so long as the information is accurate and relevant to the person receiving it. She must also ensure that the information is secure.

11896 Victoria must have a valid reason to send the document to John. There must also be adequate security in place to prevent un-authorised viewing, e.g. firewall, password protection, encryption etc.

12333 Under UK Law, Victoria may send the unedited report to John only if she has a good and justifiable cause for doing so.

12998 Victoria may send her unedited report provided it complies with the Data Protection Act of 1998 and does not infringe the Copyright laws. This means that the report should contain her own work and not contain data about an individual which John should not be allowed to access.

13055 NULL

13093 Victoria may send the email of her report to a colleague as long as it is accurate and processed in accordance with the subject's rights. It must also be relevant to send the information to her colleague otherwise the UK data protection law is contravened.

14765 Victoria must consider the principles of the UK's Data Protection Act of 1998 (and recent amendments/updates to the Act) when forwarding her report as if it contains facts or opinions about an individual then the individual has a right to access the content. She must also abide by the principles of good practice laid down in the act, in that there should be a good and justifiable cause for including the data and it should be accurate data, further both her and John should not keep the data for longer than necessary and ensure the data is protected from unauthorised access.

15126 She may send the report if it is done so securely (encrypted possibly). Also John may only keep the data for as long as he requires it (grey area??)

15430 The data transmitted by Victoria must comply with the eight enforceable principles of good practice as it contains personal data (unit 5 page 56). Has the data been legally collected; personal data collection requires registration under the Data Protection Act 1998.

15829 Victoria must take into account the principles of the Data Protection Act 1998 when sending data which identifies individuals to John. There are eight principles of this Act; they are that data must be: * Fairly and lawfully processed * Processed for limited purposes * Adequate, relevant and not excessive * Accurate * Not kept longer than necessary * Processed in accordance with the data subject's rights * Secure * Not transferred to countries without adequate protection

15943 Providing the company's network is secure, and the report is adequate, relevant and not excessive and has been processed in accordance with the data subject rights then it is acceptable for confidential report to be mailed via the company network.

16000 She should make sure that the data is secure, probably by encrypting it before emailing. She should also make sure that all data is accurate and processed in accordance with the data subjects rights.

16513 Firstly the TMA question does not state what the report and the data contained therein is going to be used for, if the data contained personal information on each of the individuals was the email encrypted to ensure the security of the document from unauthorized access? .Secondly is the data accurate?, If not she could be liable for prosecution as the data could be defamatory and damaging to the individuals reputation.

16627 She must not keep the copy of the report she has, longer than necessary. Because when processing any variety of data the probability of transmitting them to others is high.

Question 18

7660 E-mail sends messages as encoded ASCII text. The report can be encoded as a sequence of bits the bits can be grouped into bytes and as such can be sent at the end of the message.

7698 The user name is "john", the domain name is "Birmingham.office.xy.uk", and the top-level domain is "uk".

8249 MIME (Multipurpose Internet Mail Extensions). A standard that converts an attachment to a series of alphabetic characters that are added to the end of the message.

9104 The property of internet email which allows the attachments of documents is MIME (Multipurpose Internet Mail Extensions). This property allows for the attached document to be converted to ASCII code for sending with a standardised encoding method which allows the recipient to decode the attached document.

10643 The property of email that allows attachments to be sent is a standard called MIME (Multipurpose Internet Mail Extensions).

10947 Attachments can be sent be email because they are generally encoded using the MIME standard (Multipurpose Internet Mail Extensions). Most computer systems can implement this protocol and a headers in the main email document are used to indicate that the attachment has been encoded to the MIME standard and to instruct the receiving machine how to process various types of data.

11080 The internet standard for encoding mail attachments is MIME (Multipurpose Internet Mail Extensions). When an attachment is given to an e-mail it is encoded using this standard originally published in 1982. Of course the key point is that both the sender and receiver of the e-mail both use the same protocol.

11441 MIME (Multipurpose Internet Mail Extensions) allows | | 3) ii) b) | files, such as Victoria's report, to be encoded as ASCII | | | characters and included in an email as an attachment. | |-----------+----------|

11555 Explain briefly the property of Internet email that allows the contents of the report to be sent as an attachment rather than as text in the body of the email message. The property of internet email that allows the content of the report to be sent as an attachment rather than as text in the body of the email message is that although only text can be sent as email transmission, an attached file can be encoded as a series of alphabetic characters which are then joined to the end of the message. All electronic documents are encoded into bit sequences, which are then grouped into bytes, which are then interpreted as text characters. In this way an attachment can be converted and sent by email.

11612 The contents of the report can be sent as an attachment by being encoded into characters which are then added on to the end of the message. There is an internet standard for this encoding called MIME, which is required as receiver of the report needs to be able to decode the attachment in order to be able to read it.

11878 The attached file is encoded as a series of alphabetic characters conforming to the MIME standard, the receiving email client can then see the email has an attachment and can decode it.

12600 The property which allows contents of a report to be sent as an email attachment is MIME - Multipurpose Internal Mail Extensions. This is the standard for encoding mail attachments, rather than as text in the body of the email message.

12733 "Multipurpose Internet Mail Extensions", (MIME), represents the most common "Protocol" used in encoding documents to be sent as attachments to email messages.

13398 The property of email that makes attachments possible is done by encoding the file to be attached. This achieved by converting the file into alphabetic characters, which can then be represented as bytes of ASCII code, which can then be transmitted.

13930 File attachments can be sent with an E-Mail by encoding the file and including in the header file some metadata (data about data) about the file, for instance, whether it conforms to the MIME (Multipurpose Internet Mail Extensions) (Block 1, Unit 5 Section 3.4) standard, this is required by the receiving mail client so that it knows there is a file to be decoded and the format the file has been encoded in. This is another reason why there must be computer standards, to make the information exchangeable.

14234 The report can be sent as an attachment by being converted into a code suitable for e-mail transmission (e.g. ASCII) and then added to the end of the message.

14367 You can send an attachment to an Email, this is achieved by encoding the file as a series of alphabetic characters and adding it to your Email. There are a number of standards for transmitting attachments the internet standard is MIME (Multipurpose Internet Mail Extensions).

14405 The property of internet email that allows the contents of the report to be sent as an attachment rather than as text in the body of the email is its ability to encode the attachment as a series of alphabetic characters and append them to the end of the message. These characters can then be converted to ASCII code suitable for email transmission. One of several encoding standards is used, the internet standard for encoding email attachments being MIME (Multipurpose Internet Mail Extensions). Originally published in 1982, MIME has undergone numerous revisions notably the inclusion of formats to handle pictures and the use of non-ASCII character sets. When the email message is sent, details of the encoding system are included in several additional email headers (lines of information giving details about the transmission such as sender, subject and return address), including the version number of MIME, content type such as text/plain, image/jpeg, or video/mpeg, and the character set code. These additional headers are key to the receiver being able to unpack the attachment. Both sender and receiver must be able to implement the same protocol.

14576 As email is sent as ASCII text, it is possible to attach a file by converting it into a series of ASCII characters which are added to the end of the message. The client mail application can then decode this portion of the message and open the attachment.

15526 he report is able to be sent as an attachment as long as it has been encoded using the MIME (multipurpose internet mail extensions) standard. The notification of this protocol is usually indicated by use of metadata in the email header e.g. (taken from a recent e mail I sent containing 2 attachments 1 text and 1) Mime-Version: 1.0 Content-Type: multipart/mixed; boundary="----=_NextPart_000_71e1_3e7_18c7" This is not something that an individual has to select but is a standard and protocol that has been built into the programme being used.

15583 MIME (Multipurpose Internet Mail Extensions). This is the standard that enables the email attachments to be encoded and decoded. It includes methods for transferring non text extensions i.e. non ASCII characters, pictures etc.

15697 MIME is a standard which allows attachments to be transmitted by e-mail. It works by converting the binary code to ASCII text which is then decoded back to binary on arrival. Both transmitter and recipient must have the MIME protocol to encode and decode the attachment in this way.

16210 property of the internet that allows the contents of a report sent as an attachment is called a link, this can be a file from another source containing data that can be password protected known only to the sender and the receiver.

Question 19
7452) John is the username. Birmingham.office .xy.uk is the name of the domain. uk is the top-level domain.
7547 User Name = John Domain name = office.xy.uk top-level doman = uk
7585 [1]John@Birmingham.office.xy.uk User Name: John Domain name: Birmingham.office.xy.uk Top Level Domain: .xy.uk
7680 John's email address is [1]John@Birmingham.office.xy.uk Which parts of the address are: the user name, the name of the domain, the top-level domain? the user name : John the name of the domain : Birmingham.office.xy.uk the top-level domain : uk
7851 The user name is John. The name of the domain is Birmingham.office The top-level domain is xy.uk
8326 [1]John@Birmingham.office.xy.uk The user name = John Name of the Domain = Birmingham. Office. Top level Domain = xy.uk
8649 In the email address John@Birmingham.office.xy.uk a. John is the username. b. Birmingham.office.xy.uk is the domain. c. uk is the top level domain.
9143 The user name of the e-mail address is john. The name of the domain is Birmingham.office.xy.uk. The top-level domain is uk.
9333 [1]John@Birmingham.office.xy.uk The user name is John The domain is Birmingham.office.xy.uk The top level domain is uk
9694 -user name=John -domain name=Birmingham.office.xy -top level domain=.uk
10036 The user name is john. The name of the domain is Birmingham.office.xy.uk. The name of the top-level domain is uk.
10378 NULL
12107 The user name is John, the name of the domain is Birmingham.office.xy.uk and the top level domain is .uk .
12468 John's email address is [1]John@Birmingham.office.xy.uk. The user name is: John The name of the domain: Birmingham.office.xy.uk The top-level domain: uk
12563 John's email address is broken into; user name - John (this is the part before the @ sign); the name of the domain - Birmingham.office.xy (this is the part after the @ sign); and the top-level domain - .uk (this is the country of origin).
13000 The Username is John. The name of the domain is Birmingham.office.xy.uk The top level domain is uk.
13038 John is the user name @Birmingham.office is the domain name .xy.uk is the top-level domain.
13532 John@Birmingham.office.xy.uk The user name is John. The name of the domain is office.xy.uk. The top-level domain is uk.
13817 [1]John@Birmingham.office.xy.uk `john' - the user name `birmingham.office.xy.uk - the name of the domain `uk' - the top-level domain
14919 Username: John Domain name: office.xy.uk Top-level domain: UK
14957 To convert decimal 1183 to hexadecimal I would use the same principle as used for converting to binary. Each value increases by the value of x16 as shown in the table. +---------------------------+ | 4096 | 256 | 16 | 1 | |------+-----+----+--------| | 0 | 4 | 9 | 15 (F) | +---------------------------+ 1183 divided by 4096 won't go so 0 goes in the 4096 column. 1183 divided by 256 equals 4 with 159 remaining. So 4 goes in the 256 column. 159 divided by 16 equals 9 with 15 remaining. So 9 goes in the 16 column. 15 divided by 1 equals 15 with none remaining. So 15 goes in the 1 column. Because hexadecimal uses the letters A to F for the numbers 10 to 15 then the letter F is used in the 1 column. Therefore the decimal number 1183 is Hex 0[x] 4 9 F.
15090 The user name is the part before the @ symbol in this case John The name of the domain is the part after the @ symbol in this case Birmingham.office.xy.uk The top-level domain is uk at the end of the email address
15584 John, office.xy.uk, uk
16192 The user name is "John" The name of the domain is "Birmingham.office.xy.uk" The top-level domain is .uk
16496 John's email address is John@Birmingham.office.xy.uk The user name is John The name of the domain is: Birmingham.Office.xy. The top level domain is: Uk
16648 Given the e-mail address: [1]john@birmingham.office.xy.uk, The user name is `John', the domain is `birmingham.office' and the top level domain is `xy.uk'.
17313 The username of the e-mail address is John. The name of the domain is Birmingham.office.xy, and the top level domain is .uk
18795 The user name is John The domain name is Birmingham.office.xy.uk The top-level domain is uk
20011 The username part of John's email address is John The Name of the domain part of John's email address is Birmingham.office The Top-level domain part of John's email address is xy.uk
20125 The email address John@Birmingham.office.xy.uk is made up of the following parts. The user name is John The computer name is Birmingham The name of the domain is office.xy.uk The top-level domain is xy.uk
20258 The user name is 'John'. The name of the domain is 'Birmingham.office.xy'. The top-level domain is 'uk'.
20277 The user name is `John' The name of the domain is `Birmingham.office.xy.uk The top-level domain is `uk'
20296 John's email address: "mailto:John@Birmingham.office.xy.uk" the following part is his user name: John name of the domain: Birmingham.office top level domain: sy.uk
20581 John@- username Birmingham.office.xy- the name of the domain uk- top level domain
20676 John@Birmingham.office.xy.uk The user name: John (part before the @ symbol) Name of the domain: Birmingham.office.xy Top-level domain: uk (It is the second type of top-level domain which is a two letter code that normally specifies a country, in this case United Kingdom).
20828 The text before the @ is the user name (john) The Domain is - birmingham.office.xy.uk The top level Domain is - uk
21379 The email [1]XXXXXXX@birmingham.office.xy.uk is made up of different parts which are explained below: the user name: XXXXXXX (this is the part before the @ symbol in the address) the name of the domain: birmingham (this is the part that comes after the @ symbol in the address) the top level domain: uk (to relate to a specific country)
21512 The user name is "John" The name of the domain is "Birmingham.office.xy.uk" The top-level domain is "uk"
21987 The parts of John's email address are as follows: user name - John Domain - Birmingham.office.xy.uk Top Level Domain - .uk
22405 The user name is John. The domain name is Birmingham.office.xy.uk and the top level domain is uk
22709 User name: John Domain name: Birmingham.office.xy.uk Top-level domain: uk
23127 User name John Domain name Birmingham.office.xy.uk Top-level domain uk
23203 The user name is John, the domain is Birmingham.office.xy.uk and the top-level domain is uk.
23830 The user name is: John The name of the domain is: Birmingham.office.xy.uk The top-level domain is: uk
24229 User name: John Domain name: Birmingham.office.xy Top-level domain: uk
24419 The username is John The name of the domain is Birmingham.office.xy.uk The top-level domain is uk
24704 The user name is John The name of the domain is Birmingham.office.xy The top-level domain is uk

Question 20

7510 When publicising a report on the Internet there are certain elements in the report that you must take into account. One is the Data Protection Act 1988 and anyone who processes personal data must comply with the eight principles. This also means that everyone has a right to privacy and you are not allowed to publicise information about an individual nor are you allowed to breach copyright laws. There are also ten computer ethics that you may need to take into consideration. If you collect information about people you must state how you use their information and their right to privacy. You also need to take into account people with visual difficulties and language barriers.

7776 Does she have the right to publish her report e.g. does her report contradict copyright laws or have a detrimental affect on the people named in the report rights to privacy. She should ensure the information gathered for the report was obtained in an open and hones manner. She should consider the content of the report, is it a fair representation of the information she gathered when compiling the report. Any images used are displayed with the owners consent and that she has permission to use any hyperlinks contained in the document and insures that the links are properly referenced.

8080 When Victoria publishes her report she needs to be aware of people's right for privacy and the requirements of the Data Protection Act of 1998. She needs good reasons for including data on specific individuals and must be able to justify making this information public. Any facts and opinions that she includes must be relevant, accurate and properly protected with adequate levels of security. When processing her report she must be fair and obey the laws of the country where her report is published. If she transmits the report to other countries then she is obliged to ensure that it is secure and protected.

8118 Purpose of the report, what is meant to be achieve by the site. * The auditory who determinates the medium and level of information contained. * Organization and clear links as a way of providing an effortless site. * Means for navigating back and forward. * Equipment likely to be use to access the information. * Good approach in the content structure, including good paragraphing, titles and sub-titles. * That the information does not incurred against data protection laws. * Provide a way of contact with the author. * The amount of information contained in one page.

8365 Victoria must ensure that she has the right (in this case permission) to publish information on identifiable individuals and / or Copyrighted material. Victoria should ensure that the web server she uses is kept secure and protected from unauthorised access, possibly using a password. The report should be published in line with the Data Protection Act. The data published in her report may be used by other individuals in un-ethical or illegal activities. She should consider imposing a copyright restriction in order to provide legal support if her report is re-published or reproduced without her permission.

8536 Victoria has to take in account when she makes her report, its contents and who will be reading it. She should bide by the principles of data protection under UK law and consider the laws of other countries where the report may be read. She should also work within the guidelines known as Computer Ethics, which may be accessed at [5]www.cpsr.org/program/ethics/cei.htm. This is a set of moral principles that guide acts as a citizen when using the computer. She should ensure that the work in her report is her own and where it is not, she should acknowledge the author.

8574 In preparing her report for publication Victoria should insure that any information or combination of facts that could identify a person is removed or that they are included with the persons consent. She should ensure that all the facts are true and accurate, that all sources are credited and permission to reproduce material has been obtained where necessary.

9049 Victoria needs to take into account who will have access to the published document, will it be on a internal website or will it be available to anyone with internet access? If the site is to be public she must ensure that no data referring to individuals is published, she may wish to replace real names with generic titles to ensure confidentiality. She also needs to ensure she is not infringing any copyright for the report that may exist and that she is also abiding by the same rules as she did when she sent the report to John.

9486 Victoria must ensure that when publishing her data it prescribes to the legislation of the Data Protection Laws, as she is the author of the document and ultimately responsible for its contents. She should take into account: * whether restrictions are placed on the data, * the data is pertinent to its relevance, that the sources and methods used in obtaining the material are lawful. * that it does not infringe copyright laws or contain material which could be in violation of national/international laws * That the data is both current and accurate. * Whether time restrictions are relevant 99 words

9543 When making her report public Victoria must consider the individuals that she has mentioned in the report and ensure that their rights under the Data Protection Act are not breached and that she does not breach the Act herself. The Computer Ethics set out by Computer Ethics Institute should be adhered to. As this is to be published on a website laws of other countries should be considered also. The report needs to be fair, accurate, lawful ,relevant and should not breach any copyright or security. Serious consideration should be made to whether this report needs to go public.

10075 she has to take in to consideration many ethical, legal and security issues regarding privacy, data rights and the laws that govern them.

10284 Under the data protection act Victoria must remove all references that identify a particular individual unless that individual provides their permission for their identity to be used. The data and conclusions in the report may be used as long as the identities or any particular individual are protected.

10360 Before making her report public, she must ensure that all the data concerning the individuals is accurate and up-to-date. She must also take precautions that the data is not personal and the individual's privacy is kept intact. For example no home telephone numbers, home addresses or e-mail addresses should be published. Also, she should contact the people concerned and ask for their permission to publish their details and ask whether they would prefer to remain anonymous in the report.

10474 She must decide if she has a right to the data. Is the data on the individuals personal or public? If personal whether she is infinging the rights of the people identified. Is it accurate? How long it is to be kept? Is the data relevant to the report and not excessive? Will the data be secure ie. will not be copied or viewed by unauthorised people? She must ensure the data is not kept longer than necessary. Can she be sure the data cannot be retrieved in countries which do not have adequate data protection laws.

10588 Victoria must take into account who the perspective audience is and how they might interpret or use the information. Under the 1985 data protection act, she must get permission from the individuals identifiable in the report. The content must not contain sexual, ethnic or racial discriminations and must be factual, accurate & verifiable. She should state that it is copyright and take personal responsibility for the publication. There are wider implications to be considered, such as the global nature of the internet and different local publication laws that might or might not be enforced. (iii)

11215 She should consider making the individuals non-identifiable on the website to begin with. Assuming this is not a realistic option, consideration should be made as to the accuracy of the information being published about people with regard to the laws of libel and slander. If the data includes the intellectual property of other people, such as quotations from their writings, copies of their paintings or

music, then getting their permission would avoid infringing copywrite law. The identifiable individuals may have their own website, obtaining their permission to link to it may be beneficial to both parties.
11614 In making the report public Victoria would have to protect the identities of all the individuals in the report. If she published the report as it stands it would not comply with the eight central principles. This is because once it is published on a website she will no longer have control over what the data is used for, who it is accessed by and how long the data is kept by the people who accessed it.
11671 Victoria needs to take into account which parts of the document are legally allowed to be published. Does the report contain information about the Company that should be seen by employees only? More importantly, does she have permission from these individuals to publish their details on the Internet? Publishing personal information would be illegal and in breach of "The Data Protection Act (1998)". Should the report be password locked so that it is only available to the intended audience? Finally, Victoria needs to be sure that the report does not contain copyrighted material written by someone from another source.
11728 Before making her report public Victoria needs to ensure that the personal data contained in her report is a true and accurate reflection. If possible it would be courteous to contact the individuals concerned to let them know that information about them is about to be published on the web. At the minimum Victoria should make sure that her report complies with the UK data protection act, though as the web has no border restriction she could find she falls foul of privacy legislation in other countries.
11766 Victoria must take into account two factors when making her report public on a website, one a legal requirement, the other one an ethical consideration. With limited exceptions, the data must have been fairly and lawfully processed. It must also be accurate and up to date, relevant and not excessive. In general terms, it must comply with the Data Protection Act, 1998. On the ethical and moralistic side, she should consider the ten principles listed by the Computer Ethics Institute (page 57, Unit 5 of M150). In particular, the report should not harm other people, bear false witness and ensure consideration and respect for others and their privacy. (iii)
12526 Some things that Victoria has to take into account, when she publishes her report are, that it doesn't infringe the copyright of any other person, violate the privacy of another person, use material which is designed or likely to cause annoyance, inconvenience or needless anxiety, facts concerning the individuals must be accurate and verifiable, their views or opinions must not portray them in a way that could damage their reputation, it does not include defamatory material.
12621 If jane is to publish personal data she should consider wether ther or not the data in question is protected by the data procetion act and if so she is prohibited from processing personal data unless a data controller is included in the register of data controllers under section 19 of the data protection act. There are some exemption which jane may feel apply to her which include such things as national security and miscellaneous exemtions. If janes purposes when publishing her site does not qualify for an exemtion under the data protection act she could always ask for the individuals consent to include their personal data on her website. This is especially true if the inclusion of data on her site is likley to cause harm or distress to the parties invloved as she could end up having to compensate the individuals involved.
13894 When Victoria prepares her report one of the first considerations is will the data, which applies to identifiable individuals, contravene the `Data Protection Act'. As this is principally about privacy she has to be clear the purpose justifies the use of the data, e.g. it was collected for this purpose only. It has to be accurate. If there is sensitive personal data this may not be publishable. Data subjects have a right to know why data is stored about them. These rights cannot be infringed. As it is a website it may transcend UK law and this is also to be considered.
14103 Victoria needs to make sure the content is factual and accurate to the best of her knowledge before publishing on the website. She needs to take personal responsibility for the content of the report so the company can't be held accountable for false information or inaccuracies. Depending on the content she may require the identifiable individuals permission before publishing the report, in certain circumstances such as a requirement for the report to be published for legal reasons or if the individuals belong to a political, religious or trade organisation that are publishing the report, permission is not required. (1)
15262 Victoria has to take into account the rights of the individuals she has identified in her report and ensure the data is accurate and conforms to UK legislation. She needs to consider restricting access to the report to authorised users only by using a password, and making the report read-only so that it's contents cannot be altered maliciously. Victoria needs to consider who owns the information about the individuals identified in the report and their rights to access the information. She will also have to consider the length of time the report will remain on the site, as this should not be kept longer than necessary.
16250) Victoria should take into account the nature of the data on the identifiable individuals in her report. The data should not provide anyone with a means of contacting the individuals concerned and nor should it disclose anything personal about the individuals. Permission should be sought from these individuals to publish the data beforehand.
16820 When making her report public Victoria must consider the both Data Protection Act and any ethical implications as publishing the report on the web would mean that the unedited report contents would no longer be secure, nor would it be restricted to countries with adequate protection. Regardless of whether she is acting independently or on behalf of her employers she would need to ensure that there would be no legal or ethical repercussions from the contents of the report as the data it contains may adversely affect the individuals identified in it.
17257 In no more than 100 words, explain what she has to take into account when making her report public. Since the report has data on identifiable individuals she has to make sure that she has the permission of the individuals to use any data she may wish to publish about them. Any copyrighted material used must have consent from the copyright owner. Should any links to other web pages be included, it is considered good manners to seek the author's permission, although this is not required by law. The report should not contain anything that may be considered offensive by others. If she is quoting from a source she should acknowledge the original source to avoid plagiarism.
17732 Victoria must comply with the Data Protection Act (assuming the website is in the UK), and respect people's privacy. Her report should not publish private data of identified individuals unless they have given informed consent. She should not violate copyright laws; she should not claim or abuse the work of others. Her report should be truthful, not misleading, and show no malice. She should respect the cultures and people in the countries of her readers. She should provide a means for people to respond to what she's written. She should do as she would be done by.
17884 Victoria will need to take into account the style of the document in web page format, for example hypertext could be included to allow users to make choices about the parts of the document they wish to access. However, more importantly, she will also need to consider the content as, if it contains information about identifiable individuals, making the report public would be in breach of the Data Protection Act. Thus, she would need to remove all personal information identifiable to individuals before making the report public in this way, possibly replacing it with generic and/or anonymous detail.

Question 21

7454 She should publish the Company address to give the Website authority and also a means of contact for users. An e-mail address should be given so that if any data within the report is erroneous those affected will be able to have it corrected. She should not add her personal phone number as this is too dangerous for her personal security.
7549 Company Address - It is probably appropriate for the company address to be published with the report as it gives a "public" place to contact the author of the report that will not the construed as too private. Personal Telephone Number - This is highly inappropriate for publishing. Personal details like this should never be published on a website for all to see, as just about anyone with a web browser could get that telephone number and infringe on the author's privacy. Email Address - There are two answers to this one, as it depends whether the email address is private or public (eg. Business or personal). A personal email address is not appropriate for the same reason that a persona telephone number would be, but on the other hand providing a work/business email is good way of collecting feedback on the report and giving a contact address for any other issues.
7853 Victoria should publish the company address in case anyone wishes to contact the company concerning the report. It does not say whether the email address is private or business. She should not publish a private one and she should only publish a business one if suitable firewalls are in place. Victoria should not publish her personal telephone number as she may receive malicious telephone calls.
7929 The company address should be published as a point of reference and to show the nature of the report. A personal telephone number would perhaps be too private to post on a website on the internet. An email address would be often posted with the report, but there is the possibility of it being spammed as a result once known. . All "processing" of personal data (includes collection, holding, retention, destruction and use of personal data) are governed by the Data Protection Act 1998. The Act applies to all personal data - whether they are held on a computer or similar automatic system or whether they are held as part of a manual file. Personal data is defined as information relating to an identifiable living individual and can be held in any format, electronic (including websites and e-mails), paper-based, photographic etc. from which the individual's information can be readily extracted.
8157 Victoria should only publish those details relevant to the company and the report. However without knowing the remit of the report, its audience and Victoria's interaction with its reception, it is difficult to qualify. However, publishing details of the company address and providing a contact email address will suffice, if the report is to attract to feedback. Victoria would be unwise to publish a direct telephone number, especially a personal one, since the report may fall into the public domain (wider than the reports intended audience).
8252 She should publish the company address to allow for any correspondence for people without email and also an email address for people to contact with regards to the site. There would be no need to publish a personal telephone number and this should not be included.
8765 Company Address - Yes as long as the company approves it, including this does give people a direct line to the company which they may not want people to have. Personal Telephone Number - No this type of information should never be included in a company website. Email address - Yes although if it is anticipated that there will be a large response it may be better to set up a dedicated mailbox.
8917 I think the Company name should be included on the website. Because she has a requirement to email a copy to a work colleague implies that it is a work document that relates to her company's business. As she has created the report and is to publish it on a website including a work email address would enable individuals to contact her professionally about the report. Including a personal telephone number would not bare any relevance here. The report is a company related article so I do not see why Victoria would need people contacting her at home - unless she actually worked from home.
10209 In my opinion, Victoria should not publish any of that information on the website with her report. My reasons are as follows: Company address By withholding her company address, she can be sure she is not revealing any company sensitive information and also the people browsing the website will not be able to identify the individuals whose data she has included in her report. Personal telephone number By withholding her personal telephone number she reduces the risk of getting crank or unwanted phone calls. Email address By withholding her email address she prevents people from sending her junk emails.
10399 I think she should publish the company address and email address so that anyone who was interested or concerned in any way about the information could make contact. Letters to the company address could then be dealt with in a controlled and orderly way as could the emails. The publication of a personal telephone number would not be a good idea however because people ringing up would have to be dealt with on the spot and could easily swamp the telephone line and they couldn't be vetted or filtered in a way that emails and letters could and so cause serious problems.
10418 Which of the following should be published on the website with her report and why? The company address should be published with the report, this gives the report and indication of who it belongs to. Personal telephone number should never be published on the web, as this could lead to all sorts of privacy issues being breached by all sorts of parties. The company's email address should be published along with the report, this is mainly for point of contact for users of the report, for additional information.
10475 The Company address could be published as it is public data and anyone can obtain it from public archives ie phone book . It does not pertain to an individual. A personal phone number could be published if it has been obtained fairly and lawfully. For ethical reasons she should first obtain permission from the person involved. An email address should not be published unless permission is obtained first. Putting someones e-mail onto a general website puts it into the public domain where it could become a target for junk messages.
10855 The following could be published on the site: Company address :- This information could be obtained freely by phoning the local yellow pages, therefore putting the information on the site is of no harm to any individuals in the company. e-mail address: - The company would want this on the site so that people can respond with ideas from the site, however if they were going to do this then they might be better to set up a dedicated email address for this purpose Personal phone numbers: - Should not be placed on the site, some people may be ex-directory and others just might not want anyone knowing their address.
10969 She should publish both the company address and a contact email address so that anyone who objects to data about themselves being on the web site can contact the company to complain. She should not publish her personal telephone number for safety reasons.
11273 She should publish her company address on the website so that anyone wishing to contact the company, may be to do business may do so. Also, there may be someone who, in creating their own web site, wishes to create a link to Victoria's web site. If they have the company address they can contact the company to tell them of their intentions. Victoria should not publish her personal phone number for reasons of privacy. The report will be accessible to anyone who has access to the internet and a private telephone number is

something many people do not wish to become public property. Again, for reasons of privacy similar to the mentioned above Victoria, should not publish her email address on the website with her report. It may also lead to her receiving vast amounts of unwanted emails.
11577 email address The company address should be published, as it is a company report. Email address only if it is a company email address, your personal email address contain more information about yourself (Via newsgroup or mailing list).
11786 NULL
12736 Company Address In order that visitors to the web site may be able to contact the company by mail. Email Address Only if this is the company email contact. It would not be advisable to publish a personal email address.
12812 I think Victoria is safe to publish the company address on the web site. However, by publishing her private telephone number and her email address, she may be become vulnerable. For example, she may be targeted by unsolicited mail or bothersome phone calls.
13211 Company address. Generally, company addresses are in the `public domain', so publishing them is not illegal. However, there is a question here of whether the company wants to be associated with the report or not. If the report is the property of the company and the company wishes their address to be published with the report then Victoria should comply. * Personal telephone number. She could publish her personal telephone number. However she might receive a large number of unsolicited calls. For this reason I would not recommend it. * Email address. When publishing email addresses on the web there's always the problem of being put on spammers' lists. However, this seems to be the best option for enabling readers of the report to contact her. Perhaps, she should assign a new, dedicated e-mail address for feedback purposes.
13401 She should publish the company address and email address, I am making the assumption that the report is of a business nature. So her personal telephone number would not be relevant for the feedback of the report, unless it was her personal choice to do so.
13553 The information which I believe should be included on the webpage is the company's address and Victoria's email address (I am assuming that the email address is a work based one and not a personal one). As the webpage will contain information created by the company then their contact address should be given, this is so that the company can be reached by a more secure method if problems or queries may arise from the viewers. Some people affected by the report may not have access to email facilities and may need to contact the company. An email address will be required on the webpage for the main reason of contacting the reports creator, Victoria may gain feedback from colleagues or she may hear from the individuals in the report. I do not think that Victoria's personal phone number should be used on the webpage, as this could be used by anyone who has had access to the webpage to gain other information about her. For example, her address could be gained with the use of her phone number.
13876 COMPANY ADDRESS: Because the report is company based then the company address should be included. It would let the audience recognize that it is not a personal document and show where the author of the report worked. It may also encourage some business for the company. PERSONAL TELEPHONE NUMBER: Unless the personal telephone number was also her business number, then I would advise against publishing it on the web. The website may be accessed by any number of people and the implications here could be enormous, for example, she may start getting malicious phone calls, people may be able to connect her name and phone number with an address, which then could be used for criminal intent. EMAIL ADDRESS If the email address is a personal one, then I would not publish it on a website. Spammers have programs, which crawl through web pages looking for email addresses (Raz n.d.) and this would cause the person to receive junk mail or even viruses. If the email address was a company one then I would publish it on the web as there may be a reason for wishing to contact the author. There is encryption software, which can be bought that does not let email extractors find email addresses on protected websites (AtomPark 2002).
14047 All three should appear on her website.The company address because she is writing a report on behalf of her company who ultimately own the document.Her telephone number and e-mail address should be available to people who have any queries with the report
15757 Company address, personal telephone number, email address [3] The company address must be published on the website with the report as its readers should be able to tell who and where the report has been written. This will inform the readers who the report has been published by and who to contact in case of inquiries. It is unwise to publish her personal telephone number on the internet as this might lead to unwanted telephone calls at home which she would like to avoid. If she wishes to publish a telephone number, it is a better idea to publish a work number, if she has one, so that she can answer any queries regarding her report over the telephone. Victoria should publish an e-mail address with her report so that she can get response about her report from the public over the internet. Because she wishes to publish the report on the internet, most responses towards it will be through e-mail therefore, one should be made available to its readers.
15890 It is not advisable to publish personal telephone number as it could be a target for unnecessary calls. Although telephone numbers are widely available to most people through telephone directories. She shouldn't really publish email address as it could be target for junk messages. The company address should be published as most people would know this anyway and it advertises the company.
16061 Victoria should publish the company's name on the website because it is good practice to make known the agent responsible for any publication. She would be very foolish to put her phone number or her e-mail address on the website as she may be subjected to unwelcome and malicious contacts. A genuine enquirer can get in touch with her by going through the company which is another reason why its name should be published.
16118 When publishing her report Victoria does not have to publish any contact details. But, it would be advisable to have an address and email account for contact purposes. The company address for people who maybe require an address for written (snail mail) documentation to be sent to the company. For instance a lawyer wanting to write to them to let them know they are being sued for breach of the data protection act. And an email address for contact via computers. Unless she likes strange phone calls, then under no circumstances should Victoria publish her personal phone number on the web.
16289 She should publish her company address on the website and possibly an e-mail address for the company put none of her personal information should be on the web page.
16384 NULL
17030 Company address Personal telephone number Email address She can publish company's address and email address with her report on the website because clients and other reader people who are interested to contact, they can contact by postal system or via email.

Appendix D Raw Marks Given by Human Markers

Question 1	Markers				
Answer	1	2	3	4	5
79	8	8	8	8	8
87	8	8	6	6	6
243	8	8	8	8	8
273	7	7	7	7	7
329	6	6	5	5	5
345	7	7	6	7	7
999	8	8	8	8	8
1025	8	8	8	8	8
1175	6	7	6	4	5
1187	7	8	8	7	8
1231	8	7	8	7	8
1287	7	7	8	7	7
1417	8	8	8	8	8
1549	8	8	8	8	8
1567	8	8	6	6	6
1657	7	8	7	6	7
1701	7	6	7	7	8
1797	8	7	6	7	6
1885	8	8	8	8	8
2107	8	7	8	8	8
2369	8	8	8	8	8
2601	8	8	8	8	8
2661	4	5	5	5	5
2691	7	8	8	8	8
2883	8	8	7	6	7
3065	7	8	6	4	7
3067	7	8	7	6	6
3207	8	7	8	7	8
3225	7	7	8	7	8
3327	8	8	6	7	6
3717	5	5	5	4	4
3835	7	7	7	6	7
3849	8	8	8	7	8
3893	8	8	8	8	8
4239	8	8	8	8	8
4269	0	0	0	8	0
4277	8	8	8	8	8
4537	7	8	8	8	8
4589	7	7	7	6	6
4683	7	6	8	7	7
4921	7	7	7	7	8
5005	5	6	7	6	6
5049	6	6	6	5	7
5189	8	8	8	8	8
5197	8	8	8	8	7
5277	6	5	7	5	5
5349	8	7	8	8	8
5703	8	8	6	7	6
5721	4	6	5	6	5
5925	6	5	4	3	4

Question 2	Markers				
Answer	1	2	3	4	5
32	12	12	12	11	12
74	12	12	12	12	12
252	12	12	12	12	12
434	12	11	12	12	12
458	7	6	7	6	7
584	12	12	12	12	12
626	12	11	12	12	12
954	12	12	12	12	12
1066	0	0	0	11	0
1078	11	11	11	10	11
1208	12	12	12	12	12
1300	12	12	12	12	12
1372	12	12	12	12	12
1468	12	12	12	12	12
1502	12	12	12	11	12
1530	12	12	12	12	12
1698	10	9	9	7	10
1884	9	8	9	8	9
2010	12	12	12	12	12
2114	12	12	12	12	12
2306	0	0	0	12	0
2360	12	12	12	12	12
2652	11	11	12	11	11
2910	11	11	12	11	12
3002	12	12	12	12	12
3198	12	12	12	11	12
3226	12	12	12	11	11
3552	12	12	12	12	12
3594	12	12	12	12	12
3644	12	12	12	12	12
3998	12	12	12	12	12
4012	12	12	12	12	12
4066	12	12	12	12	12
4120	12	12	12	12	12
4136	12	12	12	12	12
4170	8	8	8	8	8
4218	12	12	12	12	12
4434	12	12	12	12	12
4550	0	0	0	10	0
5128	12	12	12	12	11
5150	10	10	10	10	11
5264	11	10	11	11	11
5488	10	10	11	10	10
5562	12	12	12	12	12
5686	11	11	12	11	12
5736	12	12	12	12	12
5766	12	12	12	12	12
5790	12	12	12	11	12
6080	12	12	12	12	12
6318	12	12	12	12	12

Question 3	Markers				
Answer	1	2	3	4	5
8291	3	4	4	1	3
8671	4	4	4	1	4
9241	3	4	4	3	4
10400	3	4	4	3	3
11350	4	4	4	3	4
13136	4	4	4	2	4
15644	4	4	4	4	4
17430	4	4	3	2	3
17544	4	4	4	3	4
18000	4	4	4	3	4
19577	3	4	4	4	4
19824	3	4	3	4	3
20546	4	4	4	4	4
20869	4	4	4	3	4
21154	4	4	4	2	4
21458	4	4	4	3	4
21648	4	4	4	3	4
22940	3	3	4	1	3
24745	4	4	4	4	4
25144	4	4	4	3	4
27519	4	4	4	3	4
28621	4	4	4	3	3
28792	3	4	4	4	4
29134	4	4	4	4	4
30065	4	4	4	4	4
32592	4	4	4	4	4
36487	4	4	4	4	4
37646	4	4	3	2	4
38615	4	4	4	2	4
39261	4	4	4	0	4
40990	4	4	4	4	4
43270	3	4	4	4	4
45778	3	4	4	0	2
47507	4	3	4	2	4
48020	4	4	4	4	4
49027	4	4	4	4	4
54442	3	3	3	2	3
55278	3	4	3	3	4
55753	4	4	4	4	4
56114	4	4	4	2	4
57596	3	4	4	3	4
59021	4	4	4	3	4
59268	4	4	4	4	3
59743	4	4	4	3	4
61700	2	4	4	2	4
62118	4	4	4	4	4
63049	4	4	4	3	4
64170	4	4	4	3	4
64778	3	4	4	0	4
67229	4	4	4	4	4

Raw Marks by Human Marker (continued)

Quest ion 4	Markers				
Answer	1	2	3	4	5
9090	1	3	3	3	3
9489	4	4	4	3	1
9622	3	3	3	3	0
10021	3	3	3	3	0
10230	2	3	3	3	0
11655	3	4	4	4	4
12776	1	3	3	3	2
14087	2	4	4	3	2
15588	0	0	0	3	0
15683	3	3	3	3	2
18780	2	3	3	3	2
21459	3	4	4	4	2
21972	2	3	3	3	2
23055	1	3	3	3	3
24024	2	3	3	3	3
25677	2	3	3	3	4
26285	3	3	3	3	3
26684	3	3	3	3	3
27159	2	3	3	3	4
27463	0	0	0	3	0
28204	2	3	3	3	3
28850	4	4	4	3	4
30332	4	4	4	4	2
31605	3	4	4	4	1
32289	0	0	0	3	0
34702	2	3	3	3	4
37723	0	2	3	2	2
37989	3	4	4	3	3
42568	2	3	3	3	3
44031	2	3	3	3	4
44183	1	3	3	3	2
44886	0	2	3	2	1
46083	2	3	3	4	0
46995	1	3	3	3	2
47698	2	3	3	3	3
53284	3	4	4	4	2
55279	2	3	3	3	4
56020	0	3	2	2	1
56552	2	4	3	3	2
57426	2	3	3	3	3
57958	2	3	3	3	0
59801	3	3	3	3	1
62024	3	4	3	3	2
63164	2	3	3	3	4
64095	2	4	4	4	2
65235	1	3	3	3	3
65292	0	0	0	4	0
66546	0	2	2	2	0
66964	2	4	4	4	3
67401	2	3	3	3	0

Question 8	Markers				
Answer	1	2	3	4	5
8030	2	4	4	3	4
9759	4	4	4	4	4
9968	4	4	4	4	4
10006	4	4	4	4	4
10405	4	4	4	4	4
10462	4	4	4	4	4
10899	4	4	4	4	4
11431	4	4	4	4	4
11583	3	4	4	4	4
11697	3	4	3	3	4
11868	3	4	4	3	4
11925	3	4	4	4	4
12305	4	4	4	4	4
12324	4	4	4	4	4
12400	4	4	4	4	4
12571	4	4	4	4	4
12970	4	4	4	2	4
13331	4	4	4	4	4
14357	3	4	4	3	4
14680	4	4	4	4	4
14946	0	0	0	0	0
14984	4	4	4	4	4
16333	4	4	4	4	4
16732	3	4	3	3	4
16789	2	2	2	2	2
17359	4	3	3	3	4
17378	4	4	4	4	4
17796	3	4	4	2	4
18024	3	3	3	0	2
18062	4	4	4	4	4
18765	0	0	0	0	0
19829	4	4	4	4	4
19943	4	4	4	4	4
19981	4	4	4	4	4
20551	3	4	4	4	4
20570	0	0	0	0	0
20741	4	3	4	3	4
21615	4	4	4	4	4
21786	2	4	4	4	4
22033	4	4	4	4	4
22299	4	4	4	4	4
22394	4	4	4	4	4
22527	3	4	4	3	4
22964	4	4	4	4	4
23021	4	4	4	4	4
23078	4	4	4	4	4
23914	4	4	4	4	4
24902	4	4	4	4	4
26270	4	4	4	4	4
26289	4	4	4	4	4

Question 9	Markers				
Answer	1	2	3	4	5
7594	2	4	4	4	3
7746	3	4	3	3	3
8772	4	4	4	4	4
9513	0	0	0	0	0
9817	3	4	3	3	3
10330	3	4	4	3	3
10615	2	4	4	3	4
11242	4	4	4	4	4
11261	2	4	4	4	4
11774	2	2	2	2	4
12002	4	4	4	4	4
12610	2	4	2	2	2
12914	2	4	4	3	4
13009	4	4	4	4	4
13066	4	4	4	4	4
13180	4	4	4	4	4
13408	2	4	4	4	4
13427	3	4	4	4	4
13446	3	4	2	3	3
13465	4	4	4	4	4
13712	2	4	4	4	4
15118	4	4	4	4	4
15232	4	4	4	4	4
15688	4	4	4	4	4
16961	1	2	1	1	1
17170	2	4	2	2	2
17550	4	4	4	4	4
17778	4	4	4	4	4
18025	2	1	4	1	1
18158	1	4	3	3	4
18500	3	4	3	3	3
19811	4	4	4	4	4
20020	2	4	2	2	3
20324	4	4	4	4	4
20609	4	4	4	4	4
20666	4	4	4	4	4
20894	4	4	4	4	4
21730	4	4	4	4	4
21977	4	4	4	4	4
22395	3	4	4	3	3
22737	2	4	2	3	4
22927	4	4	4	4	4
23117	4	4	4	4	4
24029	4	4	4	4	4
24485	1	4	3	3	3
24732	4	4	4	4	4
25511	4	4	4	4	4
25872	4	4	4	4	4
26100	3	4	3	4	3
26537	3	4	4	3	4

Raw Marks by Human Marker (continued)

Question 10	Markers				
Answer	1	2	3	4	5
7481	4	4	4	4	4
7614	2	4	3	3	3
8146	4	4	4	4	3
8374	2	4	4	4	4
9001	4	4	4	4	4
9799	4	4	4	4	4
10711	2	4	3	2	3
11110	4	4	3	3	3
11528	4	4	4	4	4
11699	2	4	4	4	4
12193	4	4	4	4	4
12573	2	4	3	3	3
12725	2	3	3	2	3
12858	2	4	3	1	3
12953	2	4	3	2	3
13276	2	4	3	2	4
13371	2	4	4	4	4
13732	2	4	4	4	4
14093	0	4	3	3	4
14207	3	4	4	3	3
15613	4	4	4	4	4
15670	2	2	4	3	4
16031	2	4	3	2	3
16240	2	4	3	3	3
17057	3	4	4	4	4
17532	2	4	4	4	4
17665	3	4	4	4	4
17817	3	4	4	4	4
17988	3	4	4	2	4
18007	2	4	3	3	3
19375	4	4	4	4	4
19394	2	4	3	3	4
19413	2	4	3	1	3
19432	2	2	2	2	2
20078	2	4	3	3	3
20914	4	4	4	4	4
21845	4	4	4	4	4
22282	0	4	3	3	3
23517	3	4	3	3	4
23707	2	2	2	2	3
24239	4	3	4	4	4
24562	3	4	4	4	4
24771	2	2	4	4	3
25018	4	4	3	3	3
25132	4	4	4	4	4
25379	2	2	2	2	3
25892	2	4	3	3	4
25968	2	4	3	3	3
26044	4	4	4	4	4
26177	4	4	4	4	4

Question 11	Markers				
Answer	1	2	3	4	5
8280	4	4	4	4	4
8755	2	2	2	2	2
9363	3	4	3	4	4
9515	0	0	0	0	0
9648	4	3	4	4	4
10123	1	4	3	2	2
10446	3	4	4	4	4
10902	1	3	3	2	2
11016	0	4	4	4	4
11073	3	4	3	4	2
11396	3	4	4	4	4
11491	2	4	4	4	4
11586	3	4	3	4	2
11757	3	4	4	4	4
12118	0	0	2	0	0
12802	2	2	2	2	2
13125	4	4	4	4	4
13581	3	3	4	2	3
14227	4	4	4	4	4
14284	3	4	4	4	3
14626	3	4	4	4	4
14683	4	4	4	4	4
14759	4	4	4	4	3
14816	2	4	3	4	4
14835	4	4	4	4	4
15823	3	4	3	3	3
16051	4	3	4	4	3
16450	3	4	2	3	3
17609	4	4	3	4	4
17894	4	4	4	2	4
17932	2	4	4	4	3
18502	4	4	4	4	4
18768	0	0	0	0	0
19870	0	0	0	0	0
19927	3	4	4	4	4
20212	4	4	4	4	4
20649	3	4	4	4	4
20763	2	2	2	2	2
21428	0	0	2	1	0
22074	3	4	3	3	3
22777	4	4	4	4	4
23271	2	3	4	2	2
23290	2	4	3	3	3
23309	2	4	4	4	4
23328	3	4	4	3	3
24430	4	4	4	4	4
25703	1	2	4	1	2
25798	0	2	4	1	1
26501	3	3	3	4	4
26577	4	4	4	4	4

Question 12	Markers				
Answer	1	2	3	4	5
8604	4	3	4	2	3
9250	3	4	4	3	3
9725	4	4	4	3	4
9991	4	4	4	2	4
10200	4	4	4	4	4
10257	0	2	2	2	2
10713	0	4	2	2	4
10903	3	4	3	4	4
11055	3	4	4	2	4
11663	3	4	4	3	3
11948	4	4	4	3	4
12119	0	1	3	2	0
12594	4	4	4	1	3
13696	3	4	4	2	3
14304	4	4	4	3	4
14627	4	3	4	3	4
14760	4	4	4	3	3
15406	0	4	4	3	4
15729	4	4	4	2	4
15862	4	3	4	4	4
15938	4	4	4	4	4
15957	4	4	4	4	3
16679	3	4	4	3	4
17021	0	1	3	0	4
17116	4	4	4	2	4
17458	4	3	4	2	3
17496	3	4	4	3	4
18047	2	2	2	1	2
18332	4	4	4	4	4
18389	3	3	3	2	3
18655	0	0	2	0	0
19035	2	4	2	4	4
19624	3	4	4	2	4
19776	3	4	4	3	3
20118	4	4	4	3	4
20156	2	3	4	1	3
20707	3	4	4	3	4
21239	3	3	4	2	4
21980	4	4	4	2	4
22113	2	2	2	2	2
22246	1	3	2	3	2
23291	2	3	2	2	3
23386	2	2	2	2	2
23443	4	4	4	4	4
23785	3	4	4	2	4
23823	4	4	4	2	4
24621	3	4	4	2	4
24640	3	4	4	3	4
24868	4	4	4	4	4
26065	2	4	2	2	3

Raw Marks by Human Marker (continued)

Question 13	Markers				
Answer	1	2	3	4	5
7731	2	2	2	2	2
7845	2	2	2	2	2
8586	2	2	2	2	2
9612	2	2	2	2	2
10410	2	2	2	2	2
11303	2	2	2	2	2
12348	2	2	2	2	2
17117	2	2	2	2	2
17839	2	2	2	2	2
18333	2	2	2	2	2
18409	2	2	2	2	2
18561	2	2	2	2	2
19169	0	0	0	0	0
19720	2	2	2	2	2
19891	2	2	2	2	2
20176	2	2	2	2	2
20689	2	1	1	2	2
21487	2	2	2	2	2
21981	2	2	2	2	2
22912	2	2	2	2	2
24584	2	2	2	2	2
24812	2	2	2	2	2
25116	2	2	2	2	2
25705	2	2	2	2	2
26237	2	2	2	2	2
26731	2	2	2	1	2
27187	2	2	2	2	2
28042	2	2	2	2	2
28175	2	2	2	2	2
29410	2	2	2	2	2
30341	2	2	2	2	2
30379	2	2	2	2	2
30531	2	2	2	2	2
31310	2	2	2	2	2
31880	2	2	2	2	2
32241	1	2	2	1	2
34445	2	2	2	2	2
34502	2	2	2	2	2
35034	2	2	2	2	2
35395	2	2	2	2	2
35547	2	2	2	2	2
35661	2	2	2	2	1
36687	2	2	2	2	2
36763	2	2	2	1	1
37276	2	2	2	2	2
37865	2	2	2	2	2
38701	2	2	2	2	2
38872	2	2	2	2	2
41969	2	2	2	2	2
42026	2	2	2	2	2

Question 14	Markers				
Answer	1	2	3	4	5
7561	2	2	2	2	2
7637	2	2	2	2	2
7713	2	2	2	2	2
8074	0	0	0	0	0
8511	2	1	2	2	2
8625	2	2	2	2	2
9727	2	2	2	2	2
10012	2	2	2	2	2
10164	2	2	2	2	2
10696	2	2	2	2	2
10905	2	2	2	2	2
11114	2	2	2	2	2
11874	2	2	2	2	2
12577	2	2	2	2	1
13109	0	1	0	0	0
13603	2	2	2	2	2
13983	2	2	2	2	2
14211	2	2	2	2	2
14800	2	2	2	2	2
15902	2	2	2	2	2
15978	2	2	2	2	2
16928	2	2	2	2	2
17270	2	2	2	2	2
17612	2	2	2	2	2
17783	2	2	2	2	2
18087	2	2	2	2	2
19778	2	2	2	2	0
20082	0	2	0	0	0
20272	2	2	1	2	2
20386	2	2	2	2	2
20576	2	2	2	2	2
20747	0	0	0	0	0
21583	2	2	2	2	2
21773	2	2	2	2	2
22001	2	2	2	2	2
22533	2	2	2	2	2
22951	0	0	0	0	0
23103	2	2	2	2	2
23483	2	2	2	2	2
23787	2	2	2	2	2
24167	0	1	0	0	1
24414	2	2	2	2	2
25022	1	2	2	2	2
25269	2	2	2	2	2
25326	2	2	2	2	2
25820	2	2	2	2	2
26181	2	2	2	2	2
26238	0	0	0	0	1
26428	2	2	2	2	2
26580	2	2	2	2	2

Question 15	Markers				
Answer	1	2	3	4	5
7467	0	1	0	0	1
7695	2	2	2	2	2
7961	1	2	2	1	0
8227	2	2	2	2	2
8531	2	2	2	2	2
8645	2	2	2	2	2
9158	1	1	0	1	0
10013	1	1	2	1	0
10070	2	2	2	2	2
10184	2	2	2	2	2
10792	1	2	1	1	0
11058	1	2	2	1	0
11229	1	0	1	1	0
11666	1	1	2	1	0
11970	1	1	2	1	0
12350	1	1	1	1	0
12901	1	1	2	1	0
12939	1	1	2	1	0
13015	1	2	1	1	0
13300	1	1	0	0	0
13414	1	1	2	1	0
14098	1	1	2	1	0
14155	1	1	0	0	0
14706	1	1	2	1	0
15105	2	2	2	2	2
15238	2	2	2	2	2
15675	1	1	2	1	0
16777	2	2	2	2	2
18715	2	2	2	2	2
18981	0	0	0	0	0
19000	1	1	2	1	0
19304	2	2	2	2	2
19361	2	2	2	2	2
19456	0	0	0	0	0
19950	1	1	2	1	0
20539	1	1	2	1	0
20957	1	1	2	1	0
21432	1	1	1	1	0
21470	1	1	2	1	0
21641	1	1	1	1	0
22686	2	2	2	2	2
22838	1	1	2	1	0
23047	2	2	2	2	2
23408	2	2	2	2	2
25061	2	2	2	2	2
25270	2	2	2	2	2
25289	1	0	2	1	0
25669	1	1	2	1	0
25973	1	2	2	1	0
26524	1	1	0	1	0

Raw Marks by Human Marker (continued)

Question 16	Markers				
Answer	1	2	3	4	5
7563	2	2	2	2	2
7753	2	2	2	2	2
8095	2	1	1	1	1
8475	2	2	2	2	2
8741	2	2	2	2	2
8855	2	2	2	2	2
9577	2	2	2	2	2
10014	1	1	1	1	1
10470	2	2	2	2	2
11819	2	2	2	2	2
12807	2	2	2	2	2
12978	2	2	2	2	2
13073	2	2	2	2	2
13985	2	2	2	2	2
15296	2	2	2	2	2
15714	0	0	0	0	0
15885	2	2	2	2	2
15961	2	2	2	2	2
16303	1	1	1	2	1
16740	2	2	2	2	2
17215	2	2	2	2	2
17899	2	2	2	2	2
18127	2	2	2	2	2
18222	2	2	2	2	2
18469	2	2	2	2	2
19476	2	2	2	2	2
19875	0	0	0	0	0
20027	2	2	2	2	2
20312	2	2	2	1	2
20369	2	2	2	2	2
20559	2	2	2	2	2
20825	2	2	2	2	2
21015	2	2	2	2	2
21034	2	2	2	2	2
21300	2	2	2	2	2
22079	2	2	2	2	2
22307	2	2	2	2	2
22421	2	2	2	2	2
22782	2	2	2	2	2
23371	2	2	2	2	2
23504	2	2	2	2	2
23580	2	2	2	2	2
24169	2	2	2	2	2
24321	2	2	2	2	2
24492	2	2	2	2	2
24796	2	2	2	2	2
25594	2	2	2	2	2
25898	2	2	2	2	2
25955	2	2	2	2	2
26506	1	1	1	1	2

Question 17	Markers				
Answer	1	2	3	4	5
7944	1		2	0	1
8001	1		2	0	2
8951	1		2	1	1
9844	0		1	0	0
9920	1		1	1	0
10053	1		2	0	1
10737	2		2	1	0
10870	2		2	1	1
11136	0		2	0	0
11193	2		2	0	0
11573	1		2	0	1
11706	2		2	0	1
11896	2		2	1	1
12333	1		1	1	1
12998	1		2	0	1
13055	0		0	0	0
13093	1		2	1	2
14765	2		1	1	1
15126	2		2	0	2
15430	1		1	0	1
15829	2		1	0	0
15943	1		2	0	1
16000	1		2	0	2
16513	1		2	0	2
16627	0		1	0	1
16722	1		2	0	2
17767	2		2	0	2
17938	2		1	0	0
18299	1		2	0	1
18527	2		2	0	2
18774	0		0	0	0
18793	2		2	1	1
18926	0		2	0	1
19021	1		2	0	1
19420	1		2	0	1
19857	1		2	1	2
20104	1		2	0	2
20161	2		2	1	2
20199	0		0	0	0
20503	2		2	0	2
21225	1		2	1	1
21263	1		1	0	1
21985	1		1	0	1
22707	1		2	2	2
22897	2		2	2	1
24721	2		2	0	0
24854	2		2	1	2
25367	0		1	0	0
25804	1		2	1	1
25975	1		2	0	1

Question 18	Markers				
Answer	1	2	3	4	5
7660	1	1	1	1	1
7698	0	0	0	0	0
8249	1	1	2	2	1
9104	2	2	2	2	1
10643	1	1	1	1	1
10947	2	2	2	2	1
11080	2	2	2	2	2
11441	2	2	2	2	2
11555	1	1	1	1	1
11612	2	2	2	2	2
11878	1	2	2	2	1
12600	1	2	2	2	1
12733	1	2	2	2	1
13398	2	1	1	1	1
13930	2	2	2	2	2
14234	1	1	1	1	1
14367	2	2	2	2	2
14405	2	2	2	2	2
14576	2	1	1	1	1
15526	2	2	2	2	2
15583	2	2	2	2	2
15697	2	2	2	2	2
16210	1	0	0	0	0
16343	2	2	2	2	2
16666	1	1	1	1	1
17046	1	0	0	0	1
18034	2	2	2	2	2
18756	2	2	2	2	1
19383	1	1	1	0	1
19725	2	2	2	2	2
19877	0	0	0	0	0
20029	2	1	1	1	1
20428	0	1	0	0	1
20751	2	1	1	2	1
21131	2	2	2	1	2
21606	0	0	0	0	0
21663	1	1	1	1	1
21796	2	1	1	1	1
22043	2	2	2	2	2
22309	2	1	1	1	1
24000	1	1	1	1	1
24646	1	2	1	1	1
24760	2	2	2	2	2
24798	1	1	1	1	1
24836	2	2	2	2	2
25425	2	2	2	2	1
25634	0	0	0	0	0
25938	2	2	2	2	2
26071	1	1	2	1	0
26261	0	0	0	0	0

Raw Marks by Human Marker (continued)

Question 19	Markers				
Answer	1	2	3	4	5
7452	1			1	1
7547	2			2	2
7585	1			1	1
7680	1			1	1
7851	1			1	1
8326	1			1	1
8649	1			1	1
9143	1			1	1
9333	1			1	1
9694	1			1	1
10036	1			1	1
10378	0			0	0
12107	1			1	1
12468	1			1	1
12563	1			1	1
13000	1			1	1
13038	1			1	1
13532	2			2	2
13817	1			1	1
14919	2			2	2
14957	0			0	0
15090	1			1	1
15584	2			2	2
16192	1			1	1
16496	1			1	1
16648	1			1	1
17313	1			1	1
18795	1			1	1
20011	1			1	1
20125	2			1	1
20258	1			1	1
20277	1			1	1
20296	1			1	1
20581	1			0	1
20676	1			1	1
20828	1			1	1
21379	1			1	1
21512	1			1	1
21987	1			1	1
22405	1			1	1
22709	1			1	1
23127	1			1	1
23203	1			1	1
23830	1			1	1
24229	1			1	1
24419	1			1	1
24704	1			1	1
24742	1			1	1
25122	1			1	1
26319	0			0	0

Question 20	Markers				
Answer	1	2	3	4	5
7510	1			1	2
7776	0			0	2
8080	0			0	2
8118	0			0	1
8365	1			1	2
8536	0			0	1
8574	1			1	1
9049	1			1	2
9486	1			0	2
9543	1			0	2
10075	0			0	0
10284	1			1	1
10360	1			1	2
10474	1			1	2
10588	1			2	2
11215	1			0	2
11614	1			2	1
11671	1			2	2
11728	1			0	2
11766	0			0	2
12526	0			0	2
12621	0			0	2
13894	1			0	2
14103	1			0	2
15262	1			0	2
16250	1			1	1
16820	0			1	2
17257	1			1	2
17732	1			1	3
17884	1			1	1
17941	0			0	2
18055	0			0	2
18264	0			0	3
18720	0			0	2
19176	0			0	0
19366	1			1	2
19385	1			3	1
20183	0			1	2
20487	1			1	1
20829	2			1	2
20867	1			0	2
21532	2			1	2
21722	1			0	1
22387	0			0	1
22843	0			0	0
23432	0			0	0
23470	0			0	2
24363	1			1	2
26282	1			0	2
26415	1			0	2

Question 21	Markers				
Answer	1	2	3	4	5
7454	2			3	2
7549	3			3	2
7853	2			3	3
7929	3			3	3
8157	3			2	1
8252	2			2	1
8765	3			2	2
8917	3			3	2
10209	1			2	3
10399	3			2	2
10418	3			3	2
10475	1			2	1
10855	3			3	3
10969	3			2	1
11273	2			3	3
11577	2			1	1
11786	0			0	0
12736	2			1	1
12812	2			2	2
13211	3			3	3
13401	2			1	1
13553	2			2	2
13876	3			3	2
14047	1			1	1
15757	2			3	2
15890	2			3	3
16061	3			3	3
16118	3			2	2
16289	1			1	1
16384	0			0	0
17030	1			1	1
17106	3			3	2
17258	2			2	2
17277	3			2	2
17600	0			0	0
18303	2			2	3
20868	3			3	3
21457	3			3	2
21989	0			1	2
22065	2			0	1
22179	1			0	1
22654	3			3	3
23319	3			3	2
23699	3			2	2
23775	1			1	1
23965	1			1	1
25428	3			3	2
25656	2			3	2
25789	2			3	2
26169	3			2	2

Appendix E Stop Words

a	anyhow	behind	consequently	either
as	anyone	being	consider	else
able	anything	believe	considering	elsewhere
about	anyway	below	contain	enough
above	anyways	beside	containing	entirely
according	anywhere	besides	contains	especially
accordingly	apart	best	corresponding	et
across	appear	better	could	etc
actually	appreciate	between	couldnt	even
after	appropriate	beyond	course	ever
afterwards	are	both	currently	every
again	arent	brief	d	everybody
against	around	but	definitely	everyone
aint	as	by	described	everything
all	aside	c	despite	everywhere
allow	ask	cmon	did	ex
allows	asking	cs	didnt	exactly
almost	associated	came	different	example
alone	at	can	do	except
along	available	cannot	does	f
already	away	cant	doesnt	far
also	awfully	cause	doing	few
although	b	causes	dont	fifth
always	be	certain	done	first
am	became	certainly	down	five
among	because	changes	downwards	followed
amongst	become	clearly	during	following
an	becomes	co	e	follows
and	becoming	com	each	for
another	been	come	edu	former
any	before	comes	eg	formerly
anybody	beforehand	concerning	eight	forth

four	here	inner	liked	needs
from	heres	insofar	likely	neither
further	hereafter	instead	little	never
furthermore	hereby	into	look	nevertheless
g	herein	inward	looking	new
get	hereupon	is	looks	next
gets	hers	isnt	ltd	nine
getting	herself	it	m	no
given	hi	itd	mainly	nobody
gives	him	itll	many	non
go	himself	its	may	none
goes	his	itself	maybe	noone
going	hither	j	me	nor
gone	hopefully	just	mean	normally
got	how	k	meanwhile	not
gotten	howbeit	keep	merely	nothing
greetings	however	keeps	might	novel
h	i	kept	more	now
had	id	know	moreover	nowhere
hadnt	ill	knows	most	o
happens	im	known	mostly	obviously
hardly	ive	l	much	of
has	ie	last	must	off
hasnt	if	lately	my	often
have	ignored	later	myself	oh
havent	immediate	latter	n	ok
having	in	latterly	name	okay
he	inasmuch	least	namely	old
hes	inc	less	nd	on
hello	indeed	lest	near	once
help	indicate	let	nearly	one
hence	indicated	lets	necessary	ones
her	indicates	like	need	only

onto	re	shall	than	though
or	really	she	thank	three
other	reasonably	should	thanks	through
others	regarding	shouldnt	thankx	throughout
otherwise	regardless	since	that	thru
ought	regards	six	thats	thus
our	relatively	so	the	to
ours	respectively	some	their	together
ourselves	right	somebody	theirs	too
out	s	somehow	them	took
outside	said	someone	themselves	toward
over	same	something	then	towards
overall	saw	sometime	thence	tried
own	say	sometimes	there	tries
p	saying	somewhat	theres	truly
particular	says	somewhere	thereafter	try
particularly	second	soon	thereby	trying
per	secondly	sorry	therefore	twice
perhaps	see	specified	therein	two
placed	seeing	specify	theres	u
please	seem	specifying	thereupon	un
plus	seemed	still	these	under
possible	seeming	sub	they	unfortunately
presumably	seems	such	theyd	unless
probably	seen	sup	theyll	unlikely
provides	self	sure	theyre	until
q	selves	t	theyve	unto
que	sensible	ts	think	up
quite	sent	take	third	upon
qv	serious	taken	this	us
r	seriously	tell	thorough	use
rather	seven	tends	thoroughly	used
rd	several	th	those	useful

uses	way	whenever	whole	yes
using	we	where	whom	yet
usually	wed	wheres	whose	you
uucp	well	whereafter	why	youd
v	were	whereas	will	youll
value	weve	whereby	willing	your
various	welcome	wherein	wish	youve
very	well	whereupon	with	yours
via	went	wherever	within	yourself
viz	were	whether	without	yourselves
vs	werent	which	wont	z
w	what	while	wonder	zero
want	whats	whither	would	
wants	whatever	who	wouldnt	
was	when	whos	x	
wasnt	whence	whoever	y	

Appendix F Testing for Statistical Significance and Effect Size

This appendix explains how to test for statistical significance and effect size and shows how these results for Tables 7-5 through 7-10 and Table 7-12 were determined.

The tables in Chapter 7 report the results of altering one of the LSA parameters. The analysis uses the mean of the percent improvement to determine which parameters yield the best results. However, it is necessary to determine whether the results are statistically significant in order to draw conclusions from the analysis.

The dependent paired sample t test, which compares the means of two groups, is the appropriate statistic to determine if the results are significant (Field, 2005 p. 295). The t test will answer the question of whether the two groups are significantly different.

However, the t test can be used only if the data are normally distributed. The Kolmogorov-Smirnov (K-S) statistic tests this assumption (Field, 2005 p. 93). The table below shows the SPSS output for Table 7-7 which compares the results of removing versus retaining stop words. The numbers under the "Sig." heading are the p values, or statistical significance indicators. If the K-S test is non-significant, i.e., $p > .05$, than the distribution is normal. If the K-S test is significant, i.e., $p <= .05$, than the distribution is non-normal. If the distribution is normal, then the t test can be used to test whether there is a significant difference between the means of the two groups. If the distribution is not normal, then the Wilcoxon Signed Rank test (Field, 2005 p. 306), which is the non parametric version of the t test, must be used.

Tests of Normality

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
removestop	.180	18	.125	.940	18	.288
retainstop	.147	18	.200(*)	.948	18	.395

* This is a lower bound of the true significance.

a Lilliefors Significance Correction

SPSS reports the results of the t test as shown below.

Paired Samples Test

		Paired Differences				
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference	
					Lower	Upper
Pair 1	removestop - retainstop	.07833	.07462	.01759	.04123	.11544

	t	df	Sig. (2-tailed)
Pair 1 removestop - retainstop	4.454	17	.000

From this table, one can infer that the means of the two groups are significantly different (p = .000) and that the difference between them is .07833. That is, there is a statistically significant improvement in results when stop words are retained.

The final step is to determine the effect size (Field, 2005 p. 294), which answers the question of, given that the difference is significant, how important it is. The effect size r is:

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

For this example, r =

$$\sqrt{\frac{4.454^2}{4.454^2 + 17}} = .73$$

An r value that is .1 or less is a small effect, an r value that is between .1 and .5 is a medium effect and an r value greater than .5 is a large effect (Field, 2005 p. 32).

These statistics tell us that retaining stop words has a large, statistically significant effect on the results.

Similar calculations were carried out for all of the relevant tables in Chapter 7. The following table summarizes the results.

Table	mean of the paired differences	K-S test for normal distribution	t test value	test Sig 1- tailed	sig. level	effect size
weighting function						
7-5	-.00778	normal	-.339	.3695	not sig	
number of dimensions						
7-6	-.060	normal	-5.274	.000	>99%	.79 - large
stop words						
7-7	.07833	normal	4.454	.000	> 99%	.73 - large
stemming						
7-8	.05800	normal	4.76	.000	> 99%	.76 - large
amount of training data						
7-9	-.01556	normal	-2.026	.0295	95%	.44 - medium
number of answers to average						
7-10	-.00444	normal	-1.641	.0595	not sig	
proportional weighting						
7-12	.010	normal	1.468	.080	not sig	