Human Fallibility: How Well Do Human Markers Agree?

Debra Haley, Pete Thomas, Marian Petre, Anne De Roeck
The Centre for Research in Computing
The Open University
Walton Hall, Milton Keynes MK7 6AA UK

D.T.Haley, P.G.Thomas, M.Petre, A.DeRoeck at open.ac.uk

Abstract

Marker bias and inconsistency are widely seen as problems in the field of assessment. Various institutions have put in place a practice of second and even third marking to promote fairness. However, we were able to find very little evidence, rather than anecdotal reports, of human fallibility to justify the effort and expense of 2nd marking. This paper fills that gap by providing the results of a large-scale study that compared 5 human markers marking 18 different questions each with 50 student answers in the field of Computer Science. The study found that the human inter-rater reliability (IRR) ranged broadly both over a particular question and over the 18 questions. This paper uses the Gwet AC1 statistic to measure the inter-rater reliability of 5 markers.

The study was motivated by the desire to assess the accuracy of a computer assisted assessment (CAA) system we are developing. We claim that a CAA system does not need to be more accurate than human markers. Thus, we needed to quantify how accurate human markers are.

Keywords: assessment, marker reliability, marker bias, inter-rater reliability, Gwet AC1, computer assisted assessment.

1 Introduction and motivation for the study

Subsections 1.1 and 1.2 show that educators believe assessment is important and costly and that these two factors have led to increasing interest in Computer Assisted Assessment (CAA). One of the critical questions about CAA systems is: How do you measure the accuracy of a CAA system? We believe that a CAA system has *good enough* accuracy if its results agree with humans as well as humans agree with each other. Thus, it is necessary to have reliable figures on human inter-rater reliability (IRR). Although the literature makes claims about the lack of good human IRR, we have been unable to find evidence. This paper provides results of a study to determine human IRR; these results can be used when assessing the accuracy of a CAA system.

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, 2003). 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 is that assessment can have a strong effect on student learning (Brown, Bull & Pendlebury, 1997; Berglund, 1999 p. 364; Daniels, Berglund, Pears & Fincher, 2004). 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).

1.2 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"; they claim a 63% cut in per student resources since 1973 (1997 p. 255).

Ricketts & Wilks (2002 p. 312) agreed with Brown, Bull & Pendlebury (1997) for the increasing interest in CAA – decreasing resources per student require a cost

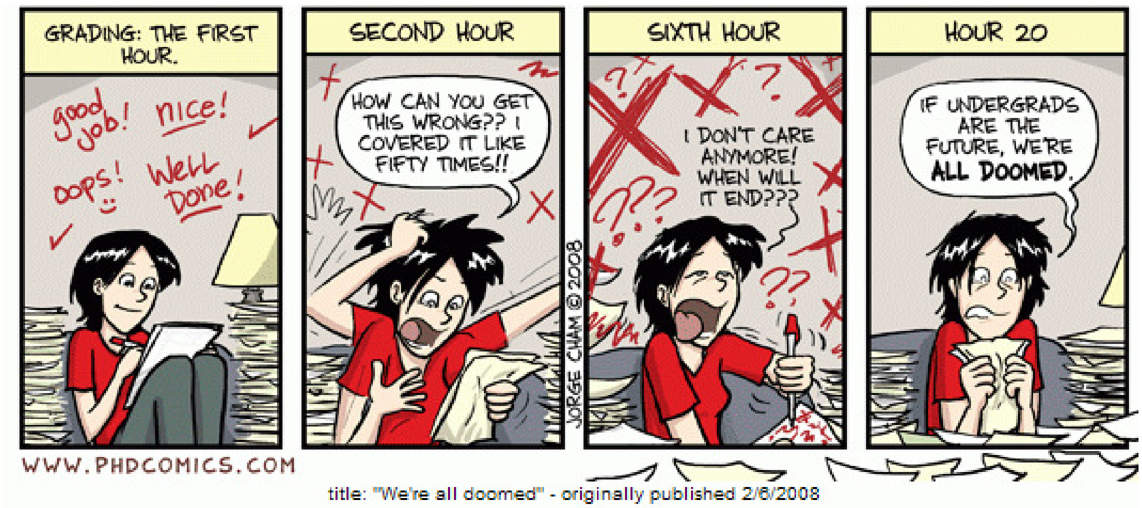

Figure 1 Human fallibility: a source of bias and inconsistency in marking

Used by permission: "Piled Higher and Deeper" by Jorge Cham www.phdcomics.com
http://www.phdcomics.com/comics/archive.php?comicid=974

savings, which can be gained by decreasing tutor marking time. A 2003 survey (Carter, Ala-Mutka, Fuller, Dick, English, Fone & Sheard) 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.

1.3 Reduce marker bias and improve consistency

In addition to the expected cost-savings, one goal of using CAA is to reduce marker bias and improve consistency. This subsection provides evidence that marker bias and inconsistency is *perceived* as a problem. Sections 3 and 4 provide *evidence* of marker inconsistency.

The papers cited used the terms bias and consistency without defining them. In the following paragraphs, we 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 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 paper.) 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 a CAA system named 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) claimed 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 (OU) follows formal procedures to address marker bias and inconsistency. We are 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 study used 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 OU believes human marking can suffer from bias and inconsistency.

This subsection gave examples of the widespread perception that human marking suffers from a lack of consistency. This perception, however, seems to be unsupported by empirical evidence and leads to the motivation for the study.

1.4 Motivation for the study

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 markers (Newstead & Dennis, 1994) provides evidence of poor marker reliability in the field of psychology. Despite these examples, we could find no literature that backed up, with evidence, the claim that CAA improves marking consistency in the field of computer science. 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 paper provides evidence that human markers are far from consistent, at least when marking short answers in the domain of computer science.

2 The Study

This section describes a study to evaluate how closely human markers agree with each other. It was part of a larger effort to develop a Computer Assisted Assessment system (CAA) to mark short answers in the domain of computer science.

2.1 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 our CAA, we needed to quantify how well human markers agree with each other. While it is often claimed that marking variability exists (see the introduction), it is difficult to find supporting evidence. This study provides evidence to support the claim that there is wide variability with human markers.

One can use the results of this study as a baseline against which to compare any CAA. If the results of a CAA closely match or exceed the baseline, then one can be assured that the CAA is *good enough*.

Inter-rater reliability (IRR) is the technical term used to describe how closely raters agree with each other. Gwet (2001 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 (2001 p. 53) and scientists classifying fish according to colour (2001 p. 98) as examples of raters. In this paper, the raters are human markers. The subjects, analogous to Gwet’s patients or fish, are student answers. The AC1 statistic was created to establish the level of agreement among raters (Gwet, 2001 p. vii).

2.2 The participants

We 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 we took the answers-to-be-marked. 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, it is customary for multiple markers to 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.

2.3 The Data

We used 18 different questions for this study (see Appendix A for the text of the questions). 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 OU's 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.

Appendix A 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. Recall that the human marker study was part of a larger effort to develop a CAA system. We removed questions 5, 6, and 7 from the study because being numerical rather than textual, they were unsuited for marking by our assessment system.)

The student answers-being-marked came from the actual student scripts to questions given in the introductory computer literacy course mentioned above. 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. We discarded the marks for the first 10 answers to each question so that the markers could become familiar with the marking scheme before we recorded their marks. To calculate the IRR of the five markers, we 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.

2.4 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. Thus, if they were interested in completing the job as quickly as possible, they could have been careless with their marking. Unfortunately, we 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 would 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 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. OU markers have years of experience carefully following a marking scheme to produce justifiably correct marks. In short, 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. However, 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, we believe the study provides valuable results. The markers were professional and experienced (in contrast to many 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.

3 The results

Figures 1 through 18 in Appendix B display, for each of the 18 questions, the IRR using Gwet's AC1 statistic. For this metric, a higher AC1 number indicates that the relevant markers are closer in agreement than those with a lower AC1 number. 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 to calculating the IRR for each pair of markers, we calculated the overall IRR for all five markers (four for question 17 and three for questions 19-21). In each of the 18 figures, the horizontal line is the IRR for all of the markers; the segmented line shows the IRR for each pair of markers.

Figure 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.

4 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.

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 system as giving the least consistent marks, then one might conclude that the CAA system is not an adequate marker.

Figure 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, 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 CAA system's marks 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.

5 Summary

The purpose of this study was to quantify how well human markers agree with one another in order to evaluate Computer Assisted Assessment Systems. 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.

6 References

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.

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

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.

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.

Christie, James (2003). *Automated essay marking for content - does it work?* **Proceedings of the 7th International CAA Conference**, Loughborough, UK.

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

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.

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

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

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

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

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

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

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

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.

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.

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

Appendix A

	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

Appendix B

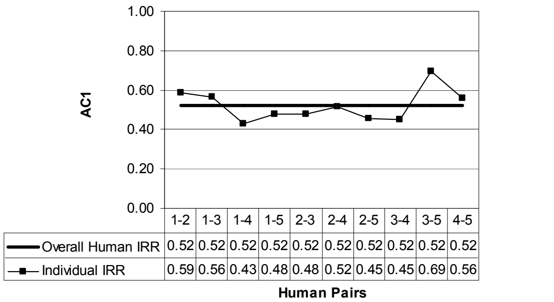

Figure 1 Inter-rater Reliability for Question 1

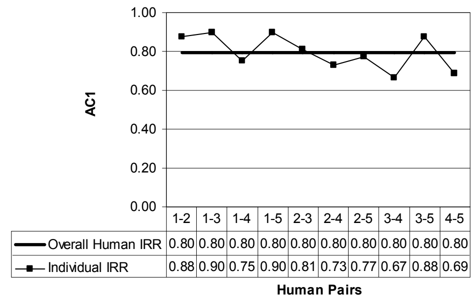

Figure 2 Inter-rater Reliability for Question 2

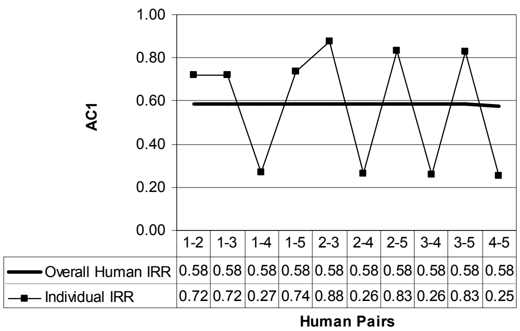

Figure 3 Inter-Rater reliability for Question 3

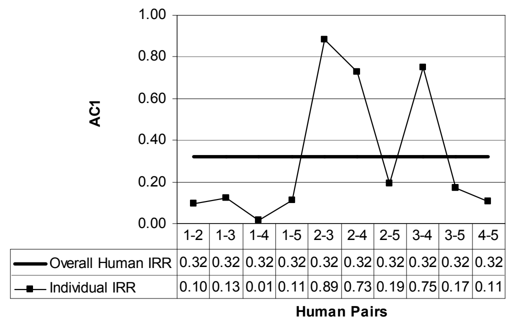

Figure 4 Inter-rater Reliability for Question 4

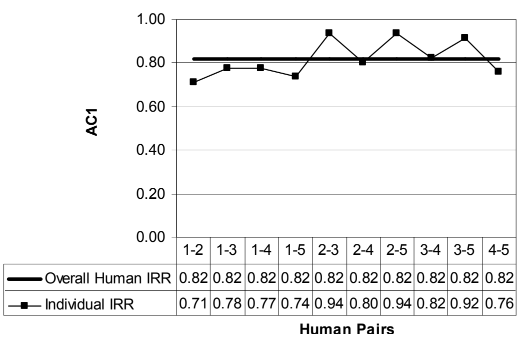

Figure 5 Inter-rater Reliability for Question 8

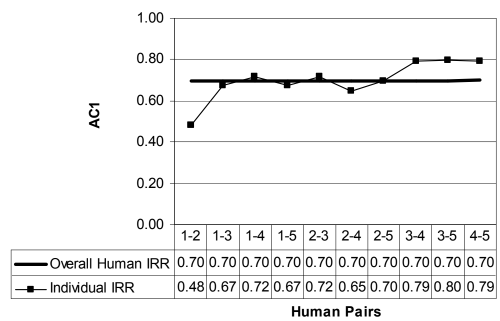

Figure 6 Inter-rater Reliability for Question 9

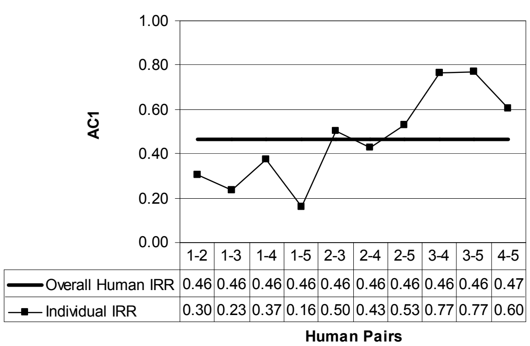

Figure 7 Inter-rater Reliability for Question 10

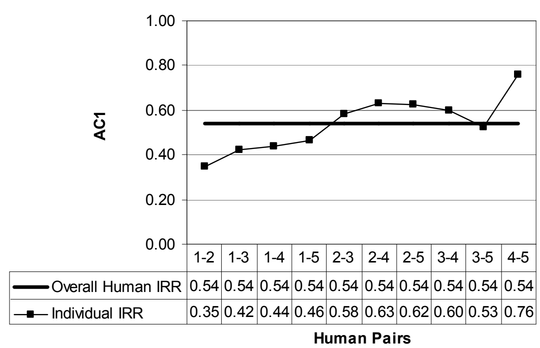

Figure 8 Inter-rater Reliability for Question 11

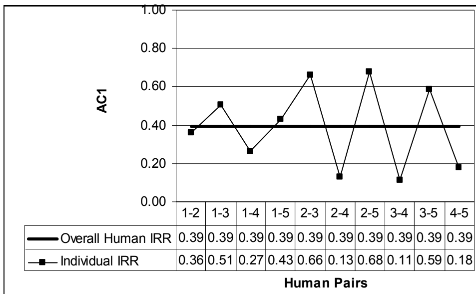

Figure 9 Inter-rater Reliability for Question 12

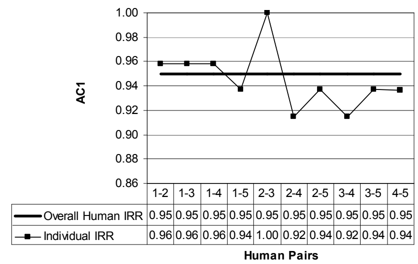

Figure 10 Inter-rater Reliability for Question 13

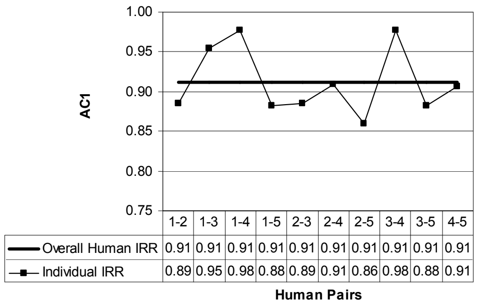

Figure 11 Inter-rater Reliability for Question 14

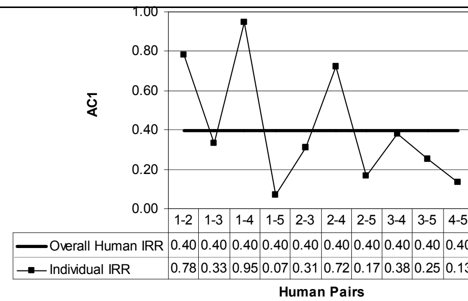

Figure 12 Inter-rater Reliability for Question 15

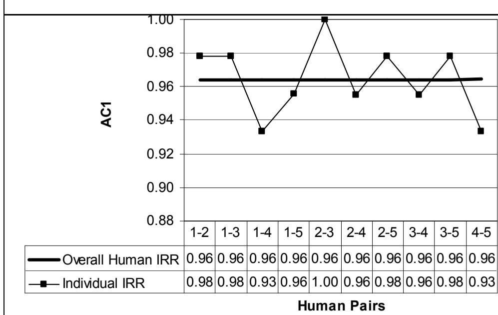

Figure 13 Inter-rater Reliability for Question 16

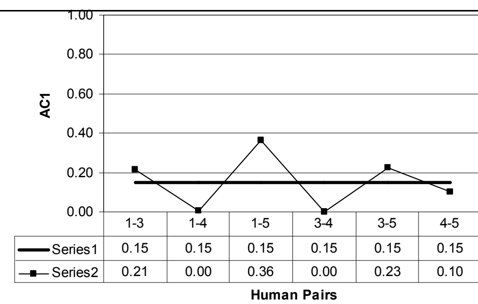

Figure 14 Inter-rater Reliability for Question 17

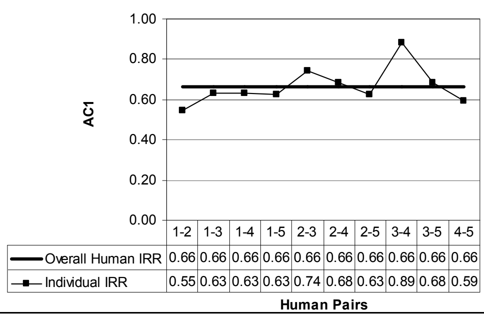

Figure 15 Inter-rater Reliability for Question 18

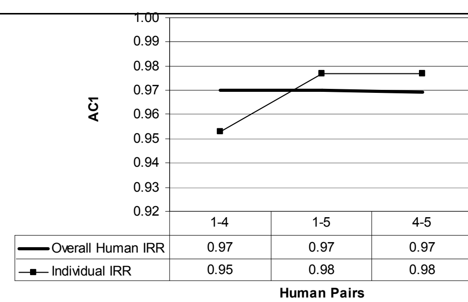

Figure 16 Inter-rater Reliability for Question 19

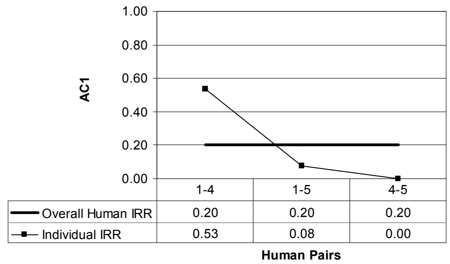

Figure 17 Inter-rater Reliability for Question 20

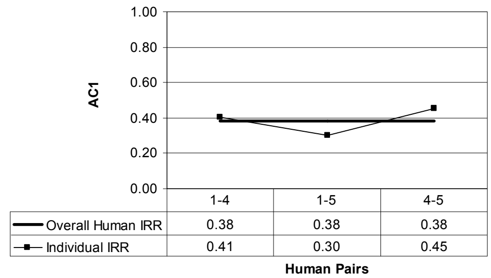

Figure 18 Inter-rater Reliability for Question 21

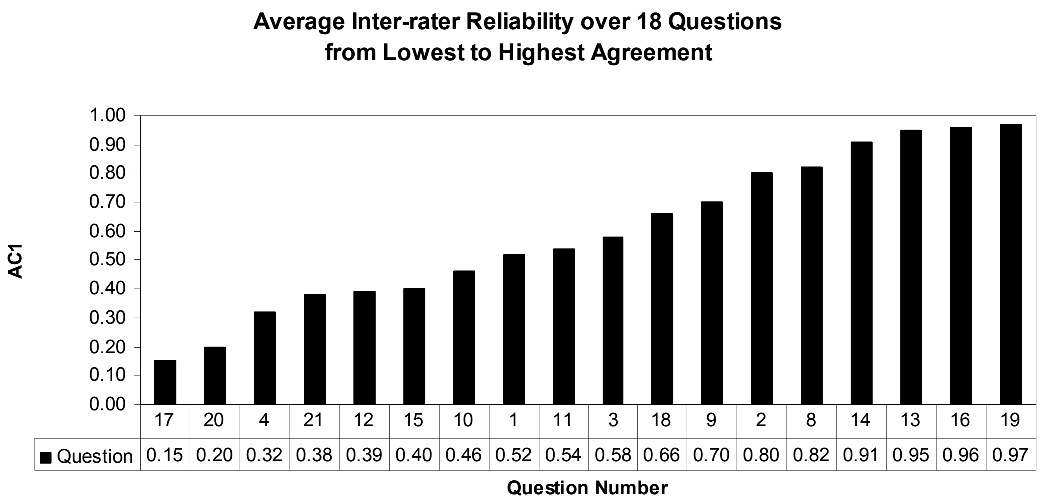

Figure 19 Average Inter-rater Reliability over 18 Questions from Worst to Best