Evaluating the Effectiveness of Live Peer Assessment as a Vehicle for The Development of Higher Order Practice in Computer Science Education

Steve Bennett

School of Computer Science
University of Hertfordshire

Thesis submitted to the University of Hertfordshire in partial fulfilment of the requirements of the degree of Doctor of Philosophy (PhD)

April 2017

Declaration

I certify that the work submitted is my own and that any material derived or quoted from the published or unpublished work of other persons has been duly acknowledged.

Student Full Name: Steve Bennett
Student Registration Number: 00011514

_____

Date: 24th April 2017

Acknowledgements

In completing this study, I would first wish to give huge thanks to my supervisor Dr Trevor Barker who encouraged me to enroll for a PhD in the first place, and supported me through the whole process. Through him I could see the bigger picture and concentrate on the things of genuine significance rather than the details I would sometimes get bogged down in. He was also a great colleague to teach with for many years and I hope I have picked up some of the wisdom and congeniality that he has always demonstrated.

I would also like to give thanks to the University of Hertfordshire generally and particularly my own department of Computer Science. Through the University generally I was lucky enough to participate in the Blended Learning Unit which gave me the time and encouragement to try out new things. There I would particularly like to thank Professor Peter Bullen and David Kraithman for their support.

Then I'd like to thank my own department of Computer Science for giving me a time allowance for this research. Particularly, I'd like to thank colleagues who supported me in different ways along the way. Firstly, I'd like to thank Professor Amanda Jefferies who first invited me onto the Cable (Change Academy for Blended Learning Enhancement) project during the 2009-10 academic year where the practices covered in this thesis were first piloted. The purchase of the first 80 EVS clickers then led directly to this thesis! Secondly I'd like to thank Dr Mariana Lilley for being a supportive colleague on a number of different courses and programs of study where I have been encouraged to try out many different approaches and with whom I have been able to discuss the various issues that came up. Without the anticipation of such support and without an environment in which natural curiosity and experimentation were encouraged I would not have had the ability to try things out to the extent I have.

I would also like to thank my parents, and particularly my mother for noticing my academic tendencies in the first place and encouraging them! And finally, I'd like to thank my girlfriend Vanessa for being such a tower of strength during this period of the writing of this thesis and being such a fabulous person to return to after a day in the office.

Abstract

This thesis concerns a longitudinal study of the practice of Live Peer Assessment on two University courses in Computer Science. By Live Peer Assessment I mean a practice of whole-class collective marking using electronic devices of student artefacts demonstrated in a class or lecture theatre with instantaneous aggregated results displayed on screen immediately after each grading decision. This is radically different from historical peer-assessment in universities which has primarily been asynchronous process of marking of students' work by small subsets of the cohort (e.g. 1 student artefact is marked by <3 fellow students). Live Peer Assessment takes place in public, is marked by (as far as practically possible) the whole cohort, and results are instantaneous.

This study observes this practice, first on a level 4 course in *E-Media Design* where students' main assignment is a multimedia CV (or resume) and secondly on a level 7 course *in Multimedia Specification Design and Production* where students produce a multimedia information artefact in both prototype and final versions. In both cases, students learned about these assignments from reviewing works done by previous students in Live Peer Evaluation events where they were asked to collectively publicly mark those works according to the same rubrics that the tutors would be using. In this level 4 course, this was used to help students get a better understanding of the marks criteria. In the level 7 course, this goal was also pursued, but was also used for the peer marking of students' own work.

Among the major findings of this study are:

- In the level 4 course student attainment in the final assessment improved on average by 13% over 4 iterations of the course, with very marked increase among students in the lower percentiles
- The effectiveness of Live Peer Assessment in improving student work comes from
    - Raising the profile of the marking rubric
    - Establishing a repertoire of example work
    - Modelling the "noticing" of salient features (of quality or defect) enabling students to self-monitor more effectively
- In the major accepted measure of peer-assessment reliability (correlation between student awarded marks and tutor awarded marks) Live Peer Assessment is superior to traditional peer assessment. That is to say, students mark more like tutors when using Live Peer Assessment

- In the second major measure (effect-size) which calculates if students are more strict or generous than tutors, (where the ideal would be no difference), Live Peer Assessment is broadly comparable with traditional peer assessment but this is susceptible to the conditions under which it takes place
- The reason for the better greater alignment of student and tutor marks comes from the training sessions but also from the public nature of the marking where individuals can compare their marking practice with that of the rest of the class on a criterion by criterion basis
- New measures proposed in this thesis to measure the health of peer assessment events comprise: Krippendorf's Alpha, Magin's Reciprocity Matrix, the median pairwise tutor student marks correlation, the Skewness and Kurtosis of the distribution of pairwise tutor student marking correlations
- Recommendations for practice comprise that:
    - summative peer assessment should not take place under conditions of anonymity but that very light conditions of marking competence should be enforced on student markers (e.g. >0.2 correlation between individual student marking and that of tutors)
    - That rubrics can be more suggestive and colloquial in the conditions of Live Peer Assessment because the marking criteria can be instantiated in specific examples of student attainment and therefore the criteria may be less legalistically drafted because a more holistic understanding of quality can be communicated

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Assessment and feedback in terms of their timeliness and effectiveness have been a perennial problem in Higher Education and is invariably the category that demonstrates the least satisfaction among students in the national student survey (2016). To assess is to judge students' work against a standard of quality, and to give feedback is to express how the student succeeds or fails to achieve the standards desired. For this to be a satisfactory process, the conception of quality needs to be shared between the tutor and his/her students, such that any feedback can be received as something embodying a shared understanding. For this feedback to be anything more than an academic exercise, it needs to influence how the student attends to future work, or if it is formative feedback within a larger project, it needs to enable the student to reflect on his/her practice and proceed differently. Moreover, it needs to be timely, received when the work being evaluated is fresh in the students' mind, or when there is still time to change things.

In a situation with large staff-student ratios, all this is difficult to achieve. In the 2016 national student survey, only 59% of students agreed with the statement "Feedback on my work has been prompt". Given those ratios are unlikely to change, what is the best way to ensure a more effective assessment and feedback process, particularly in Computer Science, or in digital topics generally? Another statement in the 2016 national student survey relating to assessment and feedback which also received low assent was "Feedback on my work has helped me clarify things I did not understand", for only 60% of students agreed. Given that in this survey students are reflecting on the totality of their university experience, it means that 40% of students do not find feedback helpful – which is a fairly sobering assessment. However, this really needs unpacking. Does the low assent mean that (a) the feedback was incomprehensible (b) the feedback was lacking or (c) the feedback was wrong? My own belief is that the likeliest explanation is that the feedback they received was not meaningful to them.

One of the most influential projects in Assessment and Feedback in the 21st century UK Higher Education Sector has been the REAP (Re-Engineering Assessment Practices in Higher Education) (2007) project. One of its major findings was that rather than seeking to make academics more efficient producers of feedback, they believed making students more effective monitors of their own work was the key to improving the assessment experience. David Nicol wrote:

> Assessment and feedback practices should be designed to enable students to become self-regulated learners, able to monitor and evaluate the quality and impact of their own work and that of others.

One of the ways this has been attempted in the past has been through self-assessment and peer-assessment. These techniques began in the 1970s and have become increasingly well established. Typically, they have involved getting students to evaluate the work of their peers, in small groups. This has become so widespread that a number of meta-studies have been undertaken to evaluate the process. In the era before LMSs (Learning Management Systems), such interventions were quite cumbersome, involving academics in distributing/collating lots of paper based responses. More recently, VLE (Virtual Learning Environment) functionality and dedicated online systems have made the act of introducing and managing peer assessment much more practicable.

However, further technological advances, particularly EVS (Electronic Voting Systems) and polling applications, means that today peer assessment can be achieved instantaneously, even with just a smartphone and rating platforms like *Poll Everywhere (2017).* This kind of peer assessment, however, is very different to what has gone on before owing to its immediacy and convenience. The fact that such platforms enable *instantaneous* and highly *scale-able* peer feedback to be obtained (for instance, getting feedback from over 200 students in a lecture theatre), means we need to distinguish it from other forms of peer assessment. Accordingly, from now on I will call it Live Peer Assessment (LPA).

My interest in LPA has been in order to:

1. Get students to evaluate the work of peers and thereby, become more competent evaluators of their own work;
2. Enable a more effective dialogue about quality in multimedia assignments;
3. To generate much faster feedback.

The objectives, therefore, of this research are:

1. To survey current thinking in the area of peer assessment by means of a literature survey;
2. To examine the use of EVS and LPA in a UK University;
3. To develop guidelines for the use of EVS and LPA;
4. To understand the potential benefits and limitations of this approach.

The overarching research questions are;

- **RQ1** How might EVS and LPA be designed and implemented?
- **RQ2** How might EVS and LPA be evaluated?
- **RQ3** What are the benefits and limitations of using LPA and EVS?
- **RQ4** How does LPA work?

The following are sub questions relating to these research questions.

RQ1 What kinds of rubrics or marking sheets are best used?
RQ1 What practices can be used to enhance the reliability of marks in this modality?
RQ1 What considerations should be borne in mind when choosing exemplars for students to mark?
RQ2 Can LPA improve marks in an assignment?
RQ2 If marks improve in an assignment, how is that improvement distributed across the cohort?
RQ2 What measures exist to quantify the success of an LPA event?
RQ2 Are students sufficiently competent to mark their peers under these conditions and are they more or less so when compared to more traditional forms of peer assessment?
RQ3 What pedagogical benefits does this kind of peer-assessment bring with it?
RQ3 Do students become more competent in judging quality in the work of others and in their own work when engaging in LPA and EVS?
RQ3 What are the limitations of this technique?
RQ4 Does the act of assessing the impact on the way students set about their own academic work?
RQ4 What might be the explanation for the effects claimed for LPA in this study?
RQ4 What are the students' opinion and feelings about the LPA process?

These questions are addressed through a study of two courses over four iterations. For each course, two of the iterations precede the start of this PhD and the final two iterations take place during the PhD. Therefore what this study is examining is pedagogical practice which was well established well before any thought of deeper examination was contemplated. The successful implementation and striking results encountered in those early iterations was what gave me the inspiration to seek more complete understanding of what was occurring. As a result, there is no real change in practice beyond refinement of the details in the final two iterations of each course. However, in order to understand how students experienced these techniques, two focus groups were conducted one in each course (carried out under ethics protocols ***1112/51*** and ***COM SF UH 00014***– see appendix).

The first of these is *Multimedia Specification Design and Production,* a 30 credit masters course. In this course, LPA was used both summatively (on assessees) and in formative exercises (where students learned how to grade others by marking previous students' work). On this course, students in groups produced multimedia artefacts in both prototype and final forms, with a proportion of their marks coming from the average score given by their peers. The other is *E-Media Design,* a 15 credit 1st year BSc course. On this course, the students did not mark their peers work, but rather, participated in exercises marking previous students' work. For the first three iterations, they received some credit for how similarly they marked to the tutors. That is to say, it was in a limited way, summative (on the assessor). This meant the marking done by any student on this course did not impact on the marks of any other student in any way.

The Masters' course involved students marking their peers' group work. That is to say, students developed their artefacts in groups of two or three, and the whole of the class marked them at a demonstration. This course had a varying enrolment between 60-30 students and so it would be difficult to draw many conclusions regarding the quality of student work over different iterations. However, the BSc course had enrolments varying between 180-240 students and did individual work and so on this course it is much easier to compare the quality of work between iterations.

The structure of the thesis is as follows.

- Chapter 1 – Introduction
- Chapter 2 - Peer Assessment Literature Survey

  This is an extensive evaluation of the research literature relating to all 4 research questions. It covers Topping's work on the variety of peer-assessment practices (RQ1) as well as the measures of peer-assessment practice propounded by Falchikov and Goldfinch and their successors (RQ2). The claimed benefits of peer assessment in the literature are also considered as well as critiques of its practice being provided (RQ3/4). In terms of LPA, I also discuss other examples of this technique as well as analogous techniques – such as the studio "Crit" used in the design sciences (RQ1).
- Chapter 3 – Multimedia Specification Masters Course: History and Results

  The structure of this course remained remarkably uniform over the four iterations, but certain aspects of each iteration brought out the particular strengths and weaknesses of LPA. (RQ3 and RQ 4). The rubric used and the difference between it and its predecessors is explained (RQ1). Moreover, the influence of rehearsal sessions and their value in the

process are covered (also RQ1). Statistical measures consistent with Falchikov and Goldfinch are used as well as new ones are promoted (RQ2).

- Chapter 4 - Multimedia Specification Masters Course: Focus Group

  In these discussions, I investigated how the students experienced LPA. The aim of this focus group was to discover how LPA works and why it may have affordances that make this form of peer assessment more reliable than others. (RQ3 and RQ4). Based on the discussion, I also return to the marking data and examine whether there was any evidence of reciprocated marking (one student agreeing to mark another's work higher in return for doing the same). This is also numerically measured (RQ2). Also, techniques to enhance the reliability of student marks are considered (RQ1).

- Chapter 5 – E Media Design: Chronology

  The E-Media Design course was much more variable in terms of the rubrics used and the incentives offered to make students participate in the peer assessment process. Since this is complicated I have devoted a chapter to it which primarily deals with the logistics of its implementation over the focal four years (RQ1).

- Chapter 6 – E Media Design: Measures of Improvement

  The E-Media Design course average for the final assignment improved by approximately 13% over the studied four years (RQ3). In this chapter, I analyse the cohort by percentiles to see which range of the cohort demonstrated greatest improvement. Also, what practices might have contributed to the results of particular years is investigated (RQ1)

- Chapter 7 – E Media Design: History and Results

  This is an in-depth look at the voting data over the four years, with the aim of determining whether certain criteria are inherently more difficult or easy than others (RQ1). Other measures of student voting patterns are considered (RQ2) along with how long it took for students to make judgements with particular criteria and artefacts (RQ2). Finally, the effect of the training set on the voting patterns of the students is examined (RQ1).

- Chapter 8 – E Media Design: Focus Group and Internalisation of Quality Standards

  This chapter is based on a focus group discussion and addresses the question as to why student artefacts, on average, improved over the four iterations. Through the

discussion with the students I investigate how they went about their own work after having participated in the LPA events (RQ3 and RQ4).

- Chapter 9 – E-Media Design: The Inspirational Element

  This chapter deals with a concept that might also be contributing to higher student achievement, namely "inspiration". I pursue this by looking into whether any properties of the exemplar training set found their way into the work produced by the students (RQ3).

- Chapter 10 – Discussion and Summary

  This chapter relates the findings and observations of the study to items of interest to computer science academics, including: MOOCs, social equity in computer science, and academic practice. (RQ3 and RQ4)

- Chapter 11 – Conclusion

  In this chapter, I return to the initial research questions and address them one by one.

Chapter 2. Peer Assessment and Feed-Forward: Literature Review

In the previous chapter, I explained that the goal of the research was to complete a longitudinal study on the effects of live peer assessment in two university courses. In order to understand the issues involved, a literature survey on peer assessment was undertaken. This was both to inform practice and to establish the conceptual frame of the research.

Peer Assessment has become a more recognised feature of university education recently, although its history is actually quite long. Arguably, the first pedagogical writing regarding this practice was that of George Jardine, who was professor of logic and philosophy at the University of Glasgow from 1774 to 1826, who extolled the practice of peer tutoring and got students to work in groups and edit each other's essays (Gaillet, 1992). However, its first mention in recent university education probably occurred in 1959, with Hammond and Kern (1959) emphasising the value of self-assessment in medical education. Specific reference to the value of peer assessment first occurs in Schumacher (1964), where he noted that peer assessment could be used to measure medical skill and also, unlike traditional measuring practices, "skill in relationships". This finding was echoed seven years later when Korman and Stubblefield (1971) found that peer ratings were the most accurate predictor of subsequent intern performance.

The next very significant paper was written by Boud and Tryee (1980), which recorded that there was substantial correlation between the marks of tutors and peer evaluated marks applied to students in a law course. Law, like medicine, is a very established profession, where peer evaluations of colleagues is commonplace and therefore, peer assessment might be considered suitable in training new practitioners. A similar level of agreement (relating to correlation between tutor and student marking) was found by Morton and MacBeth (1977), who investigated medical students rating each other. The fact that very early in the practice of university peer-assessment there was some need to prove the validity of peer-awarded marks is significant. After all, how could one responsibly ask students to receive the opinions of their peers, if one did not have faith that they would be fair and representative?

At this time, self-assessment was as important as peer assessment and one significant meta-study about student self-assessment was published in 1989 by David Boud and Nancy Falchikov (1989). This established a number of techniques for comparing different studies which would also re-emerge in Falchikov and Goldfinch's meta-study on peer assessment some 10 years later; essentially, these

pertained to correlation coefficients between student awarded marks and tutor awarded marks as well as effect size (Cohen's D). Regarding this, they found that the mean correlation coefficient between staff assessment and student self-assessment was 0.39.

2.1 Meta-Studies of Peer Assessment

Up to now, there have been five significant meta-studies on peer assessment:

- *Peer Assessment between Students in Colleges and Universities* Keith Topping (Topping, 1998);
- *User of Self Peer and Co Assessment in Higher Education* (Dochy, Segers, & Sluijsmans, 1999);
- *Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks* (Falchikov & Goldfinch, 2000);
- *Effective peer assessment processes: Research findings and future directions* (Van Zundert, Sluijsmans, & Van Merriënboer, 2010);
- *Peer assessment in the digital age: a meta-analysis comparing peer and teacher ratings(Li et al., 2015).*

2.1.1 Peer Assessment between Students in Colleges and Universities – (Keith Topping)

Topping examined 31 papers, and his paper is a broad summary of findings, largely qualitative, with its major contribution being a very rigorous classification system of the different ways in which peer assessment is undertaken. Typical modes include peer ratings between members of a group – typically done on a one to one basis (each person rates one other). Sometimes one person rates a small number of others. As well as these, Topping included categories such as weighting (summative or not), anonymity, assessors per assessee and many others. This classification system will be used for reviews of some other studies, as well as the studies in this report. Topping was also very good at critically analysing the claims made by a number of papers and synthesising them into a compact overview.

2.1.2 *User of Self Peer and Co Assessment in Higher Education* - (Dochy, Segers, & Sluijsmans)

Dochy et al. (1999) probed 63 papers covering both self and peer assessment (the fact that they covered self assessment also accounts for the larger number of papers). Their main findings were that, if implemented properly, one could expect agreement between tutor and student marking and also, that it would have positive effects on students. They associated the need for peer assessment with the needs of the labour market for more self-reflective learners capable of problem solving.

2.1.3 Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks ((Falchikov and Goldfinch)

Falchikov and Goldfinch (2000) covered 48 papers and involved a much greater statistical rigour than the previous two studies, with no study of similar ambition being attempted until Li et al. (2015) . This was a very significant meta-study, which attempted a synthesis of data to establish the validity of peer assessment. They focused particularly on two measures to compare tutor and student marking patterns: the r (correlation coefficient), and d (the effect size). In broad terms, r represents the level of similarity of the marking patterns of tutor(s) and student(s), whereas d measures the similarity of marking weights. In other words, r can tell if there is a correlation between marks given by different markers, whilst d can tell if one group of markers grades more or less generously than another. Synthesising and putting the data of the 48 studies together, they found an average r (correlation coefficient between tutors and students) of 0.69. Notably, this is much higher than the average correlation coefficient reported in the meta-study of self-assessment papers (0.39). They elicited that the average d (effect size) to be 0.24. Overall, these are very positive figures for justifying the overall validity of peer assessment.

However, within these data, they also highlighted a number of other variables influencing the level of agreement between tutors and peers. One of the codings used to differentiate the studies was the marking instrument used - based on "dimensionality" vs "globality" -, by which they meant how structured the marking instrument was. The three categories in this regard are:

1. G - meaning global - that is to say, students and staff have to award an assessee a particular overall score without much concentration on criteria or attainment descriptors;
2. G+ - means that students award a global score, but with strong guidance as to the considerations or criteria for giving an award;
3. D - means explicitly dimensioned, namely, that students follow a rubric of various criteria with potentially attainment descriptors for each.

There findings in this regard, were that the G+ types of peer assessment exercises yield the highest correlation coefficients and that D type rubrics the least. Nonetheless, this division itself might be too broad-brush since not all such marking instruments would fall easily into one of the three categories. As an example, the marking procedure described by Magin (1993) could be said to be both G+ or D, whilst

Falchikov and Goldfinch would place that paper in the G+ category. Magin's paper contained two cases studies, one of which asked students to rate other students based on *contribution to discussion* and *contribution to development of group.* The other case study involved seminar presentations, where students were subsequently rated on six criteria, broadly summarised as evidence of reading, evidence of understanding, organisation of time, creativity and variety, facilitation of student participation and critical evaluation). Each of these had to be marked from 1 (inadequate) through 3 (satisfactory) to 5 (outstanding). This appears fairly strictly dimensioned, yet it was coded under G+ by Falchikov and Goldfinch.

The other interesting findings in the paper are that cohort size (the constellation involving a large number of assessors per assessee) does not improve the level of agreement, and potentially might lead to lesser correlations. This, however, may just be a reflection of the logistical issues of having multiple assessors in 1998 (when the meta-study was undertaken). Brown and Knight (1994) also wrote:

> One danger arises in the sheer mathematics of multiple assessors. If 20 or so groups assess each other and two categories of staff are also involved, and there is an element of oral assessment too (eg, 'How effective were the group in answering questions on the poster?'), then the processing of all the assessment sheets can be a nightmare.

At the time of writing of the current study, where there are many technologies, such as EVS clickers as well as dedicated polling apps, these logistical issues, which could have impacted on the processes under consideration by Falchikov and Goldfinch may be no longer a problem. Another finding is that the rating of academic products (e.g posters and presentations) is more likely to lead to high agreement than the rating of professional practice. This may be down to the more subjective criteria in operation. A number of papers - Orsmond, Merry, and Reiling (1996), Sitthiworachart and Joy (2004) - have also demonstrate different levels of agreement amongst different criteria.

Falchikov and Goldfinch undoubtedly brought greater rigour to the study of peer assessment. In their paper, they criticise approaches which demonstrate validity by "percentage of agreement" between tutor and student marks, where "agreement" could be interpreted sometimes extremely strictly, whilst at other times extremely loosely. Similarly, their insistence on the effect size variable (Cohen's D) as being an important measure, meant that validity would not merely be demonstrated by similarity of marking pattern, regardless of the scores given, for it would also concern whether marks awarded are over-generous or the opposite.

However, a number of the other recommendations made at the end of the paper might be regarded as unduly prescriptive. For instance, their promoting of criteria informed global scores above explicitly dimensioned criteria, was made from a comparison of a small number of studies reporting different levels of correlation. They also recommended avoiding large numbers of assessees per assessor, but did not offering much in the way of a proof for this. However, their emphases, assessor population size, clarity of criteria, holistic verses granular criteria and effect size as a supplementary measure of validity, were important considerations for practice in this study. Falchikov and Goldfinch, while justifying the validity of peer assessment, did not seek any way of quantifying the benefits for students doing so. At the beginning of their study, they cited Vygotsky and social constructionism and stated that peer assessment will promote learning. However, they provide little concrete evidence for that assertion.

2.1.4 Effective peer assessment processes: Research findings and future directions (Van Zundert et al., 2010)

The next meta-study did indeed, attempt to find out if peer assessment (PA) promoted learning, and also to find out what factors are instrumental in this. In their words, they sought "to investigate how PA conditions, methods and outcomes are related" and grouped its findings across four headings:

1. Psychometric Qualities Of Peer Assessment
   (by this is meant accuracy and reliability of peer marks);
2. Domain-Specific Skill
   (by this is meant improved performance in the field where the peer assessment was used);
3. Peer Assessment Skill
   (by this is meant the differential abilities of different peer assessors and the factors contributing to this);
4. Students' Views Of Peer Assessment
   (by this is meant students overall attitudes towards the process).

In terms of *Psychometric Qualities* (in basic terms validity of judgement), this meta-study did not say anything not already said by Falchikov and Goldfinch. In terms of *Domain specific skill,* the authors pointed to a number of studies where students did a draft of some assessment, received peer feedback, and then redeveloped their work, which usually resulted in a higher score and the authors attributed this to students acting on the feedback received. In terms of *Peer Assessment Skill*, the authors pointed to better outcomes happening when students received training in peer assessment. Other findings were that high achieving students appeared to give better feedback, and that students with "high executive

thinking" (generally a willingness to follow the instruction and guidance of the tutors, rather than being more independent) appeared to give better feedback. Among *Students' Views of Peer Assessment* they found greater acceptance of peer assessment among students who had been trained in it and a more positive orientation to study. However, they also found a number of studies where students expressed negative opinions of peer assessment.

2.1.5 Peer assessment in the digital age: a meta-analysis comparing peer and teacher ratings (Li et al)

The most recent meta study by Li et al. (2015) involved doing very much the same as Falchikov and Goldfinch, namely, to cover the correlation between tutor and student marking. However, they covered a larger number of studies (70). They also described a more sophisticated technique for aggregating the various reported correlation coefficients into an overall figure. They found that the average correlation between tutor and student awarded marks is 0.63 (very comparable to Falchikov and Goldfinch's figure of 0.69). They also made a number of other observations, contending higher correlations occur when the following is the case.

1	the peer assessment is paper-based rather than computer-assisted;
2	the subject area is not medical/clinical;
3	the course is graduate level rather than undergraduate or K-12;
4	individual work instead of group work is assessed;
5	the assessors and assessees are matched at random;
6	the peer assessment is voluntary instead of compulsory;
7	the peer assessment is non-anonymous;
8	peer raters provide both scores and qualitative comments instead of only scores
9	peer raters are involved in developing the rating criteria

Table 2-1:Li et al.'s list of conditions where higher correlations occur between tutor and student marking – (my numberings)

They believed the paper-based finding (item 1 in Table 2-1) may be due to the immaturity of some web based systems for conducting peer-assessment and also, that paper-based versions might make the exercise less casual (this might also explain higher correlations when qualitative comments were added - see item 8 in Table 2-1.) Also, the finding that correlations improve when participation is voluntary might also mean that participation is more serious (unmotivated students are not participating). However, the danger in having only voluntary participation is that it might mean that those who would most benefit from participating in peer assessment might not do choose to so. The finding that correlations are higher when the subject area is not clinical (item 2 in Table 2-1) relates to the difficulty of judging professional practice (and echoes Falchikov and Goldfinch's findings) and more generally, the

difficulty of evaluating more subjective criteria. That individual rather than group work receives higher correlations (item 4) is intriguing and the authors did not attempt to explain why this is so.

This paper is valuable in that it confirms the typical correlation between tutor and student marks that had been established in Falchikov and Goldfinch. They also pointed to the different variations based on the maturity of the assessors, and the level of their buy in to the process. However, just like the other meta-studies here, there is not much in terms of the effects, that is to say, the beneficial outcomes that can be said to arise from peer assessment.

2.2 Synthesising the Findings

In order to synthesise the findings of these studies, to identify those most relevant to the study being undertaken here, I summarise them under four headings: Perception of Fairness, Measures of Accuracy, Claimed Effects and Logistics and Constellations.

- **Perception of Fairness** is typically the student's perception of the honesty and validity of the process. In this study, this is examined in the focus group with the masters students (chapter 4).
- **Measures of Accuracy**, involve the *r (*Correlation Coefficient) and *d (Cohen's d for Effect Size)* cited by Falchikov and Goldfinch, but also other measures, such as pairwise Kendall's Tau (Orpen, 1982) and also Paired T-Tests (Cheng & Warren, 2005). The coverage of the marking on the MSc course in terms of *r* and *d* is examined.
- **Claimed Effects** relate to the benefits or otherwise the practice brings to the student, both in terms of higher scores in subsequent assignments, or in terms of attitude to learning. This is investigated in the scores for the assignments on the BSc course.
- **Logistics and Constellations** relates to how peer assessment needs to be organised in order to be successful (prior training in peer assessment, negotiation of criteria, summative vs formative) as well as who marks who and number of assessors per assessee. The logistics are very specific to the method used in this study (EVS clickers) and I will attempt to convey what I believe to be the most effective methods for doing this.

These four factors do interact. The *Perception of Fairness* registered by students will be related to the *Logistics and Constellations* (the type and wording of rubric, whether staff moderation is undertaken, training in peer assessment, anonymity or not, number of assessors per assesse and the selection of who rates who) and also, to the real *Measures of Accuracy* (the general correlations and effect sizes

between tutor and student scores in any assessment, the presence or otherwise of evidence of non-academic marking).

2.2.1 Perception of Fairness

Fairness is a perennial concern and the anxiety it causes students is never completely assuaged by how many studies demonstrate high correlation between tutor and students' marks. As explained earlier, Falchikov and Goldfinch found the average correlation across all studies between tutor awarded marks and student awarded marks to be 0.69, which is substantial. However, this correlation figure describes the coincidence of patterns of marking and cannot take account of odd occasions of injustice. Moreover, a high correlation is not necessarily an indication of fairness. Highly correlated marks between two markers with highly discrepant standard deviations (where the range between the highest and lowest may be larger) would mean that potentially high performing students are not given sufficient credit for their more accomplished work when marked by markers with a narrow range. Related to this is the disinclination of students to penalise their peers. Finally, there is the fear of abusive marking. Brown and Knight, (1994) establish a typology of non-academic influences on student marking in the context of groups of project students rating each other. They particularly mention friendship marking and collusive marking. Two other factors, namely, decibel marking (individuals dominating groups and getting high marks as a result) and parasite marking (students benefit who do not participate) are only relevant to project intra-group peer evaluation and so, are beyond the focus of this study. Mathews (1994) also attempted to construct a primitive typology of marking styles (*flat*, *normal*, *finger-pointing*, *stitch up* and *out of kilter*). Being aware of these is useful for preventing any distortions that might occur in peer assessment, however, being too vigilant and trusting the students too little can also bring about its own distortion. Typically, most peer assessment practice involves some process of moderation by the tutors, such that unfair or malicious marking can be altered by the monitoring tutor. Other factors influencing the perception of fairness is training in peer assessment. If students can see its operation in non-summative contexts, they are more likely to accept it in summative ones.

2.2.2 Measures of Accuracy

Accuracy has tended to concentrate on comparison between tutor and student marks. The major measure is the correlation coefficient, although this figure is often presented uncritically in small sample sizes (Boud 1980 for example) and at times can be presented with eye-catching results, which seem impossible to justify, for instance, in Burnett & Cavaye (1980) in which a mean correlation of 0.99 between tutor and students was reported. Another common concern is under-marking or over-marking.

Freeman (1995) found a high correlation between student and staff marks, but also found evidence of student under-marking of good presentations and over-marking of poor ones (a general compression of the marking). Orsmond et al. (1996) and Stefani (1992) also demonstrated similar effects, thus suggesting that whilst averages and correlations seem to be consistent, often standard deviations are lower. However, this may be little more than an obvious statistical effect. Simply the fact of averaging multiple marks will lead to a more compressed distribution (if there are the averages of many student ratings being compared with a single tutor rating), rather than being the result of different marking patterns. This statistical effect, however, will also feed into worries about fairness (among high achieving students), who may find the excellence of their work only occasions a small premium in their marks relative to the rest of the class.

Topping (1998) looked at 31 studies concerning the reliability of peer assessment. Eighteen of these studies reported high reliability, with a tendency for peer marks to cluster around the median noted. He also found seven studies with low reliability. An early paper by Orpen (1982) had another mechanism, creating pairs of Kendall's coefficient of concordance as a measure of association. Among 21 comparisons between tutor and individual student marks few demonstrated a significant difference. Another measure is paired t tests between tutor markings and student markings (Cheng and Warren, 2005).

At this point, it is worth sounding a note of caution, particularly because there is likely to be a publication bias in favour of studies that demonstrate high-correlation and substantial agreement between tutor and student marks. Studies reporting disappointing results are few; one of these, by Bostock (2000), reported a 0.45 correlation between student and tutor marks. Swanson, Case, and van der Vleuten (1991) reported students giving uniformly high ratings to other students, and when some forced ranking was introduced in order to stop this, it elicited so much resistance that the practice had to be discontinued. However, this paper itself takes a negative stance regarding peer evaluation, confusing it with self-evaluation in its summary of research and therefore, the negative outcomes may merely have reflected the negative assumptions of the authors. What we can say is that among those studies carried out more or less successfully, there are a number of interesting commonalities and one of the most compelling is the *r* (correlation value), very often being between 0.6 and 0.7, which has been subsequently confirmed by the later meta-study undertaken by Li et al. (2015)

2.2.3 Claimed Effects

This can be described as being the change in learning and also learning self-image caused by peer assessment. Dochy et al. (1999) suggest that increased student self-efficacy, awareness of academic quality, reflection, performance, effectiveness, responsibility, students' satisfaction and an improvement in learning culture are all potential positive effects of peer assessment. Topping particularly focused on cognition and metacognition, believing that PA could lead to increased time on task (evaluating). He also suggests that students would be better able to measure deviations from the ideal, meaning some kind of norm referencing– enabling a student to locate himself or herself in relation to the performance of peers and to prescribed learning targets and deadlines. Sluijsmans, Moerkerke, Van Merrienboer, and Dochy (2001) suggested that students acquire a better self-perception of their own assessment skill as a result of participating in peer assessment, whilst Liu and Tsai (2005) reported students saying that it had helped their learning. In terms of quantitative measures of increased learning or performance, (Bloxham & West, 2004) also commended peer assessment for giving students an understanding of the assessment regime and giving them a clearer sense of the "rules of the game".

When approaching the issue of the *benefits* of peer assessment, as opposed the *validity*, the previously highly numeric and quantified evidence gives way to a vaguer and less provable set of propositions. Regarding this, Pope (2005) stated that "Some of these benefits can give the impression of being nebulous". Most of the time the proof comes from questionnaire feedback from students. In many studies, such research into student attitudes covers mainly feelings of support or resistance to the practice, whilst elsewhere there is evidence where one can clearly see some educational benefit.

Gielen, Dochy, and Onghena (2011) argued that students might gain fresh ideas from seeing other students' work. Topping contended that feedback from fellow students might be more comprehensible, since it comes from their peers. Higgins (2000) suggested that the power imbalance between tutors and students might mean that students exhibit a 'emotion-defence system' when they receive feedback from tutors, which they might not experience when it is received from peers. Peer feedback is also likely to be released more quickly and therefore, whatever limitations there might be in the marking capability of the assessor, the timeliness of the feedback will be some compensation for it (Gibbs & Simpson, 2004). Pryor and Lubisi (2002) suggested that having to evaluate other students and express those evaluations makes them cognitively operate at an evaluative level and to pose metacognitive questions. Stiggins(1987) said "'Once students internalise performance criteria and see how those criteria come into play in their own and each other's performance, students often become better performers".

Hanrahan and Isaacs (2001) identify a recurring theme of students' feedback to peer assessment as being "productive self-critique". They quoted one student in their study saying "You realise what markers are looking for (a new experience for me and very valuable) and are forced to acknowledge whether or not the factors which must be in your essay are present". Sambell, McDowell, and Brown (1997) suggested that peer assessment can help students with self-assessment: "The experience of being a peer assessor can be considered as a precursor to becoming a skilled self-assessor". As can be seen, the predominant method of data collection used to assess whether PA increases student learning has been self-report.

2.2.4 Logistics and Constellations

This relates to what is necessary for successful peer evaluation. Cheng and Warren (2005) persuasively argued for the need for a lot of training for students, if they are to become comfortable with peer evaluation. They cite Williams (1992) and Forde (1996) giving substantial induction in peer evaluation to head off the kinds of anxieties that are often expressed in surveys of student attitudes. Further logistical considerations relate to the marking instrument used. Falchikov and Goldfinch (2000) found evidence that the more specific the marking rubric, the less likely it is to result in high correlation.

As mentioned before, it is Topping who established a rigorous classification system for types of peer assessment, which is probably the most relevant for the analysis of PA logistics. The key terms for this study are:

- Focus (quantitative/qualitative);
- Product (what is being assessed);
- Relation to staff assessment (substitutional or supplementary);
- Official weight (contributes to grade or not);
- Directionality (one-way, reciprocal or mutual);
- Privacy (anonymous or confidential);
- Contact (distant or face to face);
- Constellation Assessors (individual or group);
- Constellation Assessed (individual or group),
- Place and Time
- Compulsory or Voluntary.

While Topping's classification system works well for most of the implementations of peer assessment, some recent papers on "Comparative Judgement" (CJ) have demonstrated a method that is difficult to classify according to his schemes. In this process students are merely given pairs of students' work and asked to say which the best is: a process by which (after a number of iterations and a suitable aggregating function) an overall ranking can be produced. Studies of this kind include those of Jones and Alcock (2014), Pollitt (2012) and Seery, Canty, and Phelan (2012) The theory behind this approach is best expressed in Pachur and Olsson (2012).

Thus far, the broad literature in relation to PA across all subjects and disciplines has been covered. Some initial conclusions can be made including:

- student and tutor marks are likely to be more convergent around academic products rather than processes,
- there are ways of measuring convergence,
- that certain criteria when applied may produce greater agreement than others and also
- some measures for checking the presence of non-academic factors in peer marks.

There have been many claims about the benefits of peer assessment, but evidently the level of rigour in the associated studies is questionable in many cases. Moreover, there would appear to be a wide variety of types of PA currently practiced in the university sector across the world. At this point, the focus is turned towards the field of computer science.

2.3 Peer Asssessment in Computer Science

The same drivers that led to PA in other disciplines, can also be argued to exist in computer science and IT. Firstly, peer evaluation of code is a well-established quality control mechanism in software enterprises. Secondly, a whole framework of programming, XP (Extreme Programming) has core principles of collaboration and buddy programming, and involves an almost continuous process of peer review (Beck, 2000). Additionally, like all other subjects, the discipline has had to confront the difficulty of high staff student ratios and the need for smarter forms of assessment to generate sufficient and appropriate feedback.

2.3.1 Peer Marking – Study by Sitthiworachart and Joy

An interesting study was undertaken in by Sitthiworachart and Joy (2004), who established an online system where students could write simple Unix programs, which would then be marked by a peer,

whose marking would then be evaluated by the recipient. Using Topping's characterisation of practice the following describes the process that these authors put in place.

Table 2-2: Topping's Peer Assessment Classification applied to (Sitthiworachart & Joy, 2004)

Focus (quantitative/qualitative)	Summative quantitative/qualitative
Product (what is being assessed)	Simple unix code scripts
Relation to staff assessment (substitutional or supplementary)	Not clear
Official weight (contributes to grade or not)	Yes
Directionality (one-way, reciprocal or mutual)	Mutual (assessee is marked by assessor whose assessments are also assessed by the assessee)
Privacy (anonymous or confidential)	Anonymous
Contact (distance or face to face)	Anonymous distance
Constellation assessors (individual or groups)	Another student
Constellation assessed(group)	Individual
Place and time	Online

The main objective of the paper was to evaluate the reliability of peer marks on a criteria by criteria basis. Given it was a study of 213 students its statistical outcomes are likely to have been reliable, and what they did reveal was big disparities in correlation values (between tutor and student marks), according to the different criteria. Some very straightforward criteria generated high levels of correlation, but the more subjective ones, for instance, "Easy to follow what the program does", had a much weaker correlation (It is also worth bearing in mind that this study was about first year students). Variation in agreement per criterion was also noted by Orsmond et al. (1996).

Sitthiworachart and Joy's table of correlations by assignment and criteria for 165 students is impressive.

Table 2-3: Sitthiworachart and Joy: tutor student correlations by assignment and marking criteria

Marking Criteria	Pearson Correlation		
	Assignment 1	Assignment 2	Assignment 3
Readability			
1. The Number of Comments	0.853*	0.613*	0.630*
2. The Helpfulness of Comments	0.862*	0.548*	0.470*
3. Appropriate Indented Code	0.620*	0.450*	0.281*
4. Appropriate Variable/Function Names	0.552*	0.694*	0.324*
Correctness			

1. The Program Meets the Specification		0.668*	0.618*
2. Appropriate Code Handles Errors		0.638*	0.789*
3. The Program Finishes with Appropriate Exit Status	0.672*	0.661*	0.492*
Style			
1. Appropriate utilities have been selected	0.380*	0.655*	0.628*
2. Good Programme Selection	0.388*	0.501*	0.351*
3. Easy To Follow What The Program Does	0.478*	0.472*	0.190**
*Correlation is significant at the 0.01 level (two tailed). **Correlation is significant that the 0.05 level (two tailed).			

According to Sitthiworachart and Joy, in terms of correlations, they believe:

- to 0.20 Negligible
- 0.20 to 0.40 Low
- 0.40 to 0.60 Moderate
- 0.60 to 0.80 Substantial
- 0.80–1.00 High to very high

and therefore, most of the agreement between staff and student ratings is moderate to substantial. However, clearly the more subjective categories produce lesser agreement.

2.3.2 John Hamer and Aropa

The other practitioners examining PA in computer science were Hamer et al. (2007) with a system called Aropa. In their intervention, there was routine use of PA among approximately 1,000 students enrolled at different times during a year of introductory programming courses. In this intervention, students had to submit code for a programming assignment and then review six submissions from peers. This study is particularly interesting from the point of view of the attitudes of the students and self-reporting of the effects of participation (no data in terms of assignment scores is provided). However, the kinds of benefits in terms of metacognition and sense of quality, mentioned earlier can easily be seen in some of the comments of students, for instance:

> "I learned very quickly the mistakes that I had made in my own programs, when marking the other people's work …It aided in the learning of certain aspects such as style and code conventions." (Hamer et al., 2007)

This is a rich pair of comments indicating greater subject specific awareness and also, some kind of internalisation (style and code conventions). These kinds of things can often be instructed, but are more likely to be internalised when examples of capable peers are given.

Another comment was:

> "I didn't realise the value of good comments and easy-to-understand code till I got a bunch of hard-to-understand code to mark." (Hamer et al., 2007)

This is a very interesting example of a student being able to connect a concrete example (submitted code) to an abstract concept ("good comments" and "easy-to-understand code"). Hamer et al. quantified the benefits the students reported and associated the number of comments per benefit. In the table below (Table 2-4: Hamer's table of comments with my annotations), I have attempted to further codify the responses by the causation of the benefit – did it arise from marking (M) or being marked (BM) or both (MBM).

Table 2-4: Hamer's table of comments with my annotations

Columns Directly From Hamer		***My annotations***
N Comments	**Type of comment**	***Causation***
40	exposure to a variety of coding styles	*M*
32	learning examples of good coding	*M*
32	non-specific positive comment	*BM*
35	helpful feedback received	*BM*
24	the system was convenient and easy to use	*N/A*
20	learning to identify poor programming constructs and mistakes	*M*
19	improving (debugging) their own code for this assignment	*MBM*
19	comparing own performance to peers	*M*
15	helpful reading code	*M*
10	helping others by giving feedback	*M*
5	learning by marking	*M*
2	the exercise motivated them to work harder	*N/A*
2	gaining an insight into the marking process	*M*
2	anonymity relieved concerns about fairnes	*N/A*

One very clear finding here is that *the act of marking* accounts for the overwhelming majority of comments and *the act of being marked* much fewer. Meaning it is not so much the receiving of feedback as the giving of it that appears to constitute the transformative nature of this assignment. Clearly, one of the most important features was the normative nature of the exercise. Students could

position themselves within the known range of abilities demonstrated in the class. The exposure to coding styles, examples of good coding, comparing performance and the benefits from reading code all relate to the students being able to experience their own work in the light of the work done by their peers. This kind of normative feedback seems to be one of the most powerful forms that students can get from the process.

2.4 Peer Assessment Technology

Earlier I cited Brown and Knight (1994), who talked about the logistical challenge of getting a large number of peer assessors together and averaging their marks. The two computer science based papers reviewed above used web based systems to facilitate peer assessment by very large cohorts, involving a systematic processing of peer grading data. Nonetheless, they were still largely asynchronous practices, carried out in isolation by students whose marks were later aggregated. More recent technology, namely, classroom EVS clickers allow for instant polling to take place with the aggregated scores visible immediately after voting.

Papers written so far using these technologies include Barwell and Walker (2009) and Räihä, Ovaska, and Ferro (2008) In both cases, the technology was used to gather the marks of other students and the results of marking were shown later to those being marked. In other words, it was not synchronous. Vanderhoven et al. (2015) contrasted 15 and 16 year olds using clickers as opposed to visible scorecards to see if the anonymity of marking with clickers affected the marks given and they found that they did. Raes, Vanderhoven, and Schellens (2015) produced a paper comparing student comfort and satisfaction using clickers as opposed to giving oral or written feedback. These latter two papers seemed to suggest that the condition of anonymity for the *assessor* meant that students could concentrate more on marking according to academic concerns, rather than non-academic ones. However, the effect on the assessees (who were not anonymous) was not calculated.

The practice of live public assessment, however, is not something completely new driven only by technology, for it has been a core part of assessment in the design disciplines since the 1830s in the Ecole des Beaux Arts in Paris, through an assessment procedure known as the Design Crit.

2.5 The Design Crit

The Crit essentially stands for critique by jury; also termed “design review”. Typically, students present sketches of their response to an assignment in front of a small panel of tutors and an audience of their peers and describe the ideas that underpin their work. Tutors give feedback on the design and make

suggestions for ways in which it could be further improved. Crits occur throughout a project and constitute 'the ceremonial culmination of each studio design project' (Lewis, 1998) at the end. Blythman, Orr, and Blair (2007) wrote a report on the Crit in UK education in 2007, regarding which they flagged up a number of advantages. Probably the two most relevant to this study are:

> "Everyone gets a chance to see each other's work. This is important now that students work less in studios, often do not have their own spaces/designated studios and may not have suitable spaces e.g. studios filled with tables and chairs, or no computers in studio….
> Students see that staff have a variety of perspectives and can have apparently contradictory positions and show disagreement between staff in crits. This is important since this shows there is not just 'one true way'."

As in Hamer's study, the mere exposure to the variety of their peers' work can not only give students a sense of the variety of approaches available, but also a sense of their own position within the class. However, the point about "contradictory positions" does show the difficulty of articulating and also recognising clear standards of quality in more subjective disciplines. In these disciplines the manifestation of such "contradictory positions" can be a source of contention.

This has been highlighted in papers by Christine Percy (Percy, 2004) and Charlie Smith (Smith, 2011), specifically in terms of the effects on students of arbitrariness in tutor opinion. Smith, in a series of focus groups, found students experiencing the Crit as an adversarial experience, finding criteria being used which was not in the project brief being applied, and feeling so defensive that they were unable to remember any of the points that came up. Smith particularly highlights students' calls for seeing examples of previous work so as to have a better understanding of what is required. He also suggested student led Crits might work as an antidote to the asymmetric power balance that occurs when a jury of tutors is critiquing a student. Percy specifically describes power relations and lack of transparency in the Crit, writing:

> "staff sometimes demonstrated a difficulty in articulating their opinions and values. Rather than demonstrating a virtuosity of language they would resort to the use of imprecise and general terms, unconsciously relying on their accompanying non-verbal and gestural behaviour to convey meaning." (Percy, 2004)

Percy offered the alternative of an online Crit, where tutors are forced to be explicit as being a more successful experience.

In general, the jury critique - an event not all that different from live peer assessment - does appear to offer significant value in education, in particular, exposure to a variety of approaches, reflection and getting a better understanding of the conceptions of quality in a discipline. However, the key downside of the Crit is the potential lack of accountability regarding the judgements that are made.

2.6 The Studio Method in Computer Science

The papers from computer science referenced above (Sitthiworachart & Joy, 2004) and Hamer (Hamer et al., 2007), pertained to the very traditional domains of introductory programming in computer science, whereby students are inducted in the operations of iteration, selection and conditions in writing sequential lines of code. However, the work covered in the two courses constituting the subject matter for this study are drawn from the area of user experience or HCI, where students have to design interfaces for various sets of content. Whilst this does have some programming in it, i.e. establishing the mechanisms whereby a user navigates from one set of information to another, it is of a fairly rudimentary nature and the emphasis is much more on the structuring of information and navigation through it.

These kinds of competencies, of information design and usability are, arguably, as equally important as ability in procedural coding. However, they will be judged by much more subjective criteria than such coding. Awareness of the less clear cut nature of good and bad interface design, has led some computer scientists to use some of the studio based teaching techniques more familiar in the design based subjects. A number of papers have been written trying to import techniques from the design disciplines into computer science, generally known as “studio based approaches”. Carbone et al. (2002), Hendrix et al. (2010), Reimer et al. (2012), Estey et al. (2010), Kuhn (Kuhn, 2001), Lynch et al. (2002) and Narayanan et al. (2012) have all written experience pieces describing their attempts to embed such practices. Kuhn made a long comparison between practices in architectural study and similar practices being employed in computer science, among these she mentions long projects, multiple iterations, frequent formal and informal critique, multi-disciplinarity, study of precedents (previous student work), support from staff in setting the parameters on the design process and using diverse media to sketch out ideas. Hendrix et al., using studio techniques to get students to critique each other’s work on two occasions on a course, reported gains in motivation and self-efficacy among students. In a small meta-study of studio based approaches in CS, Carter and Carter and Hundhausen (2011) described increases in student satisfaction as measured through end of term questionnaires, as well as suggesting benefits in critical thinking, whereby students over time become more effective reviewers. On the other hand, the results of

implementing a full three year IT degree using studio based methods had very mixed results (Carbone et al., 2002). However, as the authors of the paper made clear, this might arise from existing issues within the curriculum as much as from the teaching methods attempted. To implement full semesters of studio based learning is costly from the point of view of tutor effort and the organisational logistics involved. Hence, the human resource implications of continual informal critique might not be easily supported.

2.7 Exemplar Based Assessment

While many of the claimed effects for peer assessment (particularly in the Van Zundert et al.'s meta study) seem to arise through the provision of more extensive and faster feedback, it is evident from Hamer et al.'s study, that many of the claimed positive effects of peer assessment are not so much in the feedback that is *received*, but rather the act of *giving* the feedback. Bloxham and West (2004) explicitly extolled the practice of peer assessment as the mechanism through which students begin to understand the assessment process better. If this is the case, and it is the act of giving feedback, or merely of giving scores, and the exposure to the range of possible attainment is the truly transformative element, then potentially the "peer" element of peer assessment (the act of placing a numeric value on a peer's work) might be an inhibitor of learning gains, rather than the vehicle for them. By this I mean, the act of giving a score to a fellow student, with all the concomitant feelings that might arise from it (e.g. solidarity, affection, betrayal or resentment) might be less effective (in terms of learning) than marking, say, a previous student, with whom the marker has no such connection.

Recently, there has emerged a strand of education research that precisely engages the students in the act of assessment or evaluation, not of their peers, but rather of the work of previous cohorts. There is no official term for this, but the words most often associated with the practice are "exemplars", namely, examples of prior student work, and "feed forward", that is to say, effectively getting feedback in advance of undertaking an assignment, in terms of the *typical* versions of high attainment and potentially also, lesser attainment. The theoretical impetus for this came in Royce Sadler's 1987 paper, "The Promulgation of Assessment Standards" (Sadler, 1987). In it, he argued that detailed rubrics might not be the best way of communicating such standards, because students might not understand them. He believed the best way to communicate those standards to students was to let them see them embodied in exemplar pieces of work. The first real research attempting to document and assess this practice was Rust & O'Donovan's (2003) paper. In this, they organised a voluntary extra session where students could come and collaboratively grade previous students work. They found that those who had come to this session achieved higher scores.

One of the most interesting papers in terms of quantifying the effects of exemplar based elucidation of marking criteria (from now on called "Feed-forward exercises") was Wimshurst and Manning's (2013) study, where a course in 3rd Year Advanced Elective in Criminology was re-engineered to build in an assignment where students marked previous students work.

In the 2009 iteration of the course, the assessments comprised a 30% case study (2,000 word essay) and a 70% exam. In 2010, it was re-engineered such that the exam was worth 60% and an extra assignment worth 10% was added, where students had to choose three from six case studies from the previous year and write 150 words of feedback on each. These acts of marking were then marked by the tutors.

The authors wrote:

> "In the past, some assignments have been excellent, but many to be mundane with lots of description of the chosen agency and its programmes, while rarely consolidating the analytical and insightful opportunities presented by the assignment. That is, they tended to remain multi-structural, touching on aspects of the assignment superficially…The aim of the feed-forward intervention was to improve students' understanding of the kind of coherence and integration which should characterise complex pieces of assessment, and hence to improve the quality of their own work." (Wimshurst & Manning, 2013)

The authors cited Royce Sadler:

> "once students recognize quality, and the different manifestations of quality for the same assessment item, they move toward becoming connoisseurs of good work and increasingly adopt holistic approaches to monitoring their own efforts" (Sadler, 2009)

At the end, the results of the two cohorts (2009 and 2010) were compared. The prior Grade Point Average for both cohorts was more or less the same (4.38 vs 4.46). In the actual case study essay itself the results were promising – in an assignment marked out of 40, the average score was 25.7 in 2009, but rose to 28.2 in 2010.

Table 2-5: Wimshurst and Manning: Improvement of Cohort Controlled For by GPA

	2009	2010
GPA	4.38	4.46
Full Cohort		
Mean	25.7	28.2
SD	5.12	4.55

Highest 50%		
Mean	69.3	74.9
SD	12.39	11.37
Lowest 50%		
Mean	60.4	66.2
SD	11.74	11.37

Wimshurst and Manning justified analysing the data for the highest and lowest 50% on the basis that it demonstrated that feedforward exercises benefited the cohort as a whole and had the same effects for the better students as for the weaker ones. Moreover, while there was a relationship between the students' grade point average and how they did in the case study, there was none between how they graded and how they did in their own case study. The authors suggested that this was potentially because improvement went across the whole cohort, although the case does not appear that strong. Another interpretation they make, which is potentially more convincing, is that while students can accurately assess the quality of other students' work, this does not guarantee they can carry this awareness forward into similar assessment tasks.

A further paper that does offer some quantification of learning gains through feed-forward exercises is that of Hendry, Armstrong, and Bromberger (2012), where a tutor carried out some exemplar based sessions in three modes: 1) with the tutor (a) prioritising discussion; 2) with the tutor (b) mainly focusing on what was wrong with the exemplars; and 3) with the tutor (c) holding no discussion at all. Basically, the students were more satisfied with tutor (a)'s session and also achieved higher grades. From this they concluded that it is the quality of dialog that seems to be a key factor.

The studies involving exemplars have nothing like the volume and comparability visible in the literature on self and peer evaluation. However, what they do provide is a much clearer sense of the route between the exercising of critical judgement in evaluation and the subsequent production of better quality work.

2.8 Relevance to the Proposed Study

The current study involves the following.

1. A longitudinal analysis of an MSc course in computer science involving four years of genuine peer assessment (where students evaluate each other) combined with training sessions, where students evaluate previous students' work (more akin to feed-forward activities).

2. A longitudinal analysis of a first year BSc course, where the evaluation is always a feed-forward exercise.

The two studies also differ in terms of the marking instruments used. The masters' course is much more, to use Falchikov and Goldfinch's (2000) terminology, a G+ marking instrument, which is to say, more holistic criteria, whereas the BSc uses a D marking instrument (highly fragmented criteria). The kinds of re-engineering of a course that Wimshurst and Manning (2013) undertook, is similar to what I did, although the effect was only studied for one year in their case. The fact that the MSc marking events do have summative implications does mean potentially there may be the kinds of non-academic marking seen in Magin (2001). These will be researched and measured. The measures of agreement recommended by Falchikov (effect size and correlation coefficient) were used over the course of this study and I also propose new ones that not only attempt to judge the comparability of marking, but also the quality and seriousness of it. I also aim to describe in a more detailed way the actual benefits of peer assessment to the students who participated in it.

All this being said, it needs to be borne in mind that while the study was informed by the tradition of research in peer assessment, it also involved using a new method with very few antecedents in the literature, that is to say, live peer assessment using EVS technologies. In the summative mode (on the MSc course), this involved the students presenting and then receiving an instantaneous result from the vote of the class. In feed-forward mode (on the BSc course), it involved the evaluation of previous students' work in a lecture theatre, with between 180-220 participants, with the aggregated results being seen instantly by the whole class after each judgement. While this survey covers peer-assessment and feed-forward, as practiced in different disciplines across the globe, only a very few researchers have considered peer-assessment practiced in a completely synchronous way, as pursued here. As such, the practice being studied combines the instantaneity of the design Crit with the scalability afforded by technology and the democratic potential of peer review and evaluation.

2.9 Conclusion

It has been explained how the practitioners of peer-assessment typically started out with the assumption that it was the greater volume and speed of feedback that was the chief value of the practice. However, as has emerged from some of the examples described in this chapter, it seems the actual act of giving feedback is potentially more transformative. This is because it requires the students to reference what they are evaluating with some sense of the standards of the field and to try to situate that work within it. Attempting a more synthetic summation of the properties of peer learning, Kenneth

Bruffee in 1984 wrote a paper called “Collaborative Learning and the Conversation of Mankind” (Bruffee, 1984). In it, he made the point that all thought is internalised conversation, and what he called “normal discourse”, the typical conversations shared by practitioners of a field, could most easily be acquired by collaborative learning. He wrote:

“My readers and I (I presume) are guided in our work by the same set of conventions about what counts as a relevant contribution, what counts as a question, what counts as having a good argument for that answer or a good criticism of it… In a student who can integrate fact and context together in this way, Perry says, "we recognize a colleague." This is so because to be conversant with the normal discourse in a field of study or endeavour is exactly what we mean by being knowledgeable- that is, knowledge-able in that field.” (Bruffee 1984)

In this sense, it might be that peer evaluation, and exemplar based evaluative activities are much faster ways for students to become inducted into the normal discourse of a field.

Next, how PA was applied over four years on a masters’ course called Multimedia Specification Design and Production is examined.

Chapter 3. Multimedia Specification Design and Production

In this chapter, I cover how live peer assessment (LPA) was applied to the MSc module Multimedia Specification Design and Production. This is a 30 credit MSc course in computer science, which the researcher led from 2002 to 2013. In the thesis as a whole, the effects of live student marking on student achievement and student participation are investigated. In this chapter, we do so in the context of fully live peer-assessment, where student marks have small but real effects on the scores they achieve for their assignments. This course had enrollment varying between 60 and 30 students over the four iterations and the coursework that was peer-assessed involved group work (in the 4 iterations of the course under examination there were 22,19,9 and 11 groups respectively). Consequently, because of the use of group work and the small sample sizes, it is difficult to draw any conclusions regarding the level of the quality of the artefacts delivered, particularly whether the students improved or not. However, what is investigable are the patterns of marking evidenced among the students over the four iterations, whether they coincided with those of the tutors (or not), what factors led to greater or lesser coincidence as well as how the students experienced the process and how it contributed to their own subjective understanding of their own practice. These are the general goals of this and the following chapter.

3.1 Module Aims and Learning Outcomes

The following are the aims and objectives of the Multimedia Specification Design and Production MSc module.

Module Aims:

1. Apply principles of computer science to the specification and design of multimedia systems;

2. Develop an understanding of research into multimedia computing and how this may relate to current application and future development issues;

3. Appreciate the constraints imposed by interdisciplinary project working. such as are required in the development of multimedia computer systems;

4. Develop a range of advanced skills and techniques for the development of multimedia computer systems.

Module Learning Outcomes

1. Know how a range of techniques from several disciplines, including traditional software engineering, is applied to multimedia system design and be knowledgeable about advanced features of the analysis and design of a multimedia system;
2. Know how multimedia is rendered, including the data representation of audio, video, animation, text and graphics along with the standards used to this end;
3. [The student] Can apply principled methodologies to the specification and design of multimedia artefacts;
4. [The student] Can develop prototypes using a standard multimedia authoring tool for the creation of such artefacts;
5. [The student] Can explain the issues involved in their work and justify and explain the approach taken.

3.2 Course Schedule

The fundamental structure of the course was established during the 2003-2004 academic year when the four fundamental assignments were established. The course was worth 30 credits and was delivered over 13 weeks.

Table 3-1:Multimedia Specification Design and Production Course Schedule

	Week	%	Type	Medium
Assignment 1	6	20	Individual	Online Test
MCQ	A 50 question multiple choice test covering technical issues related to media storage, together with other questions dedicated to multimedia development methodologies (waterfall, information architecture, heuristic evaluation)			
Assignment 2	9	30	Group	**Flash artefact** (.swf file) with MS Word documentation
Prototype Artefact	A prototype multimedia artefact (group assignment) together with documentation			
Assignment 3	11	30	Individual	Slideshow
Slideshow Evaluation	An audio visual slide show containing a review of another group's prototype artefact (individual assignment).			
Assignment 4	13	20	Group	**Flash artefact** (.swf file) with MS Word documentation

Completed Artefact	A fully completed multimedia artefact (group assignment) together with documentation

The original proportions of marks are reported above. However, the proportion awarded for each assignment has varied over the years, initially being 30/20/20/30, but by the final iteration it this had become 20/30/30/20. This was because it was felt that the first assignment was more a precursor to the creative work on the course and therefore, its weight should be reduced.

As can be seen from assignment 3 (Slideshow Evaluation) in table 2-1 above, peer review was already used in an assignment where students reviewed other groups' work (prior to the introduction of live peer assessment with clickers). The justification for the introduction of student peer reviews was that, in so doing, they could apply the fundamental concepts of heuristic evaluation and information architecture to a concrete piece of work. They would have to map the theory of heuristic evaluation to the various features, positive or negative, of a multimedia artefact. Consequently, this is definitely peer-review, however, it is difficult to call it peer assessment, since the student does not award any grade to the group whose artefact they are viewing. Using the fundamental categories of Topping's (1998) peer assessment classification system, it can be said that:

Table 3-2:Topping's Peer Assessment Classification Applied to Assignment 3

Focus (Quantitative/Qualitative)	Qualitative
Summative/Formative	Formative (to group being assessed); Summative to assessor
Product (what is being assessed)	Multimedia artefact
Relation to Staff assessment (substitutional or supplementary)	Supplementary (students already receive staff feedback for their group projects, the student reviews supplement this)
Official Weight (contributes to grade or not)	The review itself is marked by staff, but there is no awarding of marks to students by students
Directionality (One-way reciprocal or mutual)	One way – from student (individual) to student (group)
Privacy	Open – assessor's name is know to assessees
Contact (distance or face to face)	Distance
Constellation Assessors (individual or groups)	Typically 2-3 students, each independently writing one review of the group artefact
Constellation Assessed(group)	Group
Place and Time	Within 3 weeks of assignment

The use of audio-visual slideshows as the vehicle for communicating peer reviews was chosen as it encourages more direct instantiation of the principles of heuristic evaluation. That is, the students are asked to create screen captures and comment directly on instances of usability violations.

3.3 Introducing Live Peer Assessment

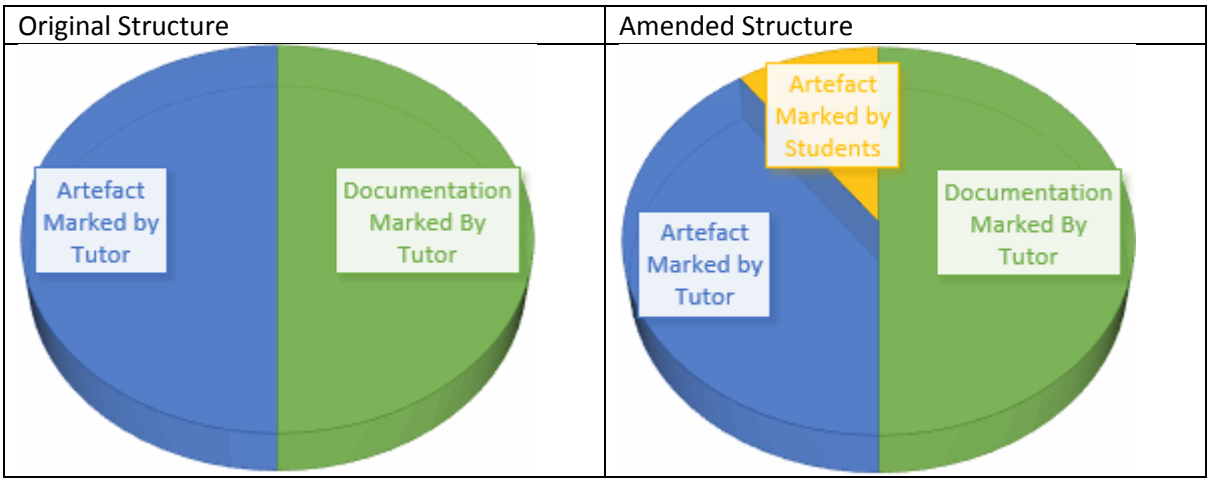
The motivation for introducing live peer assessment arose from my participation in the 2009-2010 Cable (Change Academy for Blended Learning Enhancement) Project. This was a project to bring about change in the delivery of courses using a Blended Learning Technique and its mode of operation involved encouraging academics to attempt small scale projects within university courses. My particular project was supported by the purchase of EVS (electronic voting systems) clickers to enable in-class polling and quizzing of students. My goal was to use EVS for peer assessment, specifically, getting the whole class to vote on student groupwork. As mentioned in the introduction chapter, one of the most frequently recurring points of dissatisfaction, according to the National Student Survey, was the time taken to give feedback. In 2014, The Times Higher Educational Supplement, commenting on the National Student Survey for that year, wrote:

"Assessment and feedback was again rated the lowest by students, with just 72 per cent saying they were satisfied with this, the same level as last year." (Grove, 2014)

It was believed that using live peer assessment would lead to much shorter feedback time, in fact, it could be said to be almost instantaneous. When the students present their work to the class, a score can be given within minutes of presentation.

Two assignments that were changed for these four iterations were the group multimedia artefacts, both in their prototype and in their final versions. The documentation continued to be marked solely by the tutor, but 20 percent of the marks for the artefact now came from the students. Consequently, overall for assignments 2 and 4, the original and the new structure looked as follows in Figure 3-1:

Figure 3-1:Marking Percentages of Multimedia Artefact Assignments



This meant, that whilst the student contribution to the marks for the two Flash artefacts would be substantial (20% each), that artefact itself now only constituted 50% of the assignment and therefore, the student contribution to the overall grade for the assignment would be 10%. Given that these two assignments themselves constituted 50%-60% of the overall marks for the course, it meant that the students would only be responsible for 5%-6% of the overall course mark. Although this might sound substantial on first hearing, if one were to imagine a scenario where student bias was demonstrated, and in an extreme case marks awarded might be ± 20% deviant to the marks that should have been given, that in itself would only make a 1% difference to a student's overall grade for the module.

In the first two iterations of the course there were measures and incentivisation to discourage non-academic marking. In the first iteration, we gave the 20 students whose marking patterns matched the tutors most closely an extra 5% for their assignment 2 score. (Since assignment 2 at the time was worth 20% of the course marks, this meant an extra 1% on their overall score for the course). This was probably an over-generous reward. In the second iteration, 10% of the marks for the artefact (which itself was 50% of the marks for the assignment) were given according to the correlation of the individual student's marking with that of the average of the tutors which was then multiplied by 10. That is, if a student's marking correlated at 0.6 with the average of the tutors', they would score 6% out of 10% for a "quality of marking" criterion. In practice, that year, the highest correlation value was 0.68 so we therefore added 0.2 to the correlation value for the "quality of marking" element in order to reflect a typical distribution of marks. For reasons explained later on, neither approach was ideal and the second option, while notionally fairer, actually had a deleterious effect, as also covered later.

3.3.1 Logistics

The cohort sizes for the years of this study were as follows in Table 3-3

Table 3-3:Student Numbers by Cohorts and Number of Groups

Year	Number of Students	Number of Groups
2009-10	60	22
2010-11	48	19
2011-12	28	9
2012-13	36	11

In the first year when this intervention was piloted, there were 22 groups to be marked. The original marksheets for assignment 2 had four criteria used to evaluate the artefact, which were:

- Functionality;
- Extensiveness of Scope and Fidelity to Design Documents;
- One full implementation of a set-piece interaction or animation;
- Accessibility and Usability Adherence.

For each criterion, there were five attainment descriptors. However, it was judged that organising a process where 22 groups were marked collectively on four dimensions each would take too much time, and might cause error owing to the complexity of the operation. Therefore, in the collective marking event for assignment 2 it was decided to reduce the process down to just two criteria, namely, (1) The Site Generally and (2) Quality of Animation. Given that the students were not expert assessors it was also decided to use a less formal style to list the attainment criteria – see below in Table 3-4.

Table 3-4:Assessment Rubric for Prototype Artefact

The Site Generally	Animation
1. *Really poor, very many bugs, loads of things not working.* 2. *Some bugs or very ugly pages, juddering, maybe the idea behind the site is not good enough, might have too little content* 3. *Not bad, shows promise. Mostly good, might lack something either in the idea behind the site (perhaps too general, or too obvious), the implementation (some buttons juddering etc.), or lack of content (not enough screens)* 4. *Good idea behind the site, everything looks good, the kind of content is right*	1. *Animation is extremely poor – looks like a complete Flash beginner did this* 2. *Animation is poor – only very simple Flash used* 3. *Animation is average – either not enough content, or not nice enough content* 4. *Animation is very good* 5. *Animation is really beautiful, and obviously*

5. *Great idea behind the site, everything works and looks of professional standard, the sorts and amount of content that it will contain is just right*	*uses some really clever techniques*

The final assignment was similarly rewritten with the number of criteria going down from five to three and the criteria similarly expressed in more colloquial terms, as presented in Table 3-5.

Table 3-5:Assessment Rubric for the Completed Artefact

Quality of Media	*Ease of Use*	*Appropriateness of Site To Target Audience*
1. *Poor quality media. Consistently poor screens or very little media* 2. *Not good quality media – looks like it's been taken straight off the internet. Not suitable. Also, if not enough media* 3. *Ok, but potentially dull graphics, sound and video. Some screens might lack balance or might twitch* 4. *Good graphics, sound and video. Well designed screens* 5. *Commercial quality graphics, sound and video*	1. *A nightmare to navigate – very difficult to get where you want to go* 2. *Certain bugs found – buttons occasionally trigger going on to the wrong screen – not easy to use* 3. *Sometimes inconsistent use of buttons or menus – change of colouring or size or position, but everything still works* 4. *Very good and solid effort Things work very well* 5. *Fantastically easy to use. Buttons have correct rollover behaviours – shortcuts are signalled – screens don't judder and positioning of navigational elements is uniform. Accessibility features perfectly implemented*	1. *Something very wrong about this site, or just not enough in it.* 2. *Either too little content, or with some content not really right for its audience* 3. *Either site too general, too obvious, or might be unclear about who its users are* 4. *We have a clear idea of the users, and this site suits them well* 5. *We know who the users are, and this site seems perfect for them*

In the sessions themselves, I used two data projectors – one for the students demonstrating and one for the live marking with clickers. The layout of the classroom would be broadly as below in

Figure 3-2.

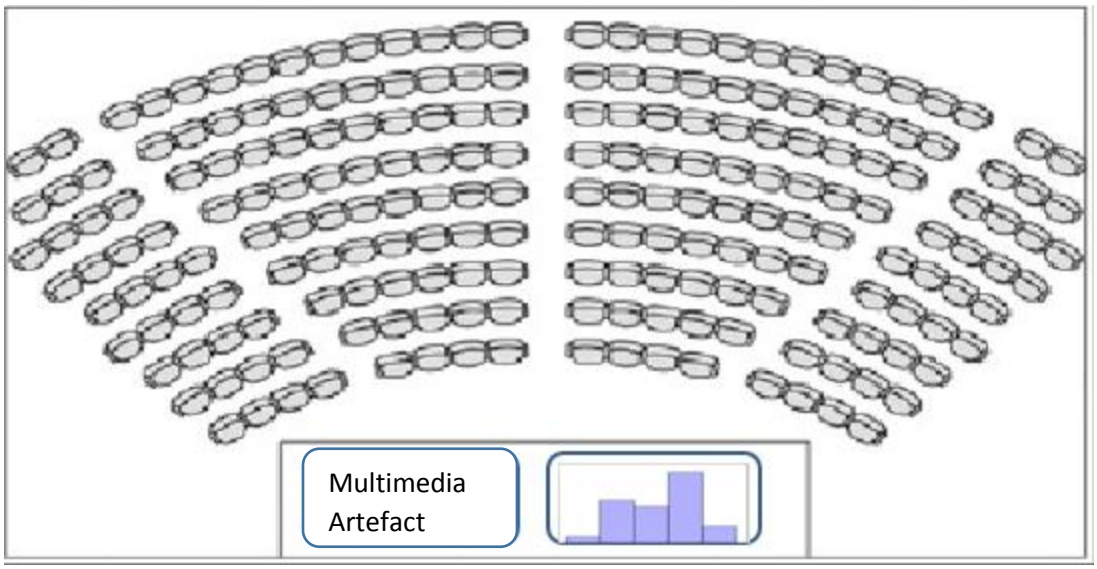


Figure 3-2:Typical Layout During a Clicker Session

In the week before the first summative event (where students would present and be marked by their peers), there would be a rehearsal event, where they would mark some examples of students' work submitted in the previous year. During that session, I would convey how the tutors would mark, the whole group would then mark an artefact according to the criteria, and after each judgement, look at the average and banding of student awarded marks. I would subsequently reply saying whether or not I agreed with that mark and why. This was in accordance with the recommendations of Van Zundert, Sluijsmans, and Van Merriënboer (2010)who believed that students would be more accepting of peer assessment if they received training in it.

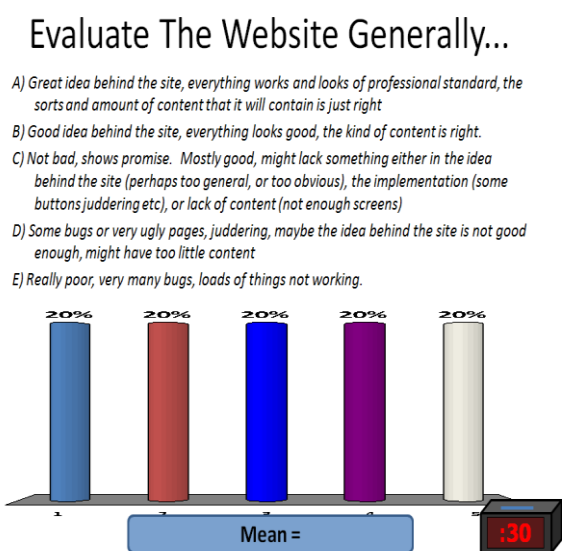
At the summative events, each group of students would present their work to the class. After each demonstration (lasting five minutes) was finished, the tutors would ask a few questions of the group (approx. three minutes), and then voting would take place (approximately two minutes – 30 seconds time for the marking grid to appear and for the students to decide on a score, followed by the "reveal", where the marks awarded would appear with an average and a histogram of the number of markers voting for each category). .

Figure 3-3 below provides an example of the screens shown to the students after each presentation, with the marking criteria. The bar charts shown appear after the marks have been awarded by the students. The tutors' marks do not appear on the screen. The marking criteria were clearly presented to the students prior to marks being awarded. A typical display would be a screen with histograms , in

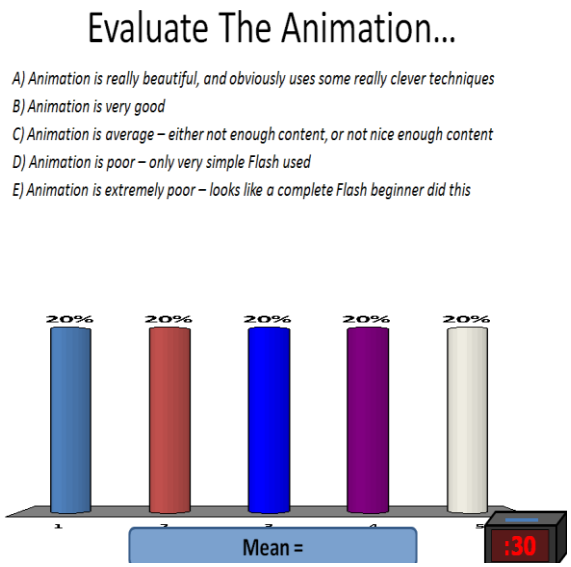
countdown mode during voting, and then with the true proportions displayed after the voting had concluded, which is shown in .

Figure 3-3.

Figure 3-3:Typical Marking Prompt Screens



Marking criteria displayed for assessing the website generally



Marking criteria displayed for assessing the animation.

After the vote had concluded, the histogram changes in proportion to the vote for each option (A,B,C,D and E) and the mean was also displayed. The box on the right which says ":30" is a countdown indicator, which shows how many seconds are left to vote.

3.3.2 Peer Assessment Marking

As aforementioned, the use of LPA ran over four iterations of the course, with it being used for assignments 2 and 4. As also mentioned above, for the first peer assessment event each year, there was a rehearsal event, where students would be asked to mark previous students' work. There are full records of summatively assessed assignments, however, the data were not always available for the rehearsal events (because of technical issues usually). The data sets are provided below.

Table 3-6:Marking Data

Dataset	2009-10	2010-11	2011-12	2012-13
Assignment 2 Rehearsal		✓	✓	✓ (but without a record of tutor marks)
Assignment 2 Final	✓	✓	✓	✓
Assignment 4 Final	✓	✓	✓	✓

Because of the highly variable recruitment on the course, and the fact that students collaborated in groups for their major assignments, it would be difficult in this study to adduce any effect on overall student outcomes based on scores in assignments (in a longitudinal fashion). Another complication is that in the first two years there were much larger cohorts and therefore, the peer assessment sessions required approximately four hours, whereas in the final two years only approximately two hours were needed This meant that in the final two years, the voting records comprised all the student markers in one sitting, whereas in the first two years, the markers would typically only be present for half of the presentations each.

3.4 The Data, Measures, and their Interpretation

The data presented below represent four years of the course. This is unusual in comparison to many of the studies in peer assessment in that very few examples of repeated practice can be found. This might indicate a weakness in some of the studies, in that the effects noted might be predicated on novelty, but also, without evidence of variation in results, it is difficult to get any sort of contextualisation of the statistics presented. A record of any course, is not just a record of the pedagogical techniques used, but also of the efficacy of recruitment, the size of a cohort, the relationships between people in that cohort and the "culture" of a programme, as established by its leadership. Also, it establishes the kinds of practices encouraged and enforced on that programme as well as those having been experienced through previous modules. Consequently, often a set of statistics is represented as being the effects of some intervention or other, when the success or otherwise thereof, might be a mediated by a host of other factors. However, seeing the data successively, while not eliminating those various mediations, allows for some kind of insight into recurring effects which might more reliably be associated with the technique under consideration.

From statistics cited previously, it has been seen that Falchikov and Goldfinch (2000) placed an emphasis on two measures: the Pearson correlation coefficient (r), and Cohen's (d) (effect size). These primarily compare the mean scores of the tutor marks aggregated with the mean scores of the student marks also aggregated. The Pearson correlation coefficient indicates to what extent the patterns of marking are the same (essentially, the extent to which the scores from both groups would deliver the same rankings). There are various interpretations attached to values (between 1 and -1), however, in keeping with the literature survey the decision here is to hold to with Sitthiworachart and Joy's (2004) interpretation, taken from De Vaus (2002)

Table 3-7: Sitthiworachart and Joy's Interpretation of Pearson Correlation Values

Coefficient (r)	*Relationship*
• 0.00 to 0.20	Negligible
• 0.20 to 0.40	Low
• 0.40 to 0.60	Moderate
• 0.60 to 0.80	Substantial
• 0.80–1.00	High to very high

The Cohen's (d) - or effect size – metric, while normally used to measure the magnitude of a variable between control and experimental groups, where typically a higher value is sought to indicate some tangible effect arising from an intervention, here, beginning with Falchikov and Goldfinch (2000), it is used to measure whether the students are over or undermarking (a positive score indicating over-marking, a negative score indicating undermarking). Effect size is calculated by comparing the means divided by a pool of the standard deviations. There are a number of variations within this technique, essentially concerning how the standard deviations are pooled. Falchikov and Goldfinch (2000) followed the recommendations of Cooper (1998) in applying the following formula (with E standing for "experimental" and C standing for "control"). Falchikov and Goldfinch understood the E values to come from students' awarded marks, and the C value to come from tutors'. The equation used by them is in Figure 3-4.

Figure 3-4:Effect Size Equation as Used by Falchikov and Goldfinch

$$\frac{(E\ group\ mean) - (C\ group\ mean)}{\frac{(E\ group\ sd + C\ group\ sd)}{2}}$$

According to Cohen (1992), effect sizes can be construed as follows in

Table 3-8.

Table 3-8:Cohen's Interpretation of Effect Size

Effect Size Value	**Meaning**
0.2	Small
0.5	Medium
0.8	Large

In Falchikov and Goldfinch's (2000) meta-study, the mean correlation was 0.69, and the mean d value 0.24. These global statistics, however, cover a wide variety across different studies – some of which cover peer assessment taking place on an almost 1-1 basis (one peer marks one student and the peer's mark is compared with the tutor's mark). Significantly, they found that the effect size became larger when the number of peers was involved in the marking was also larger.

One particularly relevant table (Table 3-9) produced by Falchikov and Goldfinch examines the effect of cohort size on the correlation and d values obtained

Table 3-9: Falchikov and Goldfinch's Table of the Effect of Cohort Size on Correlation and Effect Sizes

Number of peers involved in each assessment: mean values	1	2 - 7	8 - 19	20+	p values
Mean r	0.72 (n = 7)	0.77 (n = 12)	0.81 (n = 12)	0.59 (n = 15)	*O. 02*
Mean r omitting Burnett & Cavaye	0.72	0.59	0.77	0.59	0.02
Mean d	-0.07 (n = 6)	0.43 (n = 11)	0.24 (n = 5)	-0.31 (n = 2)	0.33
Mean d omitting Butcher	-0.07	0.05	0.24	-0.31	0. 09

Clearly, more dramatic changes in d values occur in large studies, however, it should be cautioned that this was based on a very small number of studies where d values were actually calculated. On the other hand, lower values for the r value when the number of markers was high does seem to be substantiated by a greater number of studies. That said, the p values particularly for Effect Eize are quite high, meaning that the null hypothesis cannot be ruled out.

3.5 Multimedia Specification Assignment 2

In this section of the chapter, I cover the history of the marking events and the data generated by them for the first "voted" assignment on the course (Assignment 2). Particularly, I will look at principal Falchikov and Goldfinch measures (correlation and effect size).

3.5.1 Student Marking: Correlations and Effect Sizes

When all four iterations of the course are aggregated and all the marking events regarding the prototype assignment are put together, there is a total of 146 separate evaluations made where there are both student and tutor data (73 artefacts with two criteria each), with on average 27 peers per evaluation (median 29, upper quartile 33, lower quartile 22). The r value is 0.75 and the d value 0.15. According to the literature, this is an extremely positive result – the correlation value indicates strong agreement and the effect size is small. The effect size is important because it means there was no significant over or undermarking. The global data is presented in Table 3-10.

Table 3-10: Global Data on All Marking Judgements for Assignment 2 over Four Iterations

Number of Items Evaluated (where student and tutor averages exist)	73
Criteria Per Item	2
Overall Number of Evaluations	146
Average Number of Markers per Judgement	**27.88**
Stdev (Markers Per Judgement)	**6.55**
Overall r (Tutor average vs Student Average)	0.76 (P=1.25979E-28)
Overall d (Tutor Average vs Student Average)	0.15
Tutor/Tutor r	0.68 (P=3.50698E-21)
Tutor/Tutor d	0.1

This assignment had two criteria: to evaluate the website, and to evaluate its featured animation. There is a small variation in the levels of agreement per criterion, but not much, as seen in Table 3-11 below.

.

Table 3-11:Correlation over Specific Criteria over Four Years

Criterion	R value	D value
Website	0.72 (P=5.18748E-13)	0.15
Animation	0.79 (P=8.2074E-17)	0.19

However, when looking across all the events individually, that is to say on a year by year basis, a richer and less dramatically optimistic set of data is revealed though this can also be explained by smaller samples in some iterations and events.

Table 3-12:Correlation and Effect Sizes by Iteration for Assignment 2

Year	2010	2011	2011	2012	2012	2013	2013
Event	Assessed	Rehearsal	Assessed	Rehearsal	Assessed	Rehearsal	Assessed
Comment		Only 1 tutor mark per item				No tutor marks recorded	
Number of Items Evaluated	22	6	19	6	9	8	11
Overall Number of Evaluations	44	12	38	12	18	16	22
Average Number of Markers per Judgement	33.68	28.08	30.03	16.33	22.89	19.94	28.64
Stdev (Markers Per Judgement)	4.44	2.61	5.60	0.78	0.96	1.88	0.95
Overall r (Tutor average vs Student Average)	0.87	0.45	0.71	0.85	0.83		0.76
Overall d (Tutor Average vs Student Average)	0.16	-0.35	0.01	-0.17	0.43		0.72
Tutor/Tutor r	0.70		0.53	0.75	0.84		0.49
Tutor/Tutor d	0.13		0.00	0.12	-0.14		-0.38
Tutor mean	3.29		3.24	3.17	3.36		3.75
Tutor Sdev	0.97		0.80	1.27	1.20		0.84
Student Mean	3.41	3.43	3.24	3.38	3.75	3.34	4.19
Student sd dev	1.09	0.90	0.83	1.22	1.10	1.25	1.09

In Table 3-12, there is much greater variability on an event by event basis. There is very strong correlation between tutor and student marks across all the events where tutor marks were recorded. Moreover, the effect size (*Overall d*) for the final two iterations becomes significantly larger, and for the final iteration, it is very large. The extent to which the marks awarded by the two markers (that is to

say, the correlation between the marking of the two tutors) are in synchrony overall has a respectable 0.68 Pearson correlation value (which in the case where there are only two markers is actually very high), however, on two occasions (2011 and 2013) a more modest relationship is found. There does seem to be some relationship between the standard deviation of the marks awarded by the tutors and the correlation coefficient. In a set of artefacts of mediocre standard, the standard deviation is likely to be smaller. Marking mediocre work generally is harder, which is, in fact, borne out by the marks. If we attempted to correlate average tutor and student marks but only for marks where the student average mark was in the second and third quartiles (where the average student rating was between 3.12 and 3.97), the correlation is much less than in the 1st and 4th quartiles.

Table 3-13: Correlation within Average Scores versus Correlation within Non Average Scores

When the Average Score Awarded by Students Was In Range.	Number	Correlation with Tutor Mark	P Value
3.12-3.97	73	0.54	9.65122E-08
<3.12 or >3.97	73	0.86	1.13745E-25

The variation in effect size (where in the final iteration, students appear to over mark) might be due to the different ways the marking was incentivised in 2012/2013 when compared to 2010/2011.
In the first iteration (2010), the tutors were worried about the potential for non-academic marking by students. Much of the literature suggests that students' greatest anxieties around peer assessment precisely comes from their beliefs that many of their colleagues are unfit to judge their work and also that they may award marks based on friendship rather than accomplishment. As a consequence, to incentivise academic marking, the students were told that those 20 students whose marks most closely matched that of the tutors, would get an extra 5% for their overall assignment 2 mark.

When the experience was presented at the University of Hertfordshire 2010 Learning and Teaching Conference, one comment was that this was "normative" marking – namely, giving some students marks based on their ranking in the class. Consequently, in 2011 it was decided to award students marks based on their correlation with the tutors (essentially the correlation of individual marks with those of the tutors). This initially appeared to be very effective: the correlation values in 2011 is both substantial, and the effect size (the measure of whether undermarking or overmarking is occurring) is practically zero (0.01), thus indicating that the magnitude of marks being awarded was almost identical. However,

when considering the findings for assignment 4 during this academic year, this, as will be seen, raises significant problems.

In 2012 and 2013, no mark was awarded for the level of agreement between tutors and students, and in these iterations, the correlation goes up, but the effect size also goes up, in 2013 quite highly.

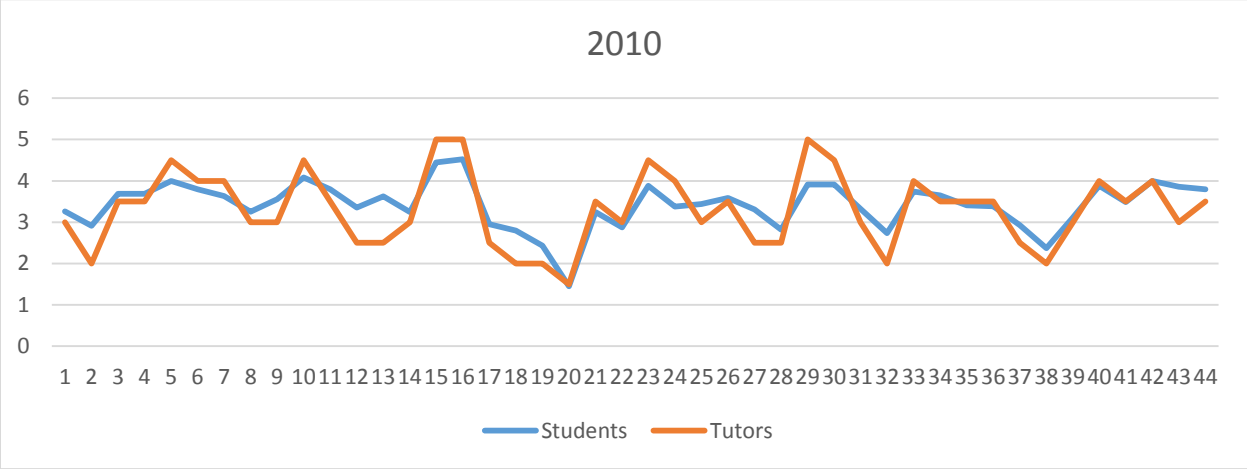


Figure 3-5:Tutor (Red) vs Student (Blue) Marks for Each Item and Criterion in 2010 – 22 groups two criteria per presentation

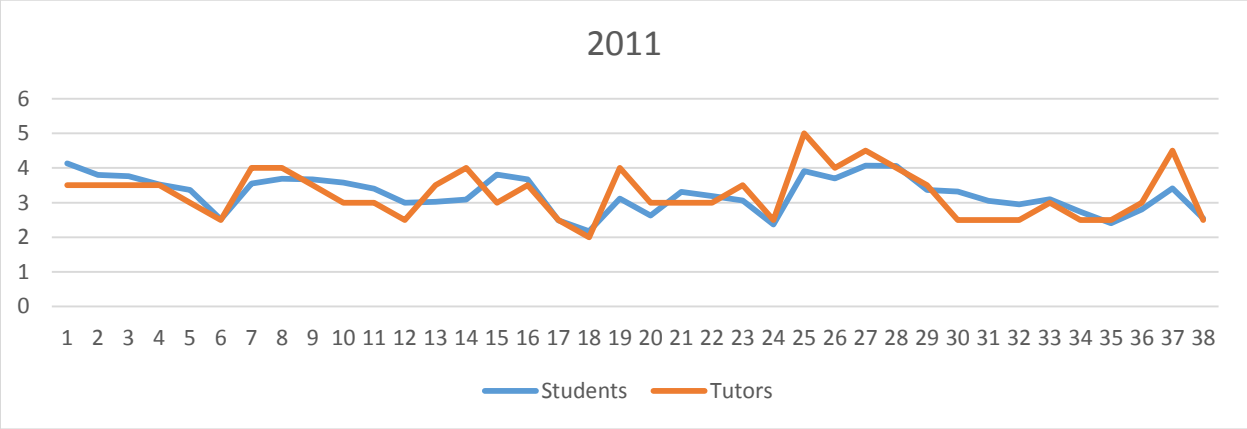


Figure 3-6:Tutor (Red) vs Student (Blue) Marks for Each Item and Criterion in 2011 – 19 groups two criteria

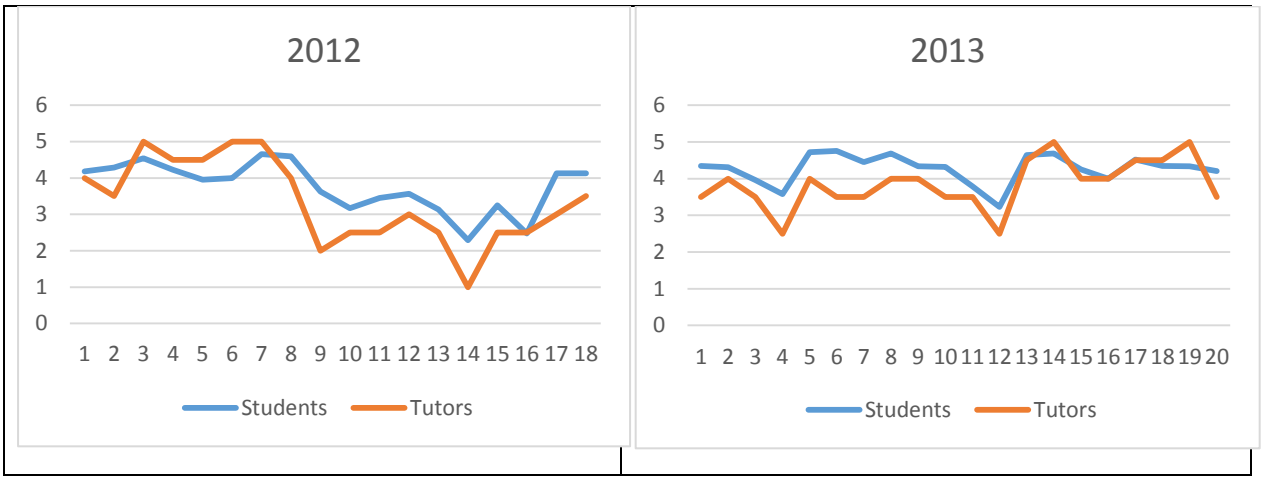


Figure 3-7: Figure 2 7:Tutor (Red) vs Student (Blue) Marks for Each Item and Criterion in 2012 and 2013

Clearly there is evidence that in the final two iterations student marking becomes more generous

3.5.2 Assignment 2: Measures of Agreement

Falchikov and Goldfinch criticised the metric of "agreement", since it is interpreted in different ways across the literature (in a continuous scale, what is the window of "agreement" – if one marker chooses 70 and the other 69, are they agreeing or disagreeing? If they can be said to be agreeing, what discrepancy needs to be there for "disagreement" to be the outcome?). This means it is difficult to compare different studies because of the difficulty of defining "agreement". However, in this case, a longitudinal study, if a definition of "agreement" is established, then how levels of "agreement" change over different iterations can be observed and what, indeed, might be causing this.

Our question is then: what is the number of times any student's mark coincides with the mark given by at least one of the tutors, as a proportion of total acts of judgement? Given there were two tutors marking each time, the average of the tutor marks will be in steps of 0.5 and therefore, to measure "agreement", any time a student's mark is within 0.5 of the average of the tutor marks, then it can be taken that the student has agreed with at least one of the tutors.

When comparing the four iterations (numerically in Table 3-14 and graphically in Figure 3-8), it can be seen that there are dramatic differences between the occasions where marking propriety was enforced (in the first two iterations when "marking quality" was graded based on correlation with the tutor scores) and those when not.

Table 3-14:Overmarking and Undermarking Proportions by Cohort

	Overmarking	Undermarking	Agreement	Total Comparisons
2010	370	307	805	1.482
2010 (%)	25.0%	20.7%	54.3%	
2011	190	210	741	1.141
2011 (%)	16.7%	18.4%	64.9%	
2012	145	56	211	412
2012 (%)	35.2%	13.6%	51.2%	
2013	265	74	291	630
2013 (%)	42.1%	11.7%	46.2%	

Viewed graphically, a gradual increase in overmarking (displayed in red) is observed over the years (green represents agreement and blue, undermarking).

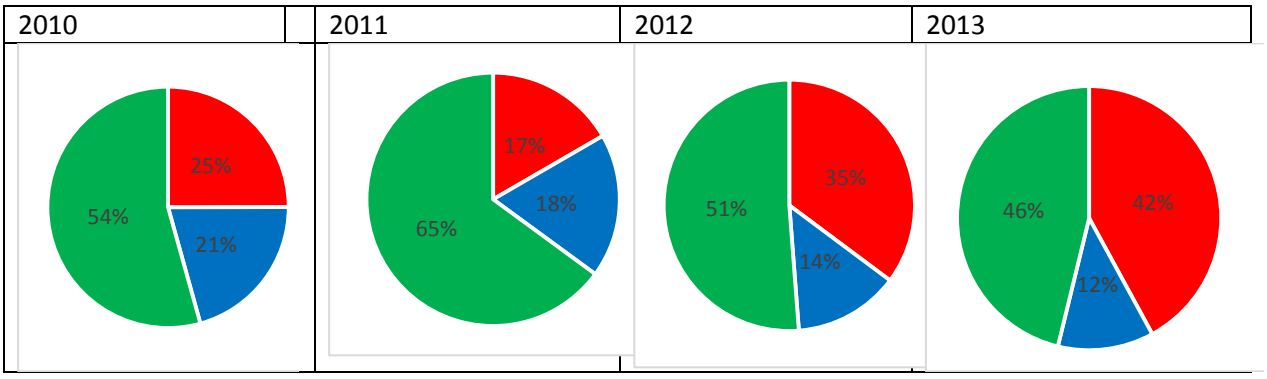


Figure 3-8:Agreement, Overmarking and Undermarking by Cohort while marking assignment 2: Agreement Green, Overmarking Red, Undermarking Blue

Nonetheless, as expressed before, notwithstanding the generosity in the marking in the final years, the correlation remains high. But is this owing to an established sense of quality among the cohort overall, or more of a "wisdom of crowds" effect, whereby the noise from unsophisticated markers is corrected by other errors in the opposite direction? In other words, to what extent are the high correlations seen between the tutor and student averages a function of the aggregation of marks (that is to say poor markers cancelling each other out), or to what extent students are *individually* putting in marks congruent with that of the tutors?

3.5.3 Assignment 2: Pairwise Correlation Measures

Up to now correlation has been considered primarily between the student average and the tutor average. When pairwise correlations (how much each ***individual*** student's pattern of marking correlates with that of the tutor average) are calculated some interesting variations over the years emerge. Below (in Figure 3-9 and Figure 3-10), the distribution of pairwise correlation values (between individual student's marking and the tutor marking) are presented as a violin plot and a box and whisker chart. The violin plot represents the likelihood of a correlation value through the magnitude of its width. There is also a notch representing the median correlation value. In the box and whisker diagram, the correlation values are presented in a box (representing the second and third quartiles, with another line within them for the median, together with lines at the top and bottom for the maxima and minima).

Here is the distribution of pairwise correlations presented as a violin plot.

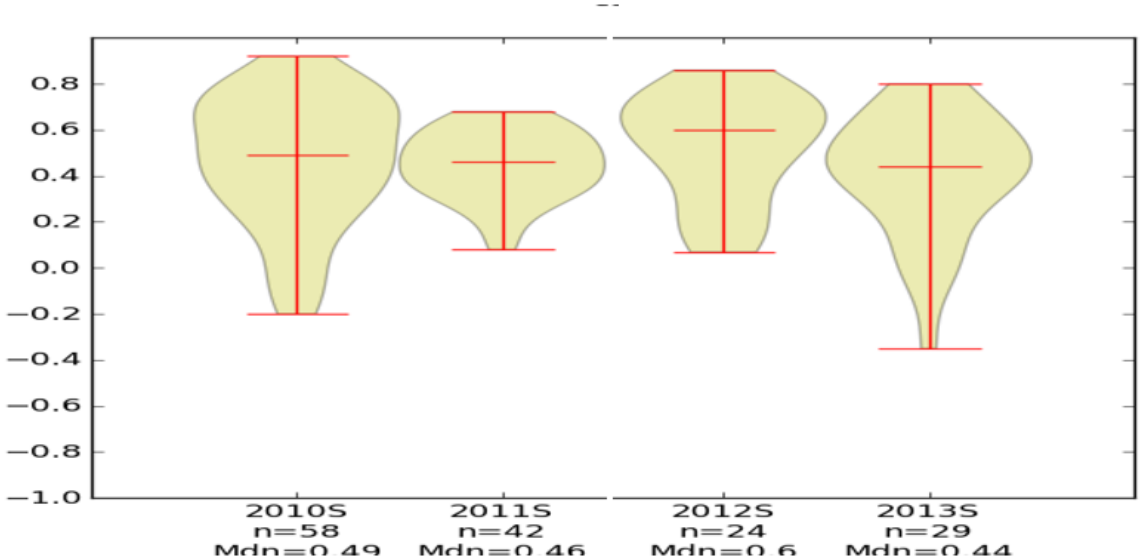


Figure 3-9:Violin Plot of the Distribution of Correlations between Individual Student's Marking Patterns and those of the Tutors for Assignment 2 Marking Events

Here are the distribution of correlations presented as a box and whisker diagram:

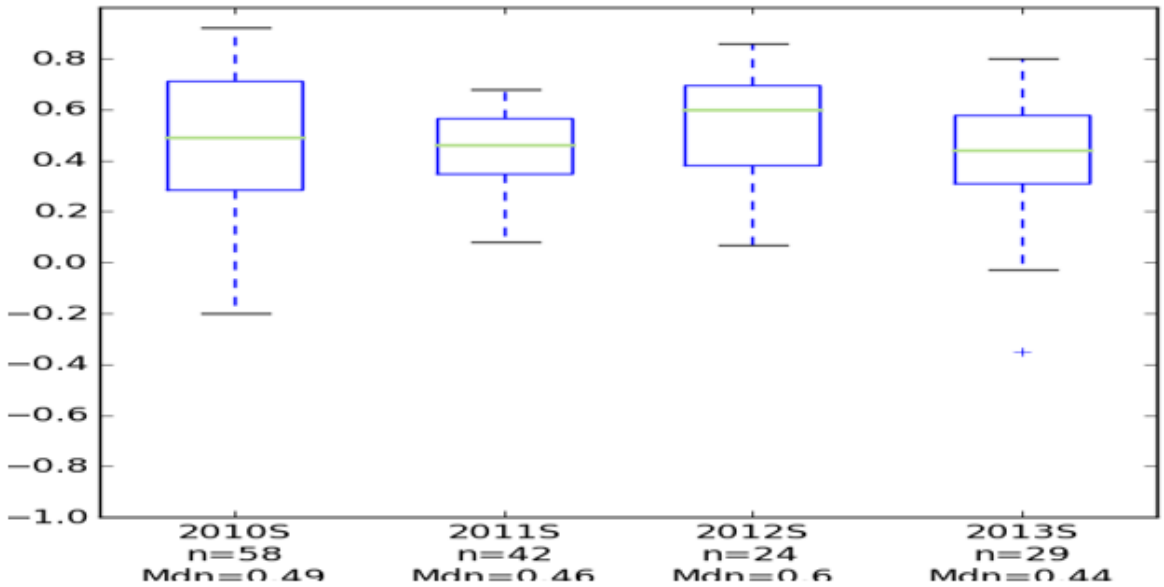


Figure 3-10:Box Plot of the Distribution of Correlations between Individual Student's Marking Patterns and those of the Tutors for Assignment 2 Marking Events

The fact that the median correlation is around the .5 mark – is so some extent encouraging – however, there are problems if we look at the level of correlation between student and tutor marks at the 20th percentile.

Table 3-15:Pairwise Student Tutor Correlation at the 20th Percentile

2010S	2011S	2012S	2013S
0.26	0.33	0.27	0.21

In practice, this means that if the marking constellation had not been the whole class per assessee, but pairwise (one student marking one other student), there was a 1/5 chance that anyone marking you would only have a very faint level of agreement with that of the tutors.

3.5.4 Inter-Rater Reliability

So far, all the comparisons have been between the tutors' marking and that of the students (be it as a cohort, or pairwise as individuals). One weakness of a lot of the literature on peer-assessment, is that it assumes the gold standard of assessment is the tutors' mark itself. However, there are other statistical

techniques, broadly known as “inter-rater reliability”, aimed at evaluating to what extent the raters agree among themselves (without reference to an exemplary set of ratings).

There are a number of techniques, all of them with particular strengths and weaknesses (comprising Chronbach’s alpha, Intra Class Correlation, Kendall’s Tau and various kappa statistics). However, most of these have great difficulty with missing data – a feature which is characteristic of much of the data collected via clickers. These missing data can result from a number of factors, among which are: (a) because students arrived late or left early and therefore, missing out on certain presentations; (b) they were involved in multiple sessions therefore not rating every piece of work (particularly the case in 2010 and 2011) or ;(c) at times they may have forgotten to click, chosen not to click, or potentially the sensors did not pick up their clicks. For this reason, missing and incomplete data is a fundamental characteristic of live peer evaluation with clickers and any inter-rater reliability metric has to be able to deal with this.

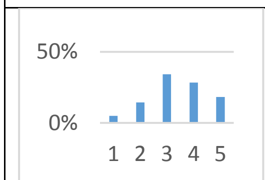
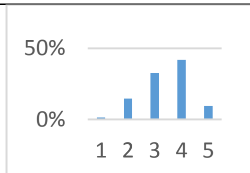
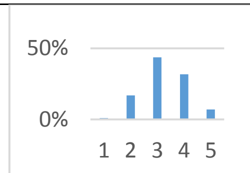
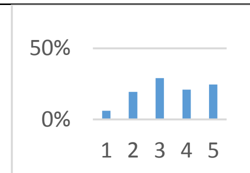
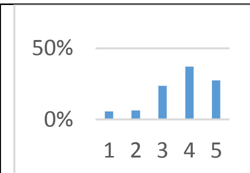
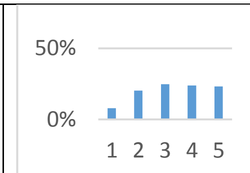
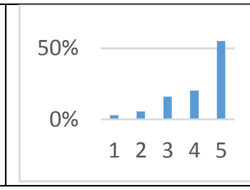
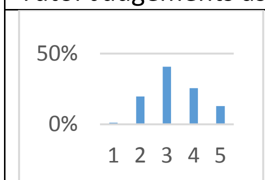
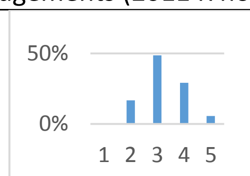
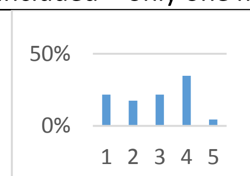
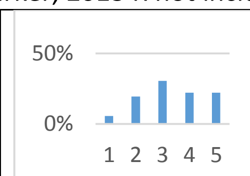
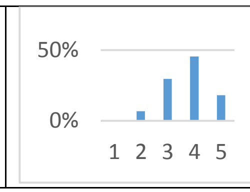
Fortunately, there is one such measurement which can do so – Krippendorf’s Alpha (henceforth to be referred to as *kalpha*) – a highly complex metric which has only recently been incorporated into a number of statistical packages (it is available in R, and there is a macro for it in SPSS). In very broad terms, it attempts to give an account of the amount of disagreement, divided by the expected level of disagreement in a population. As a consequence, it is quite a severe measure, since it might be argued in the context being used here, two scores of 4 or 5 on a particular criteria rather than being understood as disagreement on how good might instead be said to be an agreement that the work is good. Moreover, what is being evaluated in such marksheets is what is termed “higher-order thinking” – the ability to recognise appropriateness and to assign a value to it. This necessarily will generate more nuanced and arguable results than agreements over much more obvious things (for instance is the spelling correct). That caveat aside, the kalpha score for all of the peer evaluation sessions, comprising the rehearsals as well as the summative sessions is now considered. (The first rehearsal event is not included since data was not collected for that).

Table 3-16:Krippendorf's Alpha By Assignment 2 Marking Event (S=Summative R=Rehearsal)

2010 S	2011 R	2011 S	2012 R	2012 S	2013 R	2013 S
.2425	.2434	.3500	.4808	.3116	.6264	.1503

On first glance this is a surprising set of statistics. Clearly the last two ***rehearsal*** (R) sessions are the only ones where it seems there was a reasonable amount of agreement amongst markers. For all other events, the level of agreement is weak, and in the very final summative session, very weak. Perhaps this is explained by counting the actual numbers of 5s,4s,3s,2s,1s awarded by individual student markers as proportions of the total judgements made, which can be seen in Table 3-17 below.

Table 3-17:Histograms of Specific Marks Awarded By Assignment 2 Marking Event (S=Summative R=Rehearsal)

2010 S	2011 R	2011 S	2012 R	2012 S	2013 R	2013 S
Student Judgements as a Proportion of Total Judgements						
						
Tutor Judgements as a Proportion of Total Judgements (2011 R not included – only one marker, 2013 R not included – not recorded)						
						

Here the paradox becomes clearer and explains the very low Kalpha score for the 2013 summative evaluation of assignment 2, namely, large levels of overmarking, with large numbers of 5/5s being awarded. What is equally vivid is the greater levels of agreement during the rehearsal sessions rather than the summative ones. This, I suggest most, likely arose due to the level of guidance given by the tutor and also, the fact that the items chosen for evaluation by the students were "exemplary" ones, that is to say, clearly good, clearly bad and clearly average. It might be the case that the increasing experience of running these rehearsal events made the tutors choose better exemplars and structure the sessions in such a way that more agreement was achieved. Furthermore, since the rehearsal events did not require mandatory attendance, it might be argued that they were being attended by the more conscientious students.

At this point, it could be concluded that the event for which the peer assessment appeared to be at its most positive for assignment 2, characterised by a high correlation between tutor and student averages, a very small effect size along with a reasonable kalpha, was in the summative assessment during 2011. On this occasion, students were given scores for their marking based on their correlation with the tutors' marks (10 *(correlation + 0.2)) out of 10. However, all incentivisation of student marking for the 2012 and 2013 events were dispensed with, which appeared to result in more generous marking by the students. In 2012 a record was kept of which student marked which group though no use was made of this. In 2013 the EVS clickers were distributed without a record of who marked who. In 2012, this increased the effect size, but not in a way that would create concerns. However, by 2013, even though the correlation between tutor and student marks is high (0.76) it is also characterised by a high effect size and a low kalpha (indicating, therefore, that not the whole group was consistently overmarking, but that a subset of it were doing so and thus, disagreeing with the others in the group).

Of course, any cohort of students will have their own particularities and probably culture, but one unavoidable conclusion that can be drawn from this is that possibilities for inflated scores will exist. This however can be mitigated by (a) explicit mechanisms for rewarding student marking congruent with the tutors but also (b) merely keeping a list of who is marking who – making sure it is non-anonymous. This being so, why was the decision taken to remove the restrictions on student marking in the first place? This will become clear when the assignment 4 statistics are presented.

3.6 Multimedia Specification Assignment 4

Before beginning to examine these figures, it is worth highlighting some contextual features of the marking events in the case of assignment 4. Assignment 4 requires students in, groups, to produce a completed multimedia artefact in a subject of their choice. Whereas the prototype allows placeholders, *lorum ipsum* text and all manner of incompleteness, the final artefact requires that the work is publishable. For this assignment, each artefact is marked on three criteria rather than two, which makes for longer marking sessions, and potentially more boredom to set in. Secondly, it occurs right at the end of the course and therefore, it sits alongside exam preparation and deadlines on other courses. Therefore, the students might also not be as focussed in these events. Finally, because of the pressure on students at that time of the course, rehearsal sessions are not organised. Consequently, whilst they are aware of the criteria, they do not have the experience of marking according to them in a non-summative context before the event (unlike the rehearsal event for assignment 2, where students

practise marking on some previous cohort's examples). The global statistics for assignment 4 are presented below in Table 3-18.

Table 3-18: Correlation and Effect Sizes by Iteration for Assignment 4

Year	2010	2011	2012	2013
Event	Assessed	Assessed	Assessed	Assessed
Number of Items Evaluated	22	19	9	11
Overall Number of Evaluations	66	57	27	33
Average Number of Markers per Judgement	**35.97**	**30.93**	**23.11**	**24.48**
Stdev (Markers Per Judgement)	**2.87**	**8.35**	**1.34**	**1.73**
Overall r (Tutor average vs Student Average)	**0.69**	**0.90**	**0.84**	**0.43**
Overall d (Tutor Average vs Student Average)	**0.34**	**0.02**	**0.53**	**0.71**
Tutor/Tutor r	0.61	0.74	0.80	0.53
Tutor/Tutor d	-0.13	0.39	0.29	0.16
Tutor mean	3.33	3.38	3.48	3.79
Tutor Sdev	0.73	0.86	1.10	0.66
Student Mean	3.52	3.39	3.90	4.02
Student sd dev	0.42	0.70	0.46	0.41
Kalpha	0.2441	0.5480	0.1825	0.1238

What can be seen here, are two anomalous results sets. That is, 2011, characterised by extremely high correlation and kalpha, whereas 2013 is characterised by low correlation (0.43) and kalpha (0.12) along with a very high effect size (0.72). In 2011, the students are extraordinarily close to the tutors marking and moreover, demonstrate high inter-rater reliability. That is to say, as a cohort they are marking almost exactly like the tutors. However, as part of a routine questionnaire put out at the end of the course, one student wrote:

> "During the second marking sessions, i noticed people were just putting what steve and trevor marked and they would give the same marks. Thats why you might notice high correlation between students n tutors marking for second session. It was after 1st session we realised that our marking should match tutors which i think is not fair."

And in a recommendation for future sessions:

> "Yes, please hide your clicker when marking :)"

The confirmation that a number of students were, indeed, spying on the tutors' marking and merely seeking to put in the exact same value can be seen in the graphs of equivalent (green) and over(red) and under (blue) marking below.

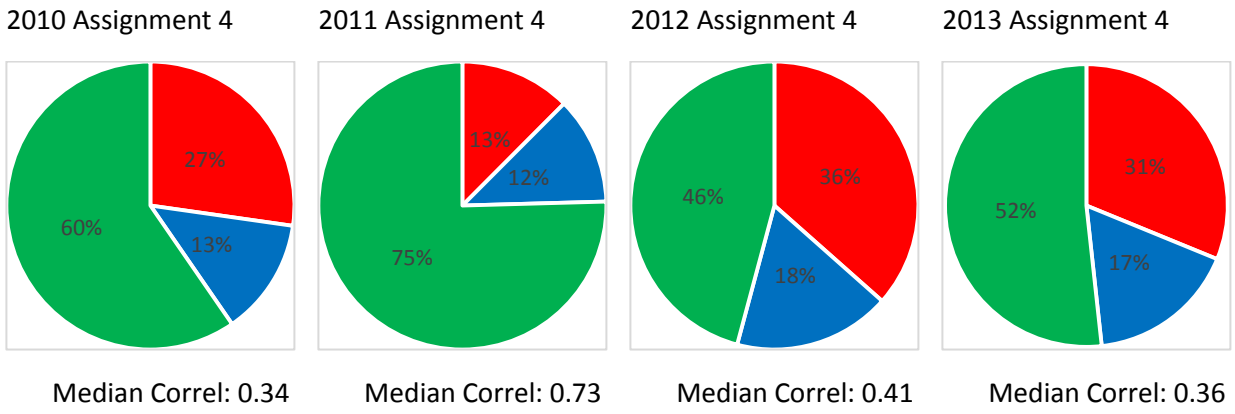


Figure 3-11:Students' Agreement and Over and Undermarking over 4 Years

Clearly, the level of agreement between tutors and student is massively high in 2011. Moreover, when pairwise correlation between each student and the average of the tutors is performed and these are sorted, then the median student would score 0.34 in 2010, 0.73 in 2011, 0.41 in 2012 and 0.36 in 2013.

A good way to see the anomalousness of the 2011 assignment 4 result, and to ensure that the same problems might not also have been evident during other marking events, is to produce a violin plot of individual student correlations against the tutor marks for each of the marking events.

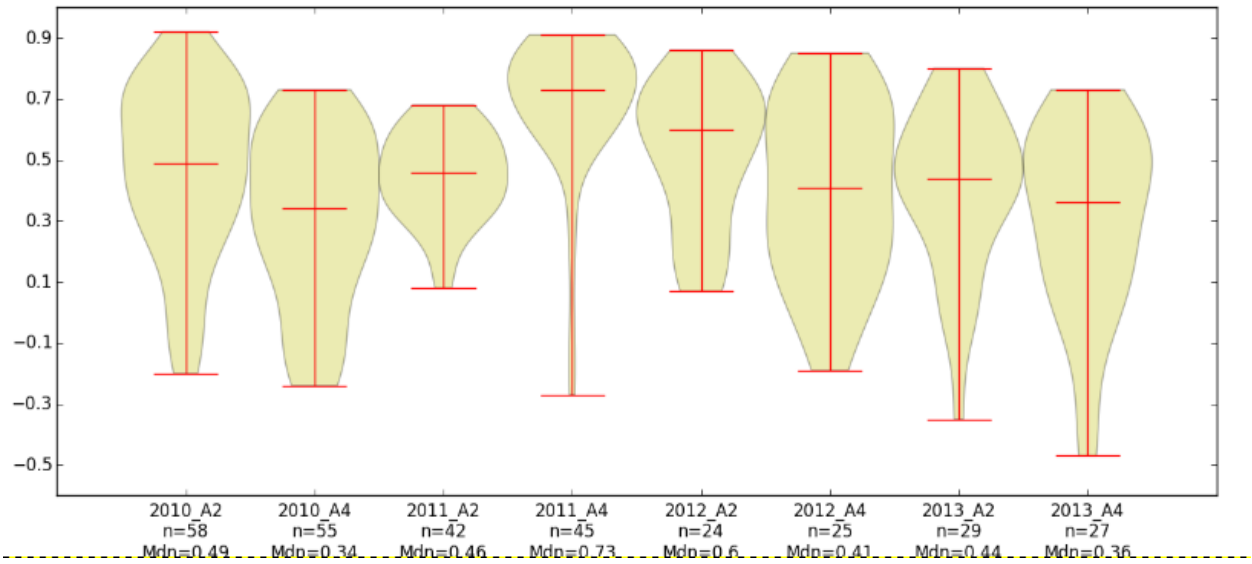


Figure 3-12:Violin Plot of the Distribution of Correlations between Individual Students Marking Patterns and those of the Tutors for Assignment 2 and Assignment 4 Marking Events

Here, from the very high median correlation the anomalousness of assignment 4 during 2011 can be seen. However, it is also worth noting, that in all years, the mean correlation goes down for assignment 4 as compared to assignment 2, as mentioned before, which could have a number of causes, including deadline pressure on other courses and/or lack of a training event with these criteria. A box plot (Figure 3-13) of these data is equally revealing.

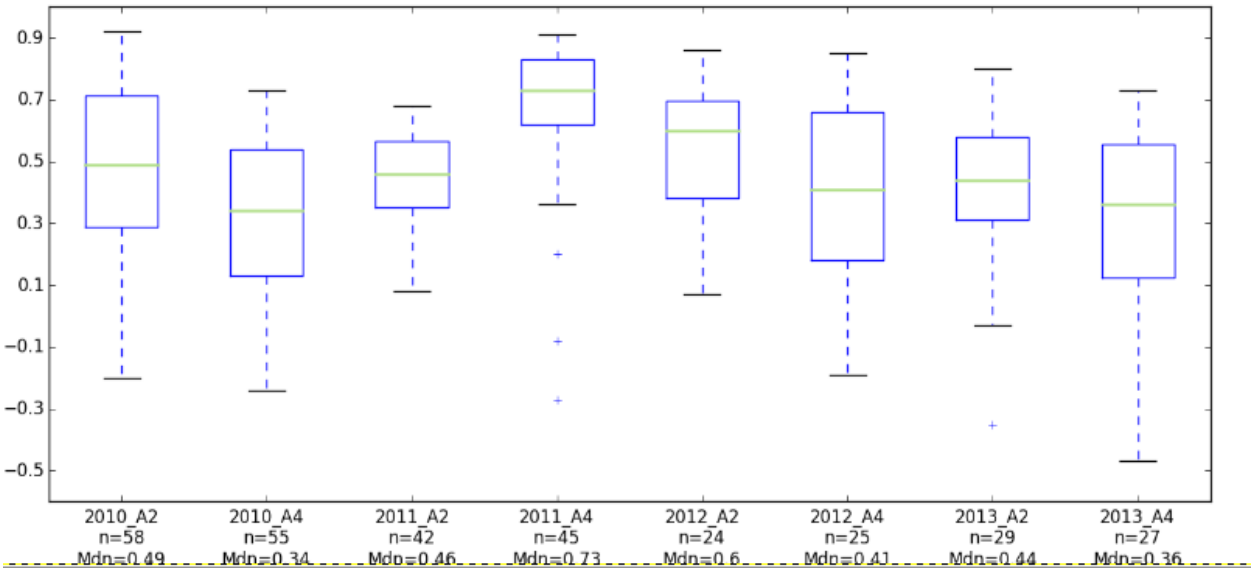


Figure 3-13:Box Plot of the Distribution of Correlations between Individual Students Marking Patterns and those of the Tutors for Assignment 2 and Assignment 4 Marking Events

Probably the most striking way to see the literal "abnormality" of this event, is to look at how *normal* were the distribution of correlations between the marking of individual students and the tutors over the different marking events. A good measure of this is the *Kurtosis* and *Skewness* of the lists of individual student correlations with tutor marking per assignment and cohort. *Kurtosis* essentially means the pointiness of the distribution, though more accurately, the weight of the tails compared to the rest of the distribution. The skewness measures the symmetry of the distribution. As can be seen below, for all the distributions besides Assignment 4 2011, the kurtosis and skewness are between 1 and -1. For that year and assignment alone there is massive asymmetry and tail weight.

Table 3-19: Kurtosis and Skewness of the Distribution of Correlations between individual student marks and tutor marks per assignment and cohort.

Year	2010		2011		2012		2013	
Assignment	A2	A4	A2	A4	A2	A4	A2	A4
Kurtosis	-0.299	-0.87	-0.206	7.135	-0.758	-0.973	0.983	0.304
Skewness	-0.667	-0.419	-0.417	-2.463	-0.639	-0.224	-0.948	-0.855

While on one level, this could be dismissed as a mere security issue, which could be resolved by concealing one's marking more effectively, equally it could be said that by determining part of the grading by *how near you matched the tutor's grading*, then it diminishes the figure of the student by effectively valuing them on the basis of to what extent they are a clone of the tutor's judgement. However, any tutor will have acquired that judgement after many years of marking, being able to recognise issues in an artefact, and having a vast inner set of examples with which to compare any new one that needs to be judged. Accordingly, it really might be that a very high correlation between tutor and student marks might not be a desirable outcome on the level of assessment for learning, because it is only through making judgements, and maybe marking them wrongly and gradually correcting, that real judgement will arise. However, as has been seen, putting no restrictions whatsoever around students' marking can also lead to unsatisfactory consequences, particularly in terms of score inflation and friendship marking.

Nonetheless, because the two cohorts where the generosity of markers began to increase were smaller, their influence on the overall correlation and effect size for all students marking assignment 2 artefacts over the whole 4 years was not strong. Taken all together, live peer assessment resulted in a very high r of 0.79 (Falchikov and Goldfinch's average was 0.69) and a respectable 0.15 in terms of effect size (Falchikov and Goldfinch cited 0.24 as the average). For assignment 4 (when excluding the data from 2011) an r of 0.73 is obtained, however, the effect size is 0.39. In all of these events, on only one

occasion did the correlation score go below the Falchikov and Goldfinch's average (2013 assignment 4). Consequently, it must be inferred that live peer assessment has some affordance, which leads it to producing higher correlations between tutor and student marks than is typically obtainable in more traditional methods of peer assessment.

3.7 Conclusion

In this chapter, the aim has been to display the kinds of metrics that can quantify the robustness of peer assessment. It has been demonstrated that the two metrics promoted by Falchikov and Goldfinch (2000) (correlation and effect size) have stood the test of time and are core to evaluating how any peer assessment exercise is successful. However, five other metrics have proven to be extremely illuminating, namely:

1. pairwise correlations between individual students and the tutor average (finding the median correlation among the students and also violin plots of those values overall);
2. counts of the number of times particular attainment scores were awarded (particularly maximums);
3. Percentages of overmarking, undermarking and agreement between students and tutors;
4. Krippendorf's Alpha as a measure of the inter-rater reliability of the student marks;
5. The Kurtosis and Skewness of the distribution of correlations between tutor and individual students' marks

The impact of peer assessment extends to questions not only of cognition and enhancement of learning, but also relates to questions of justice. Since the marks awarded by the student assessors will have impacts on student grades, clearly measures have to be taken to make sure those awarded are just. The two recommendations for practice arising from these results are as follows.

1. Keep the proportion of marks awarded by students to being a small amount of a particular piece of work.
2. Do not incentivise students by the level of similarity to the tutors' marks, but rather, establish fairly broad criteria for acceptable marking (potentially that a student's marks should correlate with the tutors marks at 0.2 and above, and should demonstrate a cohort effect of less than

0.8) and offer some notional percentage credit for marking within this range. This would mean that students would have considerable freedom to mark, however, it would ensure that their marking has at least some relationship with that of the tutors.

The major variable in the history of these assessments has been the strategies used to incentivise "honest" marking among students. In 2010 there was a premium for the 20 markers whose marking was most similar to the tutors - a procedure that appeared to work well, but was potentially suspect due to its "normative" character. In 2011, we awarded marks based on a according to the formula: $((r + .2) * 10)$ out of 10 – that is to say multiplication of individual students' correlation with the tutor multiplied by 10 with 2 added on. This produced good marking in assignment 2, but resulted in students copying the tutor in assignment 4, in effect defeating the exercise. In 2012, there was no grade awarded for students' patterns of marking – and in this year we saw a slightly higher effect size – but, certainly from my personal experience, the most successful version of the course. In 2013, however, for assignment 4 particularly, the effect size is high and the correlation is low. The major difference in this iteration from the previous ones is that the clickers were handed out anonymously, whereas in 2012, even though the student's pattern of marking did not contribute to their grade, nonetheless, the tutors did know who marked who. This matter is returned to in the following chapter.

Though these statistics tell us a lot about the validity of student marking and under what conditions we find greater or less validity of those marks, they don't tell us much about how participating in exercises such as this impacts on students' own learning and their ability to do higher order practice. In the next chapter, we will look in depth at the 2012 cohort including a focus group where students who participated in these sort of exercises experienced them and what effect it had on their work.

Chapter 4. The Experience of Peer Assessment on a Masters Course

In the previous chapter, the outcomes for four iterations of a course, and seven recorded cases of comparisons between tutor and student marks during real grade-bearing peer assessment sessions were presented. It was seen how the correlation between tutor and student marks was comparable and indeed superior to the reported measures in the major meta-studies by Falchikov and Goldfinch Falchikov and Goldfinch (2000) and Li et al. Li et al. (2015)

Table 4-1:Total r and d for all Judgements on all Assignments Delivered as Summative Assessment. (Excludes Rehearsal Events)

	Assignment 2 (Prototype) 2010,2011,2012,2013	Assignment 4 (Final Site) 2010,2012,2013 (2011 excluded)
Pearson Correlation (r)	0.77 (n=122 [61 groups x 2 criteria])	0.73 (n=126) [42 groups x 3 criteria]
Effect Size (d)	0.15	0.39

In the two peer assessment studies, the values reported were as follows.

*Table 4-2:Total r and d for all the two Major Meta-studies). *note effect size was calculated from s very small number of studies*

	Falchikov and Goldfinch (48 studies)	Li et al (269 comparisons from 70 studies)
Pearson Correlation (r)	0.69	0.63
Effect Size (d)	0.24*	(not calculated)

In any case, these results remain strikingly in line with the literature, and is nearer to the higher scoring values in Falchikov and Goldfinch (2000), rather than the lower scoring ones which appeared in Li et al. (2015)

4.1 Factors Affecting Higher Student/Tutor correlations

In the paper by Li et al. (2015), they tried to identify the factors that tend to lead to higher coincidence of tutor and student marks, concluding that:

> This correlation is significantly higher when (a) the peer assessment is paper-based rather than computer-assisted; (b) the subject area is not medical/clinical; (c) the course is graduate level rather than undergraduate or K-12; (d) individual work instead of group work is assessed; (e) the assessors and assessees are matched at random; (f) the peer assessment is voluntary instead of compulsory; (g) the peer assessment is non-anonymous; (h) peer raters provide both scores and

qualitative comments instead of only scores; and (i) peer raters are involved in developing the rating criteria

These conclusions are assessed in the context of the data outcomes of the current study, in Table 4-3 below, to ascertain whether they are pertinent to it.

Table 4-3:Factors in Li et al. Contributing to Higher Correlations between Tutor and Student marks: Whether they were of Relevance to the Current Study

Li et al.'s factors contributing to higher tutor student correlations	Presence in the Current Study
The peer assessment is paper-based rather than computer-assisted	NO: in the current case it is completely computer (clicker) based, however, it takes place synchronously. "Paper-based" might be a proxy for synchronicity and co-location
The subject area is not medical/clinical;	YES: our topic is hci and the thing to be assessed is its presentation by a group of students
The course is graduate level rather than undergraduate or K-12	YES: our course is graduate level
Individual work instead of group work is assessed	NO: it is group work that is assessed (though often peer assessment of group work in the literature is understood as intra-group rating of performance by members, rather than comparison of the outputs of groups)
The assessors and assessees are matched at random	YES and NO: in the literature, often a "sample" of the total population rates an assessee, in our case, 50-100% of students on the course rated other students. Consequently, there was no real "selection" of who marked who beyond who attended which session.
The peer assessment is voluntary instead of compulsory	In 2010 and 2011, student participation in peer assessment received some grade bearing value. Effect sizes were lower when participation was compulsory.
The peer assessment is non-anonymous	YES: a clear outcome from the research is that anonymity of reviewers is the strongest factor in the deterioration of marking quality
Peer raters provide both scores and qualitative comments instead of only scores	NO: at no time did any student provide qualitative comments
Peer raters are involved in developing the rating criteria	NO: the marking criteria were fixed over all four iterations and there was no student involvement in its creation.

Interestingly, Li et al also had a set of predictors that they eventually discarded for not achieving statistical significance at the 0.05 level.

Among the listed predictors in Table 1, eight were dropped eventually, including W2 (subject area is science/engineering), W4 (task is performance), W6 (level of course is K-12), W8 (number of peer raters is between 6 to 10), W9 (number of peer raters is larger than 10), W10 (number of teachers per assignment), W15 (there are explicit rating criteria), and W17 (peer raters receive training).

When these dropped predictors are examined, some do figure in the proceedings for the current study

Table 4-4: Non-significant Factors in Li et al. Contributing to Higher Correlations between Tutor and Student Marks: Pertinence to the current study

Factor	Is it relevant to the current study?
Subject area is science/engineering	YES
Task is performance	NO (performance in the literature typically means how well a student performed in a task as rated by their peers in a group over a period of time)
Level of course is K-12	NO
Number of peer raters is between 6 to 10	NO
Number of peer raters is larger than 10	YES (in the literature there is small improvement when there are more raters, however, Falchikov and Goldfinch believed this deteriorates when there are more than 20. However, this conclusion was from a very small sample (of studies with more than 20 raters per assessee) and in this study the average number of raters per assessee averaged around 30 for both assignments)
Number of teachers per assignment	YES – all iterations of this course had the same two teachers
There are explicit rating criteria	YES
Peer raters receive training	YES

Summing up, the correlation levels demonstrated across all of the peer assessed events on the course at the centre of this study across seven assessment events over four years are higher than those found in the papers reviewed by Li et al. and yet, a number of the features predicting higher correlations according to these authors are absent in our work (paper-based, individual work, voluntary participation, qualitative feedback and participation in developing the criteria). To try to understand what were the factors contributing to the positive experience of peer assessment across this course, and also, the reasons for the consistently and repeated high correlations, during the third iteration, in 2012, a focus group was held to gain understanding as to how the students experienced LPA.

4.2 Focus Group

Focus groups are a well known method in qualitative research. Bloor et al Bloor (2001)have suggested, among other things, they can tease out group norms and meanings, explore a topic and collect "group language", to clarify, extend, enrich or challenge data collected elsewhere as well as potentially even being utilised to feedback results to participants. In the current case, the aim was to find out group norms and meanings and also, to seek to explain further some of the numerical data gathered during the running of the assessments. Moreover, it was employed to try to contextualise the experience of peer-assessment to the students as well as the feelings it engendered in them. However, there are a number of criticisms that can be levelled at the focus group method. It can be overly led by the facilitator or certain participants can dominate it and it is difficult to generalise owing to the small size of the sample. Also, there is the criticism in the phrase "one shot case studies", that is, the focus group only represents the views of the attendees without any other group to represent a control group. There is also the related question of sampling bias: potentially those students who are the most positive about the topic are more likely to attend. However, we were not asking for a vote of confidence in the method, but rather some understanding of how it felt for any student participating in it.

The focus group was organised under research protocol 1112/51 involving nine students and three tutors. This course was very international and the focus group comprised two Europeans, three Asians and four African, with the gender split being two women and seven men. All were under 35 and the majority in were their 20s. Hence, it was a very diverse group. At the focus group a short introductory script was read out, but then we established a very wide ranging discussion in an attempt to understand how students experienced the focal assessment practices. The session was recorded by audio and subsequently transcribed.

The main purpose of the focus group was to tease out the student experience of live marking: the discussion was semi-structured. I sought to cover some of the themes of the literature (students' fear of bias, ideas of fairness, suspicious of peers' competence in marking, the value of training), however, some of the most illuminating insights came from some of the more casual and unplanned moments. For instance, when asking students whether they thought they were more generous or meaner than their peers when marking other students' work, I found out how they would calibrate their responses as they went along, being more generous to a group on a second criterion after they had been mean to a group on the previous one (comparing the score they gave to the average score given by the cohort).

4.2.1 Students' Initial Feelings about Peer Assessment

The discussion began with two questions – how did the students feel when they found out there would be peer assessment, and then, how was their experience regarding it?

> Student 4 - personally I think, I didn't feel comfortable with it, initially, because I thought that maybe a level of bias… sort of… when we actually did it. The question that impressed me was matching the clicker number with the student. That gives me a bit of confidence, that… at least you can know who is marking who, if the person is trying to be like biased, for example, somebody that did very well, and because you don't like that person's face, and you are marking that person extremely low, the teacher will know that this is not fair. So, that made me more comfortable with it, but initially I wasn't.
>
> Student 6 - initially when I also found out the marking system - like he said, I also felt like there may be some develop, bias. But when we went through the assessment criteria, like and the contribution made by students was just 20% and the tutors - like 80 - when I found out about that then my mind was put at rest, like. It's still the feeling like, it's more the lecturers than the students, and I believe the tutors won't be biased, so...I think it's better that way cos they have 80% and the students, they have 20%, so it's quite fair.

The two principles of accountability and also the proportion of the total marks given by tutors as opposed to students appeared to make students more accepting. In the literature around peer assessment, there is conflicting evidence about anonymity in peer assessment. In Lu and Bol (2007), anonymity in a purely formative peer review context appeared to be positive, because it allowed reviewers to be more critical, which resulted in better subsequent scores for the assessees. Xiao and Lucking (2008) study of anonymous peer review of creative writing assignments showed a very high correlation between student and tutor scores (0.829), however, anonymity here should be understood as "not revealed to the assessee"; the assessor still had to log into a system and make their grades. As the student in the focus group said: "matching the clicker number with the student - that gives me a bit of confidence - that at least you can know who is marking who". In other word, there was a clear felt sense of accountability in the peer assessment event.

Because the intervention was marked live (the whole class marking the presenting group), there was never a question of the anonymity of the assessees (the groups being rated), because who is being marked is overt. This might increase the likelihood of "reciprocity effects" (students agreeing to mark

generously in return for other students doing likewise). However, in the first student's remark, the fact that the user of the clicker was identified meant that there was also no anonymity at the level of the assessor (in the 2010, 2011 and 2012 iterations).

4.2.2 Fairness

The group was then asked, in practice, how did it go? Was there bias or not? Two responses to this question are particularly interesting, as one said:

> Student3: For me, most of the voting, I think, wasn't fair, not for me only, but also for the others… for I find there is some student, you know, they should have more than what they get and there is some students, unfortunately, maybe for their circumstances they didn't do very well, I think they shouldn't get that mark. So, I think most of the evaluations wasn't fair you know, especially from the students.

This response while seeming to indicate two classes of undermarking (quality being insufficiently appreciated, or lack-of-quality being excessively penalised), also indicates the centrality of a concept of "fairness" operating in the students' minds. Here, fairness not only relates to the accuracy of the mark, but also is extended beyond that to a kind of "value-added" understanding of academic performance. The student talks of "there is some students, unfortunately, maybe for their circumstances they didn't do very well".

Another student mitigated this with a sense that unfairnesses might be cancelled out in the overall coverage:

> Student2: It's obvious sometimes that there was some kind of erroneous marking. but the average on the whole levelled itself out… on the whole it seemed fair when you got the average mark - but when you look at the bar charts you wonder - how did that get there - maybe how the data was presented - I don't know.

This element of the data presentation is important. That is, the nature of the software meant that the total numbers of 1s, 2s,3 s, 4s and 5s are presented in a histogram – making it visible when, say, a mediocre presentation receives a number of 5s, or a very good one is awarded unexpected 1s. The fact that the student could use the word "obvious" to describe the phenomenon of erroneous marking means that it must have been experienced as self-evident. In some ways, this is due to the radical transparency of the practice: instant voting and breakdown of scores by votes per attainment criteria.

Error in marking can potentially be put down to three causes:

1. conscious bias or favouritism arising from students “gaming” the system, namely, agreements to award each other high marks (reciprocity);
2. unconscious bias (being more sympathetic to friends);
3. incapability (a real lack of understanding of quality in the field).

4.2.3 Reciprocity

Unprompted by the moderators of the focus group, the topic of reciprocity did come up, with one student saying

> Student7: "If I know four people, I say, please give me 5 rating ... they are doing ...
> Moderator: Seriously?
> Student7: yes its going on

Picking this up, a short time later, another student remarked:

> Student 5: Yeah, I think, if you go to someone and say we are presenting next time, I want you to give me 5, they will probably tell you yes. That I will give you a 5. But they won’t do it. They won’t do it! They will mark...based on what they see. They will tell you, because obviously they do not want to tell you no I am not going to give you a 5, they say ‘OK I will give you a 5’, since it is easy, since you wouldn’t know who are gonna click, what they will give you. So they'll just tell you yes, ok, but back of their mind they know they're still gonna give you the right...

Is this true, or just an opinion? In fact, the hypothesis of reciprocity effects (students agreeing to award a high mark to another student who reciprocates with a similarly high mark), is actually testable. Douglas Magin (2001) first attempted to measure these with a very ingenious procedure. In his case, he was evaluating how students rated each other in group projects (that is to say, intra group rating, members of the same group rating each other). These groups were quite large (10-11 people). Essentially on a presumption that each student in a group rates every other student, he made a a list of all the possible pairs in that group (e.g. student 2's rating of student 1, student 1's rating of student 2 and so on). He then computed the fisher Z score for each of those students’ ratings (in simple terms, how many standard deviations each student’s rating of another student deviated from that student rater’s own mean when making his ratings). Then Magin ran a correlation over those pairs of Fisher Z scores (correlation between each side in a list of all the valid pairs). In the examples he chose he found

evidence of such reciprocity to be vanishingly low as a proportion of all student rating practice, although there did appear odd moments where occasional pairs of students clearly had reciprocated.

Magin's data are, however, simpler than ours. In his case, one student awarded one global mark to each of the other students. In our case, the students (belonging to a group of 2-4 members) each mark each of the other groups and moreover, it is not a "global" mark. In assignment 2, there is a mark for (a) animation and then, (b) a mark for the site as a whole, whilst in assignment 4 the marks are for (a) visual effect, (b) suitability for audience and (c) usability. However, it is possible to replicate Magin's procedure by working out the average score that all the students in group A give to group B across all criteria, and vice versa. In that way, a group by group pairing can be made.

In the year in which this focus group was conducted (2012), both for assignment 2 and assignment 4, the Fisher Z score comes to approximately 0.2 - meaning there is some relationship, but again it is very small (and also, because it is just a 9x9 matrix, the sample is very small). Therefore, Magin's claim would appear to be borne out, i.e. whilst the students might have talked about reciprocating marks, there is little evidence that this did actually happen in 2012. If the levels of reciprocity in the previous years are examined, the same or even less is found.

Table 5: Magin's Reciprocity Test applied to reciprocal group marking

Year	2010		2011		2012	
Assignments	A2	A4	A2	A4	A2	A4
N Groups	22	22	18*	18*	9	9
Possible Pairs	231	231	153	153	36	36
Valid Pairs	114	143	140	117	36	36
Correlation between paired review scores	0.19	0.20	-0.01	0.11	0.22	0.19

* In the year 2011 there were 19 groups, but one group did not present live and thus, did not do any rating and hence, it has been excluded from the calculations.

In 2013, peer marking was done anonymously – and therefore it is impossible to do the reciprocity calculations.

Because the large number of groups in 2010 and 2011, the presentations had to be undertaken in two sessions, and not all students showed up for both. As a result, there are more examples for which a presenting group A did not receive any marks from members of group B, because the latter were not in the audience for the former's session. From these data, it does appear that there is a small or low level

element of reciprocation. Of course, in this context, for groups to have reciprocated it would have taken more organisation than the more casual reciprocation that might be possible among individuals (all students in group A would have to understand that they would needed to reciprocate the marks given by the students in group B and vice versa). Interestingly, in 2011, the year in which a premium was placed on the similarity of the student and teacher scoring, that element of reciprocation seems to have been reduced most effectively (in assignment 2), although it did have the negative outcome of the students copying the tutor marks in assignment 4.

4.2.4 Bias

Whilst students might have talked about reciprocating marks in the focus group, in practice this did not happen on a noticeable scale. However, there is the issue of more insidious evidence of unconscious bias. A recent study from a huge dataset of peer assessed work for a MOOC course on HCI (Human Computer Interaction) Piech et al. (2013) has found evidence of "patriotic grading" – namely 3.6% higher grades awarded to one's own country (if one excludes students from the USA the effect of own nation favouritism goes down to 1.98%). In this study, it would be impossible to check this since most groups were multi-national. However, for an on-campus course, where, unlike with MOOCs, students are known to each other and socialise together, it is not so much nationality as friendship that is of concern.

One student commenting on how their own group was marked compared to others said:

> Student 7: Yes, this happened! It's not fair, but if you take account, this is not what I want to say, but this thing happened. We are observing if people, didn't like me, they are putting in 1s, I've seen that also …I am not 100% sure they are giving the marks for the website, because some people, they have got friends, so that they are giving different marks"

There was also a case of a very high scoring group that was inexplicably awarded 1s by two of the students.

> Student 9- they must have, because someone gave us a 1
> Student 3 - that was just cruel
> Student 9 - there was two, two, two people gave us 1
> Student 5 - maybe that was an error

This is a possible explanation - students may indeed have simply pressed in the wrong number by accident. One of the dangers of live marking is that it is impossible to shield students from these elements of friendship or antipathy based bias. Also, the extent of this is almost impossible to quantify

since there would need to be some measure of affinity and antipathy between each student to measure against the overall scores and ethically, that would probably be very difficult to obtain, and if obtained, very unlikely to be trustworthy. A practical solution (in the class) might be to show the overall average without the banding of scores by number. This, however, is less informative than banded scores, and paints a less rich picture. Nonetheless, as a way of protecting the students from malevolence or errors in scoring it might be valuable and the broken down marks could, instead, be sent with the moderated feedback.

4.2.5 Marking Competence

In the literature, the third reason often cited for non-robust marking is marking competence. In this focus group, very little was said about suspect judgement on the part of peers, although one student did warn against the practice being applied to undergraduates.

> Student 4: I think that undergraduates are not academically matured enough to be assessed through that kind of process. I may be wrong but that is just my personal opinion.

Another thing that might have influenced students' marking is the order in which the presentations were given. One student said:

> Student 6: Can I say something? I think basically what happened was that people were voting based on other people's website - not really based on criteria, especially the first one… the first evaluation we did… basing on if you had a very good website and that was like 5. Then, all the other ones are marked based on that really good one… not really based on what was on the sheet… so it was over… that was my own thought for the first one.

Another student responding to this said that all the presentations were marked in comparison with the "a"s. In this case, for the sake of fairness, the policy was to establish an order for the first peer evaluation event and then reverse it for the second. However, it does also show that no matter how "criteria referenced" the marking is, normative effects will also happen, but these normative effects will nonetheless have their educational potential – namely show what can be achieved by students within the same cohort.

4.2.6 Tutor Influence

As can be seen, fairness is the absolute touchstone around which peer assessment is considered by students. However, there was one final influence on the fairness of the marking discussed in the focus

group that is specific to the live feedback event, namely, the role of the tutors in chairing the presentation and the voting. One student said:

> Student 4: What would have influenced me was if the lecturer sort of led me to have how can I put it now a prejudgment… like maybe like I say oh that's a nice site. Then probably that would have influenced me, but it didn't happen throughout. I mean the lecturers would kept quiet and left us to judge what we think about a site.

This was indeed a big consideration for the tutors when running these events. In a way, at this point the educational and the equity concerns are at odds. By giving any sort of feedback on a site prior to the collection of the marks, the tutors would be influencing the marking by the students and they would also be pointing out the flaws and achievements in the work being presented. Consequently, even while the tutors tried to maintain a poker face during these events, some influence was nonetheless transmitted. Below is an extended citation of the focus group discussion, provided here, not only since it deals with how tutors influenced student marks, but more interestingly, how some of the students "read" the teachers:

> Moderator2: - Do the questions influence you do you think, or the way we ask them? We try not to, but it's difficult, you know. We try to ask fairly neutral questions rather than leading questions.
>
> Student 3 - Usually sometimes you catch something there… especially you tell them OK can you click on that… it doesn't work. Of course, we know that we have to keep that in our mind you know?
>
> Student 4 - If you master the pattern, the questions that are being asked. I can know the kind of questions Tutor A will ask or you [Tutor B]. I talk to Student A and I say Tutor A will pick that and he picked it!
>
> Moderator 1 - That is amazing.
>
> Student 3 - We started laughing then.
>
> Student 4 - Yeah, we started laughing, when he picked it. I said I told you!... because we know the pattern, we know what he likes, we know what he will say, oh, that's not right, yes, you can study it, the pattern of questions.
>
> Student 9 - Yes, do some psychology!
>
> Moderator 1 - Did any of the other groups, did any of you, when you were doing your…
>
> Student 9 - Some of your comments also influenced our documentation, because the kind of

> questions you asked, the kind of... when you see the site and ask about it, they influence what we are going to write about it, and if I can agree with what he [Student B] is saying. He wrote so many things, like, oh these are the kinds of words Tutor A, likes, these are the kinds of words that you [the Tutors] would like us to use. So yes we picked up quite a bit of your...

This shows that the students could anticipate what the tutors would point out, and would even craft their documentation around some of the vocabulary used by the tutors. However, the comment by the student above, shows that tutors hold enormous influence even when they are trying to conceal their true opinions. That being said, it also appears that certain students seem more able to "read" tutors than others.

In fact. this ability was first pointed out in 1974 by Miller and Parlett in a famous report "Up to the mark: a study of the examination game" Miller and Parlett (1974). In this they categorised students as "cue-seeking", "cue-conscious", and "cue-deaf". The first category being those students who would actively engage tutors, try to tease out the likely questions in exams. The second group were ones aware of there being a "technique" to high academic performance, who were not quite as proactive as the first but who nonetheless thought on similar lines. The final group, the "cue-deaf", merely believed in hard work. This "cue-deaf" group, for example, believed in revising comprehensively for exams (covering all the material), while the first two groups ("cue-seeking" and "cue-conscious") were more keen to know where to focus their attention. What is important to recognise here in this categorisation is that it does not do so by virtue of their level of engagement (active, passive or strategic learner), but rather, the extent to which students believe high attainment is a game that rewards a focussed strategy. In the table produced by Miller and Parlett, the cue-seekers did the best, the cue conscious came next and the cue deaf did the least well. One of the students (in 1974), whilst talking about preparing for an examination, spoke in a remarkably similar way to the students in the focus group:

> "The technique involves knowing what's going to be in the exam and how it's going to be marked. You can acquire these techniques from sitting in the lecturer's class, getting ideas from his point of view, the form of his notes and the books he has written – this is separate to picking up the actual work content"

Another student from that report also said "everybody's looking for hints, but some of us look for it or probe for it, rather than waiting for it to come passively. We actively go and seek it…"

However, it maybe that it is not only any unconscious "tells" coming from the tutors that may have influenced the students, but also, the selection of a particular part of the students' presentation as a topic for a question. One student said:

> I think it would be worse - I think just small little comments - because not everyone gets the same concepts as everyone else… everyone has their own perception. So, it wasn't sort of leading questions at all, as for me, *it was just like identifying the main things that you saw*, in case, it wasn't like leading to us, was just, wanted to know exactly what they did, for you, because you are also voting, wanted to give the marks, *so I am sure you were asking to assist you in your marking*. For us, what we saw, we give our own marks was influenced or not influenced, so I think it's still ok to give the small comments that you gave… wasn't too much, wasn't leading.

Hence, it is not just any unconscious betrayals of opinion, for it is also the mere isolation of items of interest that can establish a context for marking, and the act of putting some features into relief above others. So, despite the practice of live peer assessment attempting to render visible and transparent all the factors in evaluation, there still remain some things that require a skilled eye to read.

There is potentially a level of openness in the social marking process, which means that the tacit elements of quality become manifested more directly and the things that matter get given greater relief. However, these tacit things are capable of being read by some students more readily than others.

At this point it is worth also considering the extent to which the descriptions of quality in rubrics and marking sheets with their attainment descriptors, and never really get to the reality of "quality" which appears to be transmitted "behind the scenes" or tacitly. In the discussion cited above, we come across statements such as: "words which Trevor likes" – the "pattern of questions" the "kind of questions you asked" "we picked up quite a bit of your…" , "you catch something there": noticeably there is no reference at all to the explicit rubric, or even the prompts which appeared onscreen to which the students gave ratings. Instead it is a decoding of the tutors' response to artefact being presented which enables the students to construct their own judgement matrix.

4.2.7 Training

Here, the prospective rather than retrospective aspects of peer assessment, and the effect on the rater rather than the ratee are considered. From the last exchange, it can be seen that the act of peer marking is not merely some enfranchisement of the students or some labour saving device for tutors,

for it also can contribute to the learners' own conceptions of quality. We have seen how the chairing of the event can influence students' sense of quality, but the live experience of marking will also put students under the influence of their peers, because every time they mark they also see the marks of the others. So, having investigated the negative influences on assessees (reciprocity, bias and incapability), it is now time to examine the positive ones: guidance from the tutor and the effect of their peers marking at the same time (therefore having the opportunity to calibrate one's own judgement).

The effect of explicit tutor guidance (rather than reading of tutors' unconscious "tells") is exercised most clearly in the training or rehearsal sessions organised before the peer assessment events. The difference between the two events was put most clearly by one of the students as follows:

> Student 3: "Actually when we do the answer, there is something interesting, because whenever we do the voting, after you say, this is the same mark we get from the students last year. It was more fair that I know that the students they are voting very well and they are voting closely to what they get last year. But when we vote for each other, of course there is no reference mark for others and that make me uncomfortable about that. So, whenever you said this is really near or the same mark as last year, that was really fair, but in the real situation when we are voting for each other we don't have that...situation"

In this quote, the most important line is "*when we vote for each other, of course there is no reference mark for the others*". The training or guidance event is firstly important for establishing the confidence of students in the practice (who of course come to courses without a grounding in the literature of peer assessment).

Most of the literature on peer assessment mentions the importance of training. In all four iterations of the course, before the first marking event we ran a training event where students were asked to mark previous students' work. On these occasions, there was no need for the tutors to be uncommunicative about the work. because it was precisely here that we wished to point out what quality consisted of, where it was revealed, what kind of effort it involved and awareness of what kinds of principles needed to be carried out to produce an excellent piece of work.

How did students experience the training? One student said:

> Student 5: Because we also practice and look through the criteria as well and OK, after the rehearsal then I know that OK… for this argument, 5, that means it's like an excellent work, like it's satisfied all the criteria. So, by the time I get to the real thing, when I click I have to go back

> to the sheets to… 3 means, good animation and 3 means good whatever, which would like slow down the pace and we only have like 20 seconds to vote. So, after we did the demo one, I have everything in my head. So, once I just see the work, I know what mark to give… it's like 2 or 3 or 4 or 5. It was pretty helpful.

This is extremely interesting, because it shows the process of internalisation of criteria; how at a novice stage the student is seeking correspondence with explicit marking criteria and how subsequently, it becomes instantaneous. In the training session, the students said that while marking he had to consult the sheets to get the score right, but then eventually he said "*once I just see the work, I know what mark to give*". Much of the literature on peer assessment, particularly Cheng and Warren (1997), emphasise the importance of training to the success in the use of peer assessment and here, a key reason why it is necessary is revealed. That is, to establish the internalisation of the criteria, which allows for a very swift judgement. This judgemental immediacy is something that ideally should also inform the student's decisions and practice when evaluating their own work – that is to say, an instinctive "feel" about when something is right or wrong.

4.2.8 Peer Influence

If the training sessions then give this immediacy, there is something else that checks and informs it during the live voting, that is, the opinions of the rest of the class. When asking the focus group if they considered their own voting more severe or more lenient than the rest of the class, two students responded very interestingly:

> Student 2 - Now, I dunno whether sometimes, I felt if I gave maybe a low mark and I saw the class gave a high mark that may alter my opinion for the subsequent question they deserve a bit more because they, I don't know, maybe I am underestimating the site… that's what everyone else thinks, so I thought, if you maybe ask all the questions first then show the results afterwards.

> Student 9 - Um, in regards to what he was saying, I think that could be true, because, for example, you voted the first page that you see you have given maybe a 5 and many people have given 2. So, the next time you think I can't give a 5 again and I have to give a 3 or maybe…

These are possibly some of the most interesting statements during the focus group. It shows the social power of the live marking effect, which induces almost immediate reflection regarding the students' own judgements. This means that they were not just making immediate decisions, for they were also

reflecting on these decisions. In a sense, what can be seen here is Kolb's feedback cycle taking place at the most minute and intimate level of personal intuition or gut-feeling. The interplay of the concepts of fairness and equity in the students' minds would appear to influencing the refining of their understanding of quality.

Ultimately, the one major difference between the practices on this course and much else in the literature around peer assessment is its synchronous and public quality. This has a number of inevitable corollaries: namely a much higher assessor to assessee ratio (typically 30:1), immediate (unmoderated) response, very fast full (moderated) response (where the tutor mark is added and weighted against the student one), but also feedback on judgements made. In traditional peer assessment, an assessor does not see the scores given by the other assessors, so there is no way to know if one's own assessments are more lenient or severe than those given by others. In the literature, there are a number of reports of correlations higher than those obtained here, and of course, many lower too. What, however, has not to my knowledge been reported, is the consistent and repeated obtaining of these kinds of correlations over four years and seven assessments, with larger and smaller cohorts. The fact that this could happen so consistently must have something to do with the live nature of this peer assessment, and the fact that students must be self-calibrating as they participate.

A surprising affirmation of this point might be found in studies relating to "inter-rater reliability". One particular study, evaluating the marking patterns of figure-skating judges during in the 1984 Olympics, was completed by Weekley and Gier (1989). The judging of figure-skating constitutes a highly analogous activity to peer evaluation using clickers, in that it involves numerically judging the results of a performance in a public fashion immediately after a performance. Moreover, the judgement and the results are revealed instantaneously, both to the performers and the crowd. They found, notwithstanding the strong partisanship evident at the height of the cold war, the multi-national judges (who may have comprised antagonistic nations), nonetheless, managed to achieve a remarkable level of agreement while marking. As to why this should be the case, the authors wrote:

> "One answer may lie in the availability of constant feedback. Recall that after each skater's performance the ratings from all judges were posted for public scrutiny. As a result, judges could see how their evaluations compared with those of their peers. Judges with idiosyncratic standards should have been easily identifiable and the direction of their divergence from the norm readily apparent. This self-others comparison process, which is the basis of much of the judges' training"

This effect and how it was also experienced in a humanities course, which is outside of the scope of this study, can be found in the following paper by Bennett et al (2015)

4.2.9 Student Experience

Finally, how much did the students enjoy the whole process overall? And what claims can be made for peer assessment in courses dealing with the creation of rich technical artefacts?

One student said:

> Student 9: I think if you compare this with the other modules ... because you are not going to present it to anyone; it's just the lecturers who are going to see it. Here you are going to present to your peers, so it gives you more motivation. You have to do a spectacular, you know, website, for everyone to see, so if I imagine we weren't going to do it in front of other people, yes, we would probably have been good, but the motivation of working together, the enthusiasm, would not have been as high as expecting to be there to illustrate what you have done.

So, while the students did have a fear that there may have been unfairness in peer evaluation, there does seem to also have been real excitement about presenting to the class. Another student mentioned the speed of feedback:

> Student 6: Another thing is, um, the practice… I didn't know the feedback was going to be immediate. I thought it was going to be get the result like a week later and so, when I got the feedback immediately I think it was really good. That's something you don't really get from most other modules… you don't get instant feedback; it takes weeks.

Another thing that was valued was exposure to the work of peers and to see what other students are capable of. Commenting on the value of the rehearsal for establishing standards, one student described their response as:

> Student 9: But if you see, say ok, like the way we were saying, oh wow, we have to do that, we have to try and do that, and the way we did the rehearsal, we saw the kinds of things like the x button to close, like all these other things. Oh my god, we try and do that, we can try and if it doesn't work, then at least we have sort of an idea of how to expand our project or assignment.

Undoubtedly, the 2012 iteration was the high water mark of the course. The marks for similarity to the tutor had been removed, but a log of who gave who what was still kept and the work produced that year was of a very high standard.

4.3 Conclusion

Summing up, the benefits of live peer evaluation as reported by the students were:

- Engagement arising from the sense of occasion;
- Speed of Feedback;
- Exposure to Exemplars.

Moreover, it has been shown how it delivers better work through the internalisation of an idea of quality established through the exemplars. This internalisation seems to be the result of a rich ensemble of influence effects that take place through training and practice in peer evaluation, including the illustrated guidance of the tutors, the establishment of an immediacy of judgement on the part of the students, the unconscious influence of tutors in the isolation of particular parts of assignments or in even the terms used as well as the impact of the rest of the class on the student when they compare how the class judged with how they themselves did so. The potential pitfalls of peer evaluation have also been discussed, which pertain to: reciprocity effects (occasionally present but not common), friendship bias and also a kind of incapability on the parts of some students to distinguish the good from the bad. However, it has also been demonstrated how the collective nature of many markers led to some of these potential problems being diminished. Given this is the case, it is important to consider why the 2013 iteration of the course produced marking of a much less trustworthy quality?

The answer to this was very clear in this focus group itself, but unfortunately, something we did not place sufficient emphasis on during the following iteration of the course. Repeating again what one student said:

> Student 4: I thought that maybe a level of bias… sort of… when we actually did it. The question that impressed me was matching the clicker number with the student. That gives me a bit of confidence, *that… at least you can know who is marking who, if the person is trying to be like biased*, for example, somebody that did very well, and because you don't like that person's face, and you are marking that person extremely low, the teacher will know that this is not fair. So, that made me more comfortable with it, but initially I wasn't.

During the 2013 (and final) iteration of the course, we merely handed out clickers without keeping a record of which person had which clicker. This almost certainly is the principal factor determining the anomalously high effect sizes in the difference between tutors and students in that iteration of that course (although the correlation remained high) Consequently, it is clear that some sense of accountability of student raters for their ratings remains important, even if it is not measured against a canonical tutor mark.

What can be seen from the above, is the absolute centrality of the tutor in the establishment of rating competence: through the interactions between tutors and students through training events, through the selection of exemplars, through the highlighting of certain features of those exemplars, student competence in evaluation begins to grow and become more automatic. Moreover, it is through seeing their own evaluations in the context of the evaluations of others that that competence can be refined.

As was shown in summary of the most recent meta-study of peer assessment conducted by Li et al. (2015), one of the surprising findings of this work is that paper based peer assessment is more likely to be reliable than computer based. It is my belief that this variable (the paper medium) could potentially be a proxy for other variables, rather than just the idea of paper keeping people honest or deliberative. What it might be potentially is this concept of *likely oversight or seriousness*. If students are given a sense that their ratings will be scrutinised and that these ratings will say something about them (the rater) and not just the ratee, then that is going to result in more *conscious* rating. Nonetheless, I have demonstrated that it is possible to achieve high levels of correlation in a purely electronic medium.

It is concluded that, live peer assessment using clickers is an effective technique for enabling students to get a better understanding of quality in a technical discipline. However, it needs to be undertaken according to the following guidelines:

- Always make sure you know who rates who;
- Simplify the rubric to make it more comprehensible and restrict the number of dimensions;
- As a rule of thumb, do not give extra marks for closeness to the tutors, but potentially penalise plainly aberrant marking (e.g. less than 0.2 correlation with tutors, very low standard deviations, excessive maximums);
- Always run training events through a careful selection of exemplars (ideally ones of a variety of levels, but ideally without high divergence among previous markers);
- Only do live peer evaluation with student groups, not individuals;

- In live peer evaluation show averages, rather than banded scores (when live). Send out banded scores with subsequently moderated feedback;
- It is recommended to keep peer scores to no more than 20% of the assignment as a whole and no more than 10% of a module;
- When chairing a live event, try to be as neutral as possible;
- Evaluate peer evaluation scores by referencing
    - Correlation (ideally around or above 0.65)
    - Pair-wise Correlation Median (again, around 0.65)
    - Effect size (ideally below 0.3)
    - Krippendorf's Alpha (ideally above 0.3 on rubrics of highly subjective and synthetic criteria. Higher thresholds can be used on rubrics of a more mechanical nature)
    - Magin's Fisher Z Reciprocity Test (ideally below 0.3)
    - Skewness and Kurtosis of the distribution of individual student vs tutor correlations; neither should be outside the range -2 to +2

It has been shown how live peer assessment can contribute to greater engagement on a course and that its positive factors are not just through the speed and quantity of the feedback received, for it provides benefits to the process of learning how to evaluate, and in those acts of evaluation themselves. As Graham Gibbs (2004) wrote:

> Much of the literature on the use of self- and peer-assessment is about the reliability of such marking, and assumes that self- and peer-assessment is primarily a labour-saving device. But the real value may lie in students internalizing the standards expected so that they can supervise themselves and improve the quality of their own assignments prior to submitting them.

In the next chapter, I turn to reporting on the outcomes from the other course used that these techniques (live social evaluation of technical artefacts using clickers), not for the grading of peers, but rather, for the grading of previous students in a process designed to help students engage with the assignment criteria, in particular, to "internalize the standards expected". What has characterised this study (of the master's course), is the evaluation of a large number of submissions (61 prototypes and 61 final artefacts in the purely summative events) by 61 groups, in comparison with the marks of two tutors who remained constant throughout the process, and a set of marking criteria which were also constant throughout four academic iterations. All of this was for grade bearing, summative, live peer assessment. In the next study, investigating the creation of multimedia artefacts on a first year undergraduate

course, we will have instead a small pool of exemplar work and a much larger cohort (>180 students per iteration). In this case, we will see students marking exemplar work, not to give grades, but to help them better understand assessment criteria, and more broadly the concept of quality as it applies to multimedia development.

Chapter 5. E-Media Design – Longitudinal Analysis

So far, how live peer evaluation is experienced by masters students in computer science has been investigated. This was undertaken on a medium sized course with a cohort size varying between 30 and 60 over four iterations. A higher than expected correlation between student and tutor marks was found, an increase in engagement, and also a sense from the focus group that the act of grading other students, and the practice of grading previous students, led to a heightened and intuitive sense of quality in the discipline. Having applied this technique for the MSc course, how to do so with a much larger course (average of 200+ students) of first year computer science students was then considered. Whilst for the MSc course, the primary driver behind the practice was speed and immediacy of feedback (students receiving feedback from their peers instantly and moderated feedback from the lecturer within two days), for the undergraduate course the aim was to address an issue of poor design practices in student artefacts.

5.1 The Course and Differences from the Previous Study

The student artefacts on the course were multimedia CVs (résumés). Students had to use the tool Flash, to produce a visual CV with a very small file size, thus exposing them to the constraints around multimedia development. In terms of the artefacts produced, put simply, despite the students being told in the lectures of the importance of the alignment of buttons, the consistency of fonts, consistency of a navigational interface and the placing of images, in practice, many of them just did not apply them to their artefacts. This did not apply to all students, but there was a very high proportion of substandard work. In the two years prior to the use of live peer assessment, the overall spread of marks was as presented below in Figure 5-1.

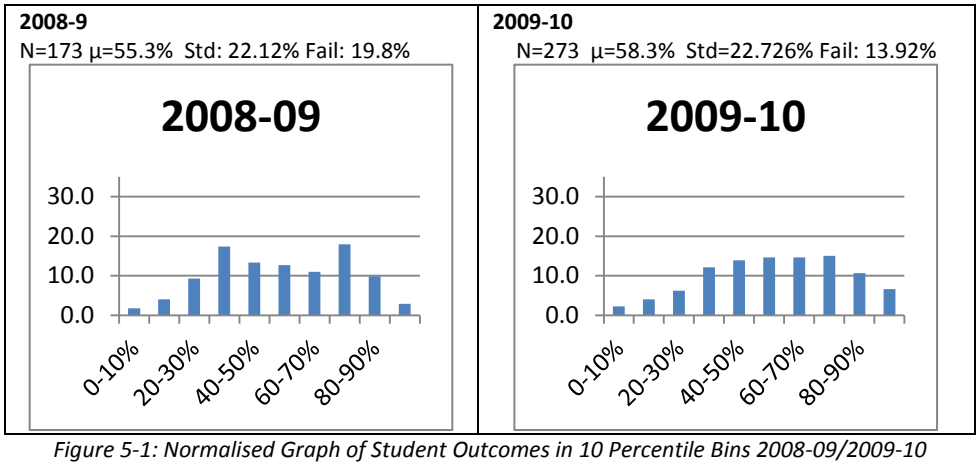


Figure 5-1: Normalised Graph of Student Outcomes in 10 Percentile Bins 2008-09/2009-10

In these years, a small improvement in the overall average, as well as the proportion of failing students was attested to, but it was still not particularly satisfactory.

In the masters module, the practices described relating to the use of clickers were largely stable over the four iterations. Certainly, the rehearsal sessions appeared to become more polished, but there was variation in results owing to particular defects in specific iterations, namely: the copying of the tutors in assignment 4 of the second iteration of the course (2011-12) and potentially the anonymity (and thereby the reduction in accountability) in the fourth iteration (2013-14).

The practices used for the BSc course were less uniform; student satisfaction rose and waned, rubrics were edited and amended and the methods for ensuring participation in the feed-forward events also varied. Moreover, the training sample used was very different in the final iteration to all the previous ones. However, in one sense it was also much clearer. Unlike the previous study, the outputs were individual pieces of work (multimedia CVs) and therefore, provided a clear measure of the effect of feed-forward on individual students' work in terms of actual attainment. There was also a prior individual piece of work (multiple choice test) that could operate as a baseline.

In order to structure the report in a way that takes account of its evolution and mutability, in this chapter I describe how the course ran over the four iterations, the changes introduced and the reflections which occasioned them together with the instruments and practices that were used. This thus takes the form of a narrative, as experienced by one of the course tutors, and represents a reflective understanding of the challenges and responses to them. A more rigorous analysis of the data will take place in the next chapter.

5.2 E-Media Design: The Module

The aims of the module were to understand what motivates design decisions, to appreciate the importance of creating systems that are fit for their intended purpose and for learners to make straightforward design decisions of their own. Arising from this, the course had the following learning outcomes:

Table 5-1:E-Media Design Module - Learning Outcomes

The learning outcomes (knowledge and understanding) were to:
• understand the reasoning behind some of the decisions that have been made in the design of existing techniques and technologies for storing, transmitting and processing information

- understand the relationship between form, function, content and aesthetics, and their importance in the design of documents and the systems that manipulate them
- understand some of the options that are available to those designing and implementing systems for the storage, transmission and presentation of information

The skills to be developed included to:

- be able to make straightforward design decisions, taking into account the relationship between form, function, content and aesthetics
- make an informed choice between different means of representing, transmitting and processing information

In order to achieve these objectives, learners were required to follow a programme of lectures and readings related to the design, implementation and evaluation of electronic media, supported by practical work pertaining to the designing of a multimedia CV. In the 2008-09 and 2009-10 academic years, the assessment regime worked as follows

Table 5-2: Weightings By Assignment2009-10

1	In class multiple choice test (technical)	25
2	In class multiple choice test (theory)	25
3	Flash CV (developed independently, but required students to make final amendments as specified by the tutors in a one hour sitting under exam conditions)	50

The learning outcomes were assessed by two multiple choice objective tests and a practical assignment requiring the students to undertake substantial independent work, and then make some last minute amendments in a lab (under exam conditions). This was introduced to ensure the learning outcomes had been reached and to discourage any kind of "contract cheating" (students getting others to do their work for them). On this course, students attended 12 practical sessions, where they developed the necessary skills to create their multimedia CV, which involved coverage of most of the design and animation features of Adobe Flash. These were practical sessions usually with 40 or so students each. The learners also attended six lectures on the basic theory of electronic media and multimedia production along with six lectures on aspects of software usability, screen design and software evaluation. Each week, they were given set compulsory set reading as well as some optional readings. In addition, the learners had one lecture supporting the development of their multimedia CV, including making explicit the marking scheme for the final practical assignment. This included the presentation

and discussion of examples by tutors. The learners were then given the brief for the final assignment, practical test 3, whereby they were required to develop a multimedia CV based on it.

The intention with the re-engineered module, was to replace the second multiple choice test with a feed-forward evaluation exercise, where the students would mark previous students' work. This assignment was to be scored according to how similar the students' marking was to the tutors'. This was based on the level of agreement per item (criterion applied to artefact) between the student and tutor marker. The way this level of agreement was defined developed over the different iterations of the module. Just as for the MSc module, there was also a training session where the students were given guidance regarding how the tutors graded the various multimedia CVs.

5.3 First Iteration with Clickers and Exemplar Marking: 2010-11

This year, as indicated above, the lecture supporting the development of the CV artefact was replaced by an EVS session evaluating the previous year's multimedia CVs. Soon after the EVS session, the students had an evaluation exercise, instead of the previous year's coursework 2. In this first iteration using feed-forward, there weren't enough clickers to perform the evaluation exercise in a lecture theatre and so Questionmark Perception (an online quiz, survey and questionnaire platform) was used. In this exercise, students were required to peer review a sample of last year's multimedia CVs, using the tutors' marking scheme and were scored according to how near their marks were to those given by the tutors. The assessment regime now looked as follows.

Table 5-3: Weightings by Assignment 2010-11

1	In class multiple choice test (technical)	25
2	In class evaluation of previous students' work (using QuestionMark Perception)	25
3	Flash CV (developed independently, but required students to make final amendments, as specified by the tutors, in a one hour sitting, under exam conditions)	50

During the EVS "practice" session in the lecture theatre, where students practiced marking previous students' work, they worked in groups of about four. The tutors presented sample CV artefacts from the work of the previous cohort and displayed them on screens. These artefacts had been anonymised to remove all indication of the student who had originally produced it. Moreover, all graphics were replaced by Simpsons' cartoon character graphics, and details regarding schools and jobs were replaced

with fictional ones. The students were asked to mark the multimedia CVs per the criteria used in the marking scheme, using the EVS. This was a fairly simple rubric, which contained a series of statements to which the students had to say *Yes*, *No* or *Maybe*. Below is the full list of criteria (used by the tutor when marking). Four of these criteria were not recorded in the student test: those relating to sound since sound was not working in the labs at the time of the test (two criteria). There we also two other criteria - width/height and file size - also omitted, because information about these could not be reliably presented to the students during the test (the width/height of the file could have been scaled by the browser, and the file-size would not have been revealed in the browser since the file was hosted externally).

Table 5-4: 2010-2011 Rubric used by Tutor + Restricted Used by Students: Henceforth referred to as 2010T and 2010S, respectively

Full List of Criteria	Used by Students	Notes
1. Has correct number of pages with correct headings on each	x	
2. Correct background colour	x	
3. Correct width and height of the Flash file		Not used by students, because width and height is fixed when in an html page
4. Correct number of buttons with correct colours for them	x	
5. Make buttons navigate to correct frames using simple action script	x	
6. Contains at least two images of you	x	
7. Small file-size		Not used by students, because data unavailable to students during test
8. Motion Tweening Of Position of Things in the animation welcome screen OR Motion Tweening of Visibility of Things (for fade ins and fade outs) in the animation welcome screen	x	
9. Correct and nice positioning of buttons and content	x	
10. Good easy on the eye content (text and image); not too little not too much and all relevant	x	
11. Button clicks have small sounds associated with them		Not used by students, because sound didn't work in labs
12. Some very clever and visually elegant animation (using Shape Tweens or Motion Guides or Masking) in the animation welcome screen	x	
13. Extremely well positioned and pleasant looking buttons	x	

14. Extremely well judged content	x	
15. An immediately visible and functioning background music toggle		Not used by students, because sound didn't work in labs

In the earlier "practice" session, after each presentation, the tutors then discussed their marks and compared them with those of the students. Whilst we did not take an attendance register during that session, large numbers of students turned up and it was obvious to the tutors that the groups of learners were actively engaged in marking and discussion among themselves and with the tutors. It was clear from those discussions that the learners were actively applying the marking criteria, not only to the work of their peers, but also to their own. At times, there were quite significant differences of opinion between the student cohorts and the tutors, which I considered to be a positive aspect of the initiative. An early paper covering the initial practice in more detail can be found at (Bennett & Barker, 2011). The purpose of this session, was not to get a totally accurate reading of the class's opinion, but rather, to initiate a discussion of the assessment, and how marks were allocated. Doing so, required the students to clarify their idea of the meaning of the various criteria and to compare these interpretations with the tutors' own. Owing to the lack of experience of the tutors with the clickers, and a number of technical issues, no data regarding student voting was collected at this session. In the week, prior to the true feed-forward evaluation session, the students were also given access to an online version of these evaluation rubrics in Questionmark Perception, where they were given four other pieces of prior student work to evaluate (formatively), which they could compare with the tutor judgement.

For the summative test of the students' evaluative ability, they were to receive marks based on how near their mark was to that given by the tutors. Given that some of the statements were self-evident (see table 5.4 above), this was not a particularly searching assignment. The average score was 69.44% with a standard deviation of 12.58%, with only four students being below 35% (the pass-fail boundary for the module overall). The generosity of the test was important such that students would not receive large penalties for disagreeing with the tutors. Nonetheless, some of the students were unsatisfied. The discussion forum had a long thread, where students wondered about the legitimacy of marking someone based on coincidence of opinion. However, alongside that was an extraordinary increase in the marks for the multimedia CV assignment submitted at the end of the course. Using the same criteria as the year before, an increase in the student average of over 7% was observed along with a reduction in the number of students scoring <35% (the pass fail boundary for that assignment at the time) to

4.81% (from 13.92%). It appeared, that notwithstanding the fact that the rubric was a little basic and, the technology not optimal (having to use both clickers and Questionmark Perception), that the exposure to the marking criteria and its use seemed to have led to real improvement among the students. Here is the normalised graph showing student attainment in bins of 10% across the cohort.

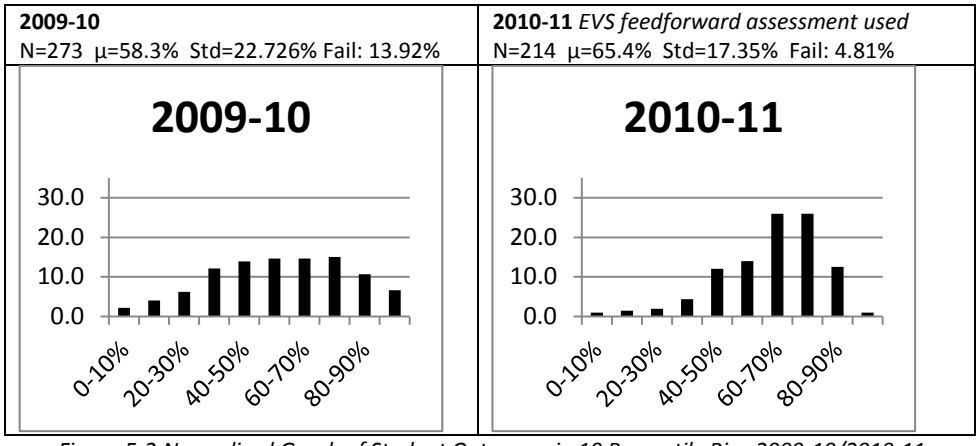


Figure 5-2:Normalised Graph of Student Outcomes in 10 Percentile Bins 2009-10/2010-11

5.4 Second Iteration with Clickers and Exemplar Marking: 2011-12

While undoubtedly, the introduction of exemplar marking had been a success during the first iteration, there was some concern about the fact that it did produce some negativity on the VLE forums. Despite the fact that the students who did complain were a small percentage, some of the complaints were valid. Take, for instance, the statement in the marking rubric: *Contains at least two images of you*. How is one to apply this to a case where an exemplar only has one image of the student? Clearly, one cannot agree and say yes. If one is to be literal about it, then the answer has to be no. However, most academics would probably want to give some credit and so, as in this case, the answer *maybe* was given. But to say an artefact "maybe" has two images of the student (when it has one) does not really make logical sense. But, that is very much a product of the mindset of the experienced academic who seeks to give credit to the student, even when the rubric does not appear to offer that option. This kind of confusion, in fact, encapsulates the essence of the difficulty of communicating marking criteria, i.e. the literal text says one thing, but typically how it is interpreted belongs to the community of practice engaged in the exercise. Wenger (2011) represented this as the participation/reification duality of such communities. While all such communities have their rules, policies, documents, the things that encode the practices they try to bolster and support, nonetheless, these always receive dynamic interpretation of nuances of meaning and foregrounded elements or conversely, are regarded as "small-print".

In this iteration of the course, I and my fellow tutor were still working under the mind-set that the nature of the rubric itself was fundamental to how the students would be able to mark and evaluate their work, and that the clearer, the more compact and expressive it was, the more reliable would be the evaluations based upon it. That is to say, we had the belief that the rubric could be, in some sense, self-sufficient and any misunderstanding the students had when applying marks to things would merely be the result of an imperfectly drafted rubric. However, as per the example above, the rubric could always contain items that require some kind of interpretation over and above the explicit text and while that text could have been improved, that is to say, we could have offered "only 1 image" as an attainment descriptor, how would we have dealt with say, two images, one of which was corrupted such that the image did not appear perfectly? It seems impossible to characterise all possible levels of satisfaction a criterion might occasion in advance. Consequently, while clarity is certainly important, it doesn't quite tell the whole story, that is to say, the truth is that a rubric or marking scheme can never be entirely self-sufficient. However, my realisation of this only came later.

In the second implementation of the course, the first difference to the previous year was that the proportions in the marking were rearranged.

Table 5-5: 2011-12 Weightings by Assignment

1	In class multiple choice test (technical)	30
2	In class evaluation of previous students work (using clickers)	10
3	Flash CV (developed independently but required students to make final amendments as specified by the tutors in a one hour sitting under exam conditions)	60

Another change was that in this academic year, the University of Hertfordshire began distributing EVS clicker devices to all new students. This meant that Questionmark Perception no longer needed to be used to perform the evaluation feed-forward exercise, and that both the rehearsal and the summative feed-forward evaluation exercise could take place in the same format. A further change was that we decided to take more care regarding the selection of exemplars to be used. Using a google form, the tutors attempted to evaluate, from 20 exemplar CVs that we were using, the ones which produced the greatest convergence in the judgements of the markers. As we have seen in the example of the "two images" case, it is not necessarily the rubric, or the artefact in itself that produces difficulty in marking, but rather, it is the relationship between the two, where none of the quality descriptors seems to quite encapsulate the reality of the artefact being evaluated. Accordingly,, a more detailed rubric was also written for the tutors to evaluate the exemplars with.

This exercise not only helped in the selection of exemplars, but also helped with the rewriting of the rubric into a final form for use with the students. In practice the criteria remained the same, but the attainment descriptors, rather than being given a simple yes/no/maybe, were given instead, much richer characterisations of the attainment level. In the table below there are the two rubrics used, with the changes highlighted compared to that which was used by the students in the evaluation exercise.

*Table 5-6: Evolution of Rubrics 2011-12. Tutor Exercise Rubric: Henceforth referred to as **2011TE** and the Tutor Marking Rubric: Henceforth referred to as **2011 TM**.*

TUTOR EXERCISE RUBRIC	**MARKING RUBRIC USED BY STUDENTS**
Has correct number of screens with correct headings on each • No headings/Way too few viewable screens • Many Incorrect headings/Some screens missing • Some wrong headings but all screens there • A few problems [visual/spelling] but mostly OK • All OK Correct background colour • Background colour completely wrong and bad to look at • Background colour completely wrong but OK to look at • Many screens wrong color but some have the right colour • Good attempt at most screens but a few not correct • All OK Correct width and height of the flash file • Totally wrong size • Right size but screen elements don't fit • Seems that screen elements weren't really designed for this size • A few minor issues with resizing • Everything OK Correct number of buttons with correct colours for them • No buttons • Some buttons but no real attempt to follow the brief in their design • Wrong number of buttons – or wrong colours – but tried to follow brief	Has correct number of screens with correct headings on each • No headings/Way too few viewable screens • Many Incorrect headings/Some screens missing • Insufficiently Prominent or Bad Spelling • A few problems [visual/spelling] but mostly OK • All OK Has Correct background colour • Background colour completely wrong and bad to look at • Background colour completely wrong but ok to look at • Background colour correct but occupies too small a percentage of the space • Good attempt at most screens but a few not correct • All OK Correct Width and height of the flash file (600x300) • Totally wrong Size • Right size, but screen elements don't fit • Seems that screen elements weren't designed for this size • A few minor issues with resizing • All OK Correct number of buttons with correct colours for them (purple with white text) • No buttons • Some buttons but no real attempt to follow the brief in their design • Wrong number of buttons – or wrong colours – but tried to follow brief

• Almost correct – just a few problems • All OK Make buttons navigate to correct frames using simple action script • No navigation • Lots of problems • Some buttons navigate well • Most buttons navigate well/minor issues • All OK Contains at least two images of you • No images • Poor image • Just one good image • Two images but some problems • All OK Small file-size • Vast file size (>x10) • File too large (x5) • Not bad (X2) • Just over (<x2) • OK Motion Tweening Of Position of Things in the animation welcome screen OR Motion Tweening of Visibility of Things (for fade ins and fade outs) in the animation welcome screen • No animation • Animation but very simple • Fair animation but not well performed • Quite good attempt showing good technique • Good animation with good Layout and positioning of buttons and content • No buttons / content • Poorly laid out buttons or other major problem with content • Buttons and/or content competently laid out but not visually attractive • Buttons and content quite well laid out but maybe lacking coherence • Very well laid out buttons and content	• Almost correct – just a few problems • All OK Buttons navigate to correct frames using simple action script • No navigation • Some wrong navigation • Navigate correctly but often with problematic transitions • Navigate correctly but sometimes with problematic transitions • All OK Contains at least two images of you • No images • Has a very poor image • Has only one good image • Two images but some problems • All OK Small file-size (Less than 200k) • Vast file size (>x10) • File too large (x5) • Not bad (X2) • Just over (<x2) • OK Motion Tweening Of Position/Visibility in the welcome screen • No animation • Animation but very simple • Evidence of technique but does not look good • Quite good attempt showing good technique • Good animation with good technique Layout and positioning of buttons and text • No buttons, or laid out so badly and inconsistently it is disorienting to navigate • Poorly laid out buttons /not appear altogether on same screen, or other major problem • Buttons and/or content competently laid out but not visually attractive • Buttons and content quite well laid out but maybe lacking coherence • Well laid out buttons and content

Well designed and appropriate content (text and image) • Poor content with low quality images • Images poor OR content poor • Content and images fair • Content OR images are GOOD • Content AND images GOOD	Choice of material, text and tone appropriate for a CV (text and image) • Poor content with low quality images • Images poor OR content poor • Content and images average • Content OR images are GOOD • Content AND images GOOD
Button clicks have small sounds associated with them • No sound • Sound almost imperceptible • Inappropriate sound • Most buttons have appropriate sounds • All buttons have appropriate sounds	Button clicks have small sounds associated with them • No sound • Sound almost imperceptible • Inappropriate sound • Most buttons have appropriate sounds • All buttons have appropriate sounds
Some very clever and visually elegant animation (using Shape Tweens, or Motion Guides or Masking) in the animation welcome screen • Poor or absent • Fair • Satisfactory • Good • Excellent	Has either very clever or visually elegant animation • Strongly disagree • Disagree • Not sure • Agree • Strongly agree
Buttons: Extremely well positioned, elegant, and suitable • Poor or absent • Fair • Satisfactory • Good • Excellent	Extremely well positioned and pleasant looking buttons • Strongly disagree • Disagree • Not sure • Agree • Strongly agree
Content: text relevant, appropriate length and evenly distributed across CV, images appropriate to CV, images of very high quality • Poor or absent • Fair • Satisfactory • Good • Excellent	Extremely well judged content The text is relevant, of appropriate length and evenly distributed across the CV, images are appropriate to CV and of a high quality • Strongly disagree • Disagree • Not sure • Agree • Strongly agree
An immediately visible and functioning background music toggle • No toggle	An immediately visible and functioning background music toggle • No toggle

• Toggle not immediately visible • Not on all screens/Not functionally perfect • On most screens but functionally perfect • On all screens and functionally perfect	• Toggle not immediately visible • Not on all screens/only plays per screen • Abrupt transitions between loops • On all screens and functionally perfect

While the rubrics used with students contained explicit attainment descriptors, the one used by the tutor when marking assignment 3, while having the same stem or statement, did not explicitly specify the attainment descriptors. This was because when giving a mark that constitutes a large part of the marks for the course as a whole some greater flexibility was needed. The redevelopment of the rubric, was partially completed in the rubric used for the tutor exercise and then redeveloped further in the new rubric for use students, thus showing an evolution motivated by different factors.

Let's look at some specific examples of changed criteria. The first shows a reduction in scope and disambiguation between the rubric used in the tutor exercise compared with that used by students.

Table 5-7: Appropriateness of Content Criterion 2011-12

Well designed and appropriate content (text and image)	vs	*Choice of material, text and tone appropriate for a CV (text and image)*

However, others demonstrate enrichment of criteria by the citation of typical cases of attainment evidenced in previous years, thus demonstrating that the rubric was becoming enriched by examples of its application.

Table 5-8: Screen Design Criteria 2011-12

Many screens wrong colour, but some have the right colour	vs	Background colour correct, but occupies too small a percentage of the space
Some wrong headings, but all screens there	vs	Insufficiently prominent or bad spelling

As can be seen here, just as was the case with the multimedia specification course, when one designs rubrics for students to be able to mark from, one necessarily tries to be comprehensible and to ensure the terms are in language they will understand. In fact, the language used is particularly colloquial in an attempt to give recognisable judgements that the students will be able to assign to any particular example of quality.

Summing up, the changes for the 2011-12 iteration were:

- reducing the percentage of the evaluation exercise to 10%;
- selecting a less divergent subset of exemplars from the pool of 40;
- rewriting the rubric and fleshing out the attainment descriptors in easily readable form;
- using clickers both for the rehearsal and the final evaluation exercises, such that all the work was done at once in the lecture theatre.

The result of this in the final assessment was a further improvement by 3% in the student average for the multimedia artefact assignment.

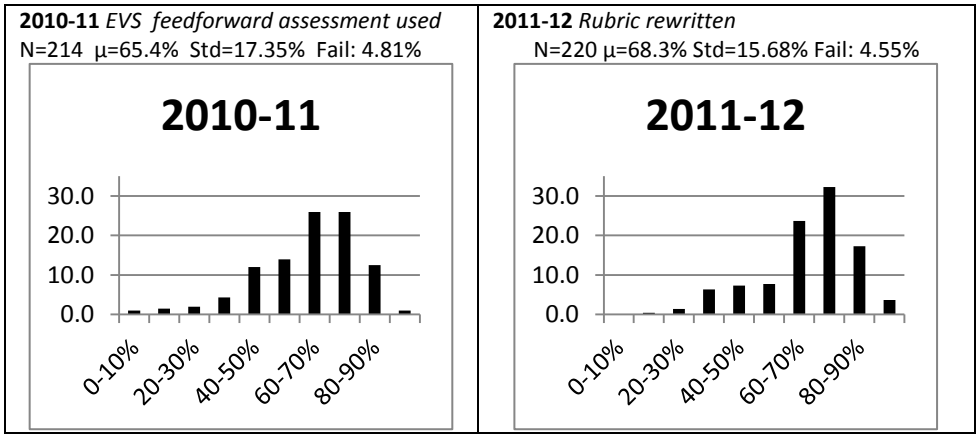


Figure 5-3: Normalised Graph of Student Outcomes in 10 Percentile Bins 2010-11/2011-12

As can be seen, the direction of travel is constant, continuing improvement, with more students getting into the 70-80% band and there is less variation. Moreover, the kinds of opposition encountered in the first year seemed to have gone away, although they returned again in the third iteration.

5.5 Third Iteration with Clickers and Exemplar Marking: 2012-13

During the third iteration of the course, the focus of the tutors became that of improving quality still further through the provision of a marking scheme with more comprehensive characterisations of attainment at the upper end. For example, the final two criteria on the list were:

- The CV is suitable for being viewed by someone who might give you a job;
- The CV's design expresses a kind of brand identity of you as a person.

As before, there was a rehearsal session and a final session, but because of technical issues during the former no data were collected. Moreover, there was some discussion among students about the two highest order criteria (quoted above), which produced a lot of divergence among the students' voting. In the final session, because we still wanted students to evaluate three different pieces of work, and time had represented a significant issue in the rehearsal event (together with a number of technical problems that day), a cut down rubric of only nine categories was used. This was a hastily rewritten rubric, edited according to the issues that came up in the rehearsal session. However, it constitutes something of an anomaly compared to other rubrics used (being much shorter) but nonetheless also demonstrates some thought in rewriting. However, these did not feed into the subsequent rubrics (for instance the one used by the tutor to mark the the final assignment during this iteration) which returned to the highly dimensioned style seen in all the others.

Table 5-9: 2012 Rubric Used by Students - henceforth referred to as 2012S

1 Has all required screens (welcome, details, hobbies, employment, education) and they are always accessible by button navigation

- Does not have all screens
- Has all screens but not all always accessible
- All screens but very large variations in location of heading
- Some variations in location of heading
- Each screen clearly headed by correct word(s) at the expected location

2 Sufficient contrast in colours

- Some pages unreadable because of colours
- Some illegibility because of colour problems
- Poor contrast in colours or large variations in contrast
- Mostly clear contrast in colours
- Clear contrast in colours

3 Buttons are readable and have sounds

- Buttons are unreadable and have no sound
- Buttons have sound but lack readability
- Buttons are readable but none have sound
- Buttons are all readable but some may lack sound
- All buttons are readable and all have appropriate sounds

4 Layout is harmonious, regular and consistent

- Very inharmonious or random screens
- Some big disharmony or inconsistency between screens
- Small positional twitches between screens
- Some minor disharmonies on screens
- All screens harmonious and balanced

5 Good grammar and spelling and use of language

- Many glaring spelling or grammar errors
- A glaring spelling or grammar error
- Minor spelling errors
- Minor grammar errors
- No spelling/grammar errors

6 Animation demonstrates meaning and visual appeal

- Has no animation
- Has only the most basic animation which lacks meaning
- Displays some creativity but lacks meaning
- Is meaningful but lacks visual appeal
- Is both meaningful and visually appealing

7 Aligned and uniform sized buttons

- Extremely poor buttons
- Chaotic arrangements of buttons across and within screens
- Not all buttons visible on all screens or alignment issues
- Some small alignment, spacing or text size issues
- Good alignment, spacing and text size on all screens

8 Text of appropriate length and evenly distributed across CV (ignore questions of suitability of text here)

- Clearly insufficient text
- Only the bare minimum of text
- Sufficient but unevenly distributed
- Mostly evenly distributed
- All text evenly distributed of appropriate length

9 A functioning widget for turning music on and off independent of navigational functionality

- No music widget
- Music widget has obvious and immediately apparent flaws
- Has flaws but not immediately apparent ones
- On most screens but functionally fit for purpose
- On all screens and functionally fit for purpose

For the purpose of completeness, below is a comparison of the rubrics used to mark the final assignment between the rubrics used by the tutor to mark the 2012/13 iteration compared with that used for the 2011/12 iteration.

Table 5-10: Evolution of Rubric 2012-13. Rubric used for Tutor Marking: Henceforth referred to as 2012TM

TUTOR CRITERIA 2011-12	TUTOR CRITERIA 2012-13
BASICS (40 – 49 MARKS - ATTAIN SATISFACTORY ACHIEVEMENT) 1. Publish an SWF file and upload it to Studynet (5) 2. Has correct number of pages with correct headings on each (5) 3. Correct background colour (5) 4. Correct width and height of the Flash file (5) 5. Correct number of buttons with correct colours for them (5) 6. Make buttons navigate to correct frames using simple action script (5) 7. Contains at least two images of you (5) 8. Small file-size (5) 9. Motion Tweening Of Position of Things in the animation welcome screen OR Motion 10. Tweening of Visibility of Things (for fade ins and fade outs) in the animation welcome screen (5) 11. Correct and nice positioning of buttons and content (5) **INTERMEDIATE (50 - 69 MARKS - ATTAIN GOOD TO VERY GOOD ACHIEVEMENT)** 12. Good easy on the eye content (text and image) - not too little, not to much and all relevant(10) 13. Button clicks have small sounds associated with them (10) **ADVANCED (70 – 100 TO ATTAIN EXCELLENT TO OUTSTANDING ACHIEVEMENT)** 14. Some very clever and visually elegant animation (using Shape Tweens, or 15. Motion Guides or Masking) in the animation welcome screen (10) 16. Extremely well positioned and pleasant looking buttons (5) 17. Extremely well judged content (5) 18. An immediately visible and functioning background music toggle (10)	**SATISFACTORY WORK 40% - 49%** 1. Publish an SWF file of under 150k and upload it to Studynet (4) 2. Has correct number of screens with correct headings on each (4) 3. Appropriate choice of screen colour – providing good contrast (4) 4. Correct width and height of the Flash file (4) 5. Correct number of buttons with good colour selection (4) 6. All buttons navigate to the correct frame script (4) 7. Contains at least two images of you (4) 8. Good spelling and use of language (4) 9. An animation in the welcome screen (4) 10. Aligned and Uniform sized Buttons (4) 11. Text content is relevant and expressive and compact (5) 12. Buttons have appropriate sounds on click event (5) **GOOD WORK 50% - 59%** 13. Images of self show high production values (5) 14. Text and image presented well on the screen (5) **VERY GOOD WORK 60% - 69%** 15. Animation demonstrates originality and visual appeal (10) **EXCELLENT WORK 70% - 100%** 16. Background music is appropriate and is controllable by user (10) 17. The CV is suitable for being viewed by someone who might give you a job (10) 18. The CV's design expresses a kind of brand identity of you as a person (10)

In this evolution of the criteria (for the tutor), three major strategies can be seen:

1. Reorganisation and aggregation (putting separate criteria into one new unified criterion). For instance, the 2011-12 criteria 1 and 8 were merged into criterion 1 in 2012-13;
2. More generic characterisation of attainment level (broad characterisations like “disharmony” and “appropriate”);
3. Higher order criteria (17 & 18).

However, whilst there was a move towards abstraction in the tutor criteria (essentially to encompass more cases of attainment without detailing them schematically), the criteria used for the student exercise (although far fewer) had more detailed attainment descriptors that those used in the previous year. Certainly, this was necessary to help the students concretize examples of a particular attainment descriptor, but having done so, a more abstract representation could be used for the tutor rubric.

At the end of this course, the quality had further increased.

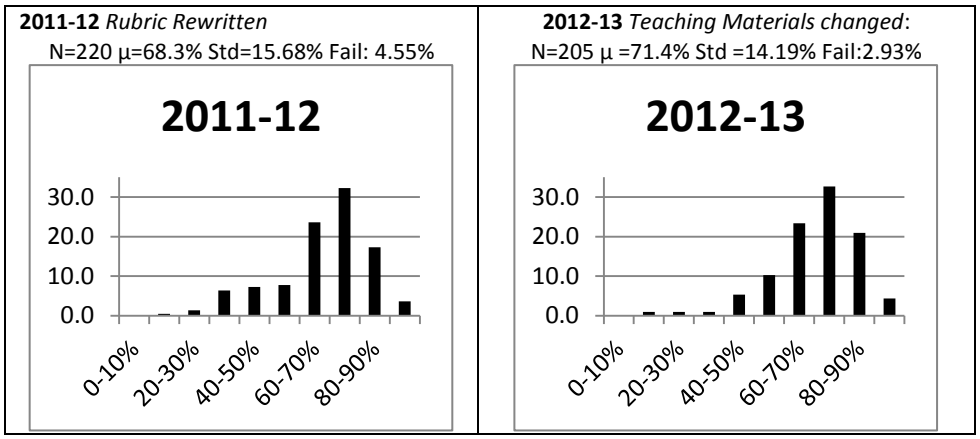


Figure 5-4: Normalized Graph of Student Outcomes in 10Percentile Bins 2011-12/2012-13

This year was a mixed one in terms of outcomes, with the technical problems in the rehearsal and the truncated rubric used in the summative feed-forward evaluation exercise meaning that the rubric the students marked to was different to that used by the tutor, however, the improvement nonetheless continued and moreover, the number of students failing (falling below 35%) had become vanishingly low.

5.6 Fourth Iteration with Clickers and Exemplar Marking: 2013-14

The two major changes for the final iteration were:

1. Rethinking of the evaluation exercise;
2. Improvement of the training set.

Perhaps the major weakness of in the prior iterations of the course had been the quality of the training sample. The initial decision to anonymize the CVs by replacing personal details and images with fictional detail and cartoon images (of the Simpsons), meant that the exemplars were not necessarily holistically unified and whilst they formed, as has been seen, a useful benchmark to separate artefacts of high quality from others of lower quality, they did not necessarily represent something to aspire to. As a consequence, we decided to focus on six exemplars specifically chosen because they represented different standards of quality. Below are the anonymized versions with the score they obtained in their original format.

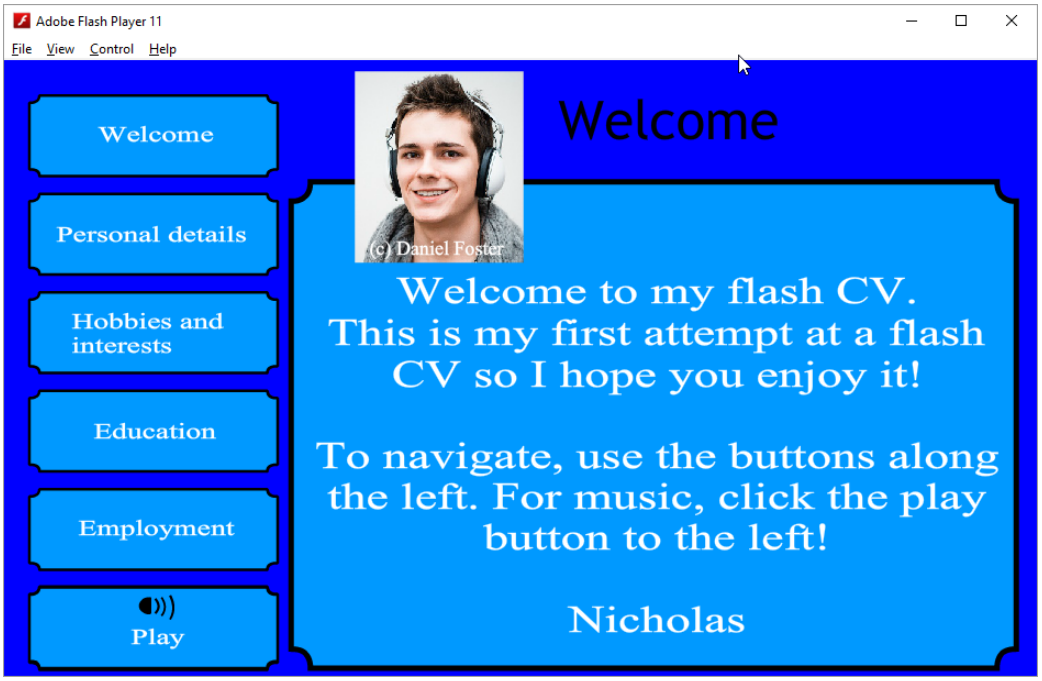


Figure 5-5:CV1.swf – Scored 80.9%

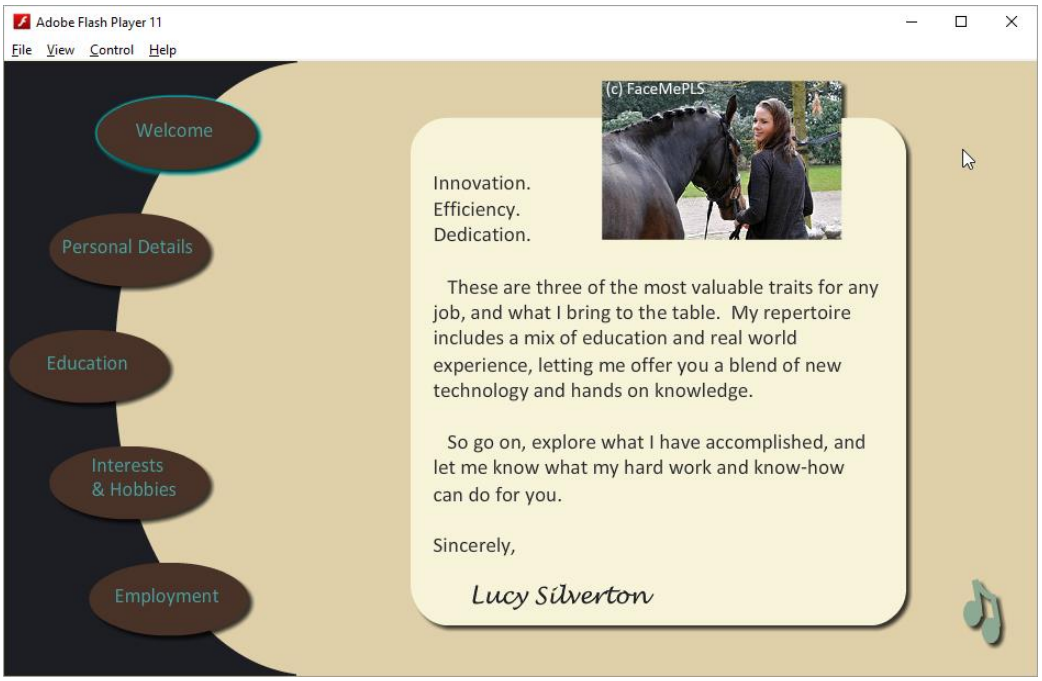


Figure 5-6:Cv2.swf = Scored 93%

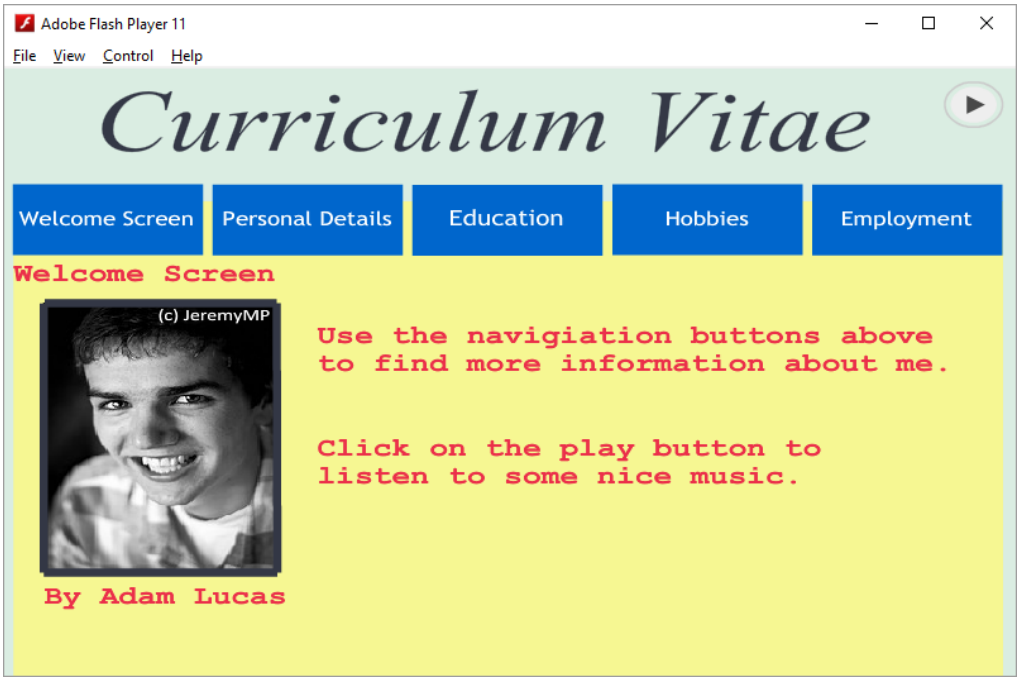


Figure 5-7:Cv3.swf = 88%

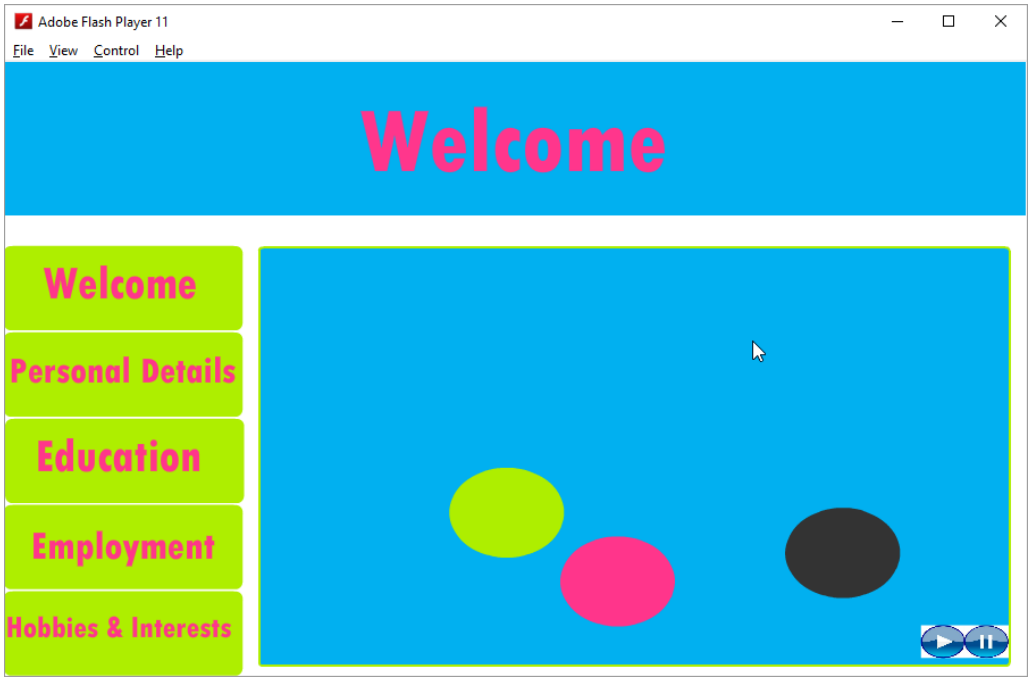


Figure 5-8:Cv4.swf = 69%

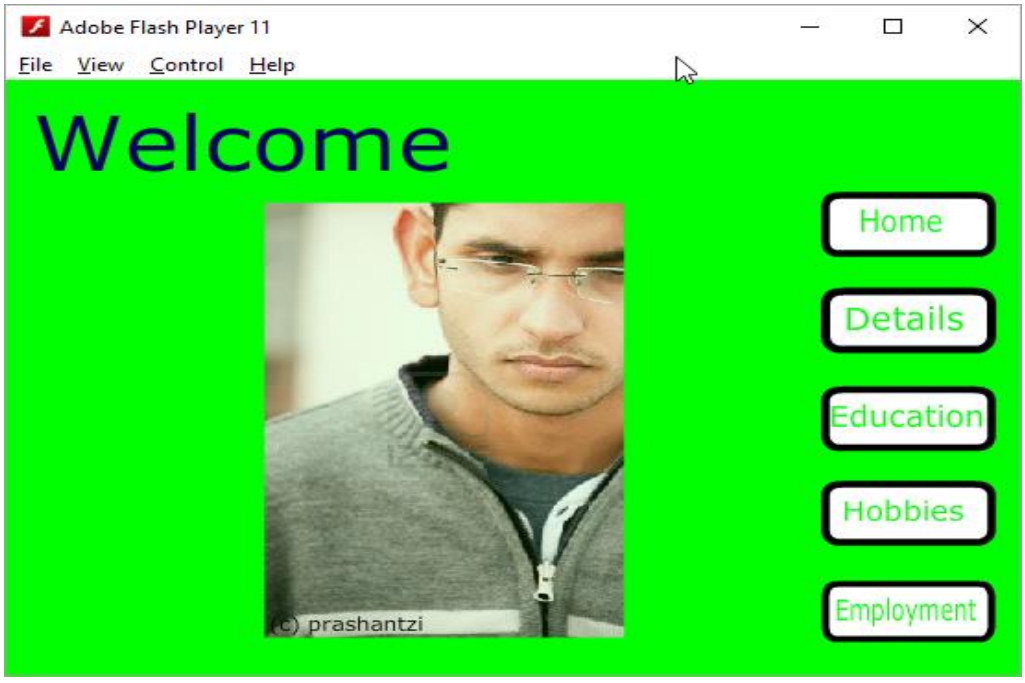


Figure 5-9:Cv5.swf = 55%

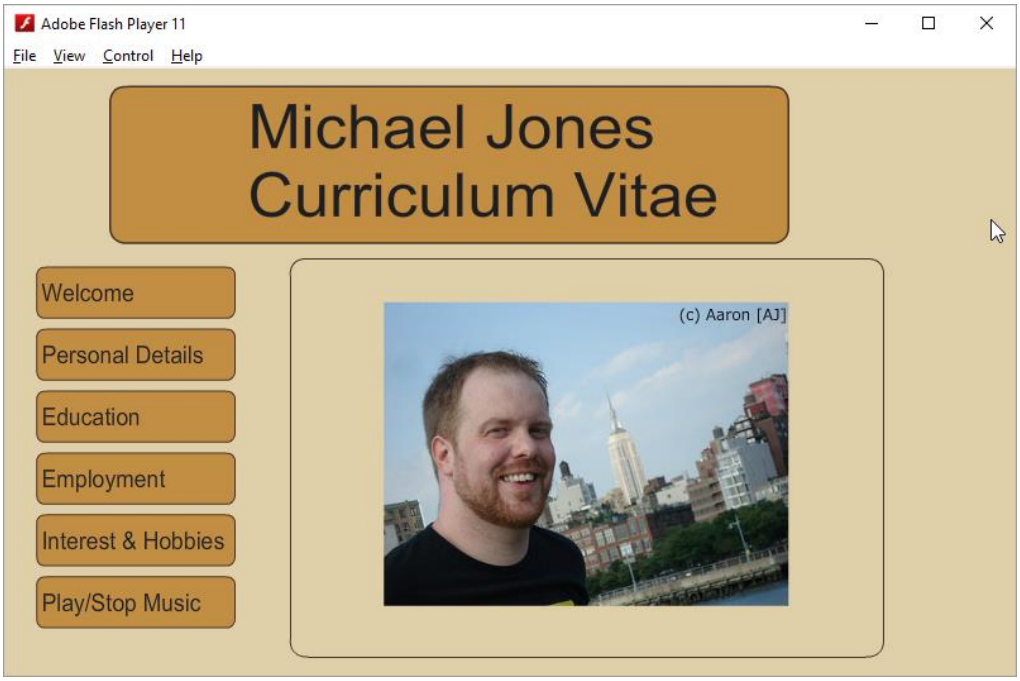


Figure 5-10:Cv6.swf = 82%

The second change was to deal, finally, with the issue of subjectivism in the evaluation assignment. The problem up then had been that we had given the students marks regarding the extent that they agreed with us. While this appeared in some ways unfair, at least it did ensure that the whole cohort participated in the evaluation sessions. In the final iteration, we sought a method that would ensure full cohort coverage, yet not be grade bearing for students. Accordingly, we developed a hybrid assessment, where we looked at two previous CVs and marked them according to the criteria (and no numeric value was awarded for the pattern of scoring made by the student), but at the end of that session, we asked a number of objective test questions about the CVs being evaluated (how was any particular effect achieved), which were grade bearing. The questions in the evaluation session thus no longer needed to be so detailed (since student scores would not depend on them) and were in fact identical to those used by the tutor in the final artefact marking; the only difference being explicit attainment descriptors when used in the evaluation session. In the subsequent chapter, I will go into more detail, but for now, here are the stems of those questions and these were applied to two different CVs

Table 5-11: 2013-14 Student Rubric (exactly the same stems as 2012TM) – referred to henceforth as 2013S

1. Publish an SWF file of under 250k and upload it to Studynet
<250k
250k-300k
300k-350k
>350k

2. Has correct number of screens with correct headings on each
All screens correct headings
Not exact but meaningful headings
Odd Headings

3. Appropriate choice of screen colour – providing good contrast
Strongly agree
Agree
Neutral
Disagree
Strongly disagree

4. Correct width and height of the Flash file
Yes
Proportions the Same, but not size
No

5. Correct number of buttons with good colour selection
Strongly agree
Agree
Neutral
Disagree
Strongly disagree

6. All buttons navigate correctly
Yes
Mostly
No

7. Contains at least two images of you
Yes
Yes, but poor quality
Only 1 Image
No Images

8. Good spelling and use of language
Perfect
Some imperfections
Many imperfections

9. An animation in the welcome screen
Yes
No

10. Aligned and Uniform sized Buttons
Yes
Slightly imperfect
Very imperfect

11. Text content is relevant and expressive and compact
Strongly agree
Agree
Neutral
Disagree
Strongly disagree

12. Buttons have appropriate sounds on click event
Yes
Have sounds, but not appropriate
Some missing sounds
No sounds

13. Images of self show high production values
Strongly agree
Agree
Neutral
Disagree
Strongly disagree

14. Text and image presented well on the screen
Strongly agree
Agree
Neutral
Disagree
Strongly disagree

15. Animation demonstrates originality and visual appeal
Strongly agree
Agree
Neutral
Disagree
Strongly disagree

16. Background music is appropriate and is controllable by user
Strongly agree
Agree
Neutral
Disagree

Strongly disagree

17. The CV is suitable for being viewed by someone who might give you a job
Strongly agree
Agree
Neutral
Disagree
Strongly Disagree

18. The CV's design expresses a kind of brand identity of you as a person
Strongly agree
Agree
Neutral
Disagree
Strongly disagree

Following this, there was a set of multiple choice questions about those CVs and about the platform (Adobe Flash), in general.

Table 5-12: Post Evaluation Objective Test Questions

1. Which objects do the classic tweens control the appearance of?
2. How is the animation of the banner achieved?
3. On which frame does the initial animation stop?
4. How were the corners of the buttons most likely created?
5. The animation here tweens?
6. Which statement is true, regarding fonts?
7. The animated rectangle is on which layer?
8. Audio will play from which frame of which layer in this symbol?
9. The Banner Animation Tweens what property?

The answers to these objective questions became the score for each student's participation in this assignment and the preceding evaluative questions were considered purely formative. This had the effect of eliminating any controversy about the lecturers' opinions, and on whether being marked in terms of proximity to them was justifiable, whilst also ensuring full-cohort coverage in the evaluation exercise. In the practice session for this assignment the same format was followed, i.e. two formative evaluations of CVs followed by five objective test questions. The big difference, however, was that 102 students attended the purely formative session, whilst 181 attended the formative and summative session. The comparison between this year and the previous year is as follows.

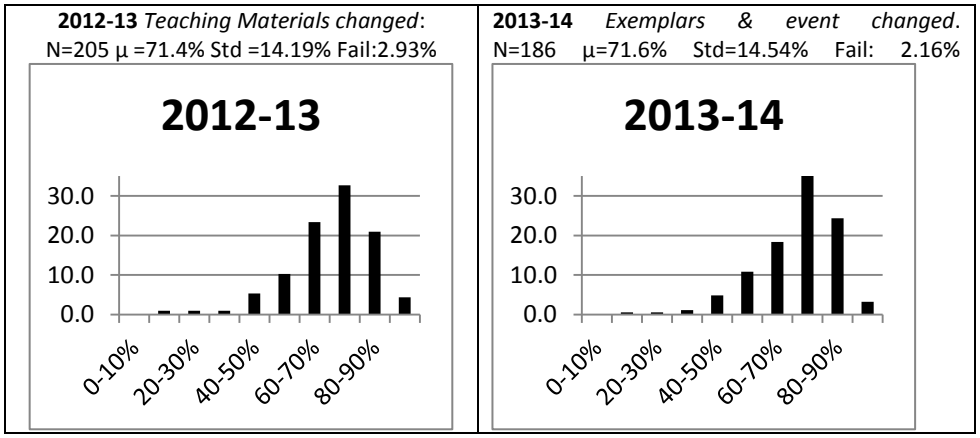


Figure 5-11: Normalised Graph of Student Outcomes in 10 Percentile Bins 2013/2014

Here the difference is tiny, which leads one to believe, whilst the final iteration of the course represented the ideal form of the course, probably most of the beneficial effects had been established already. Despite the exemplars with the Simpsons' characters in them having probably been less than polished, they nonetheless offered enough contrast between themselves to be the basis for establishing an intuitive sense of quality among students, which then impacted on their own practice.

Summing up, over the course of the module the following evolution was observed.

Table 5-13: Summation of Evolution of Module by Year

Year	Summary	Rubrics
2010-11	Introduction of the use of exemplars. Clickers used for one training session in class. Only 80 clickers, so students shared them. The evaluation exercise was performed instead using Questionmark Perception (online evaluation) in computer labs. Overall average score in final assignment improved from 58% to 65%. The grade value of this exercise was 25%, based on how near the students' marking was to the tutors. Some negativity expressed about the subjectivity of the exercise.	2010TM 2010S
2011-12	Full distribution of clickers to all the cohort. The evaluation exercise was performed in two sessions, formatively and summatively. Effort was made to establish a set of exemplars resulting in the least divergence among tutor markers. Average student scores went up a further 3% to 68%. The rubric was improved by adding more options beyond yes/no/maybe. The grade value of the evaluation exercise was reduced to 10%.	2011TE 2011TM 2011S
2012-13	Attempt to establish richer marking criteria to produce higher standards. This led to complications during the evaluation exercise in that there were problems during the rehearsal exercises. In the summative evaluation exercise these were not used, and a temporary much reduced criteria set was deployed instead. There was further improvement in the overall scores by 3%.	2012S 2012TM

2013-14	A more polished and credible set of exemplars was produced with images of real people and not cartoons in them. The evaluation exercise was made hybrid formative/summative, with the summative element coming from objective questions.	2013S 2012TM

Over the whole of the course, a gradual refinement of the criteria used in the rubric has been evidenced and this has been characterized by a case by case process of modifying the attainment descriptors, making more abstract or more detailed attainment descriptors and inserting higher order attainment descriptors, such as *making the CV suitable to someone who might offer you work.* However, the fact that the direction in improvement continued over the four iterations perhaps suggests that it is not the precise actual wording in the rubric which is instrumental in the improvements in student attainment, but rather the concentration on the concept of *quality* itself, for which the rubric represents a more schematic proxy

Next, the statistics arising from the use of these rubrics over four iterations of a very large course are presented and analysed. Owing to the dynamic evolution of the course over time the data are not as longitudinally uniform as was the case for the MSc course. Moreover, the fact that a number of the criteria were typically binary in outcome (yes/no/maybe with very few maybes), this means the kind of analyses carried out before based on correlation of marking patterns between tutors and students is a less illuminating measure. However, the artefacts students were producing were individual pieces of work, and the cohort in which they did so was very large, thus making it easier to examine, numerically, improvements in attainment.

Chapter 6. E-Media Design Course – The Map of Improvement

As has been seen in the previous chapter, over six years of the course, the last four with EVS exemplar evaluation being used, a remarkable increase in scores for the final assignment among students was witnessed.

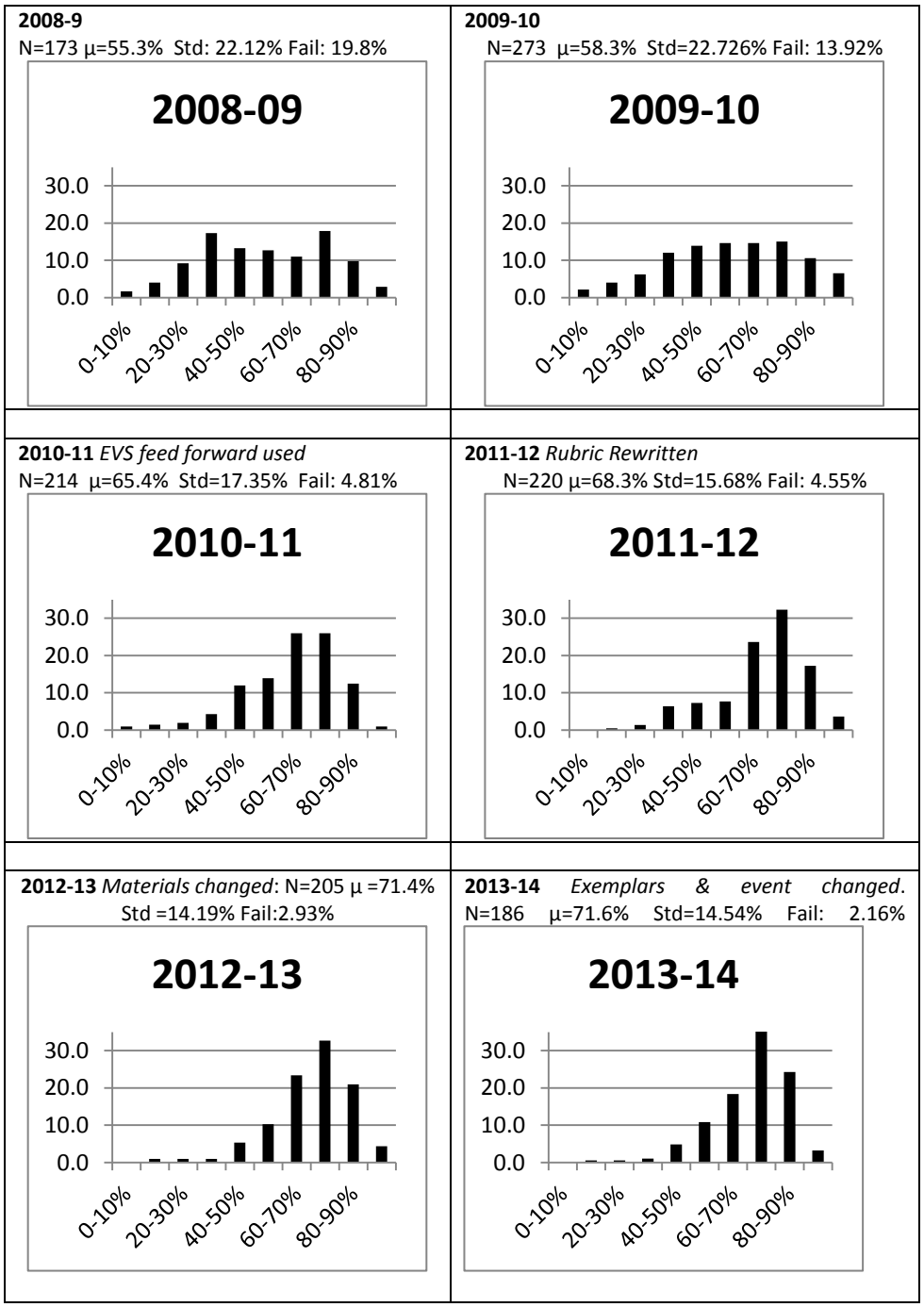


Figure 6-1: Marks Distribution by Year

T-tests for each pair of successive years from 09-10 and 10-11 were carried out and following was found.

Table 6-1: T-Tests for Successive Year Averages

	09-10	10-11	10-11	11-12	11-12	12-13	12-13	13-4
Mean	58.32	65.42	65.42	68.32	68.32	71.29	71.29	72.13
P(T<=t) one-tail	0.0001		0.03		0.02		0.27	

This demonstrates that the result (indicating an improvement on a year-by-year basis) is significant at the $p<0.01$ level for the first year, and at the $p<0.05$ level for the two succeeding years, whilst the improvement in the final year is not significant. It can be said with a fair degree of confidence that this was not a cohort effect, since student scores in the first multiple choice test assignment remained very static over those six years

Table 6-2: Averages of Assignment 1 and Assignment 3 over time

	Without LPA		With Live Peer Assessment			
	2008	2009	2010	2011	2012	2013
First Test	55.67	58.89	56.41	54.16	**N/A***	52.28
Final Artefact	55.3	58.3	65.4	68.3	71.4	71.6

*The first test in 2012 was, of necessity, in a different format from previous studies, so cannot be included in this research.

However, the dramatic effect it has had on student attainment in the Flash artefact can be seen in the normalised graph of student attainment below.

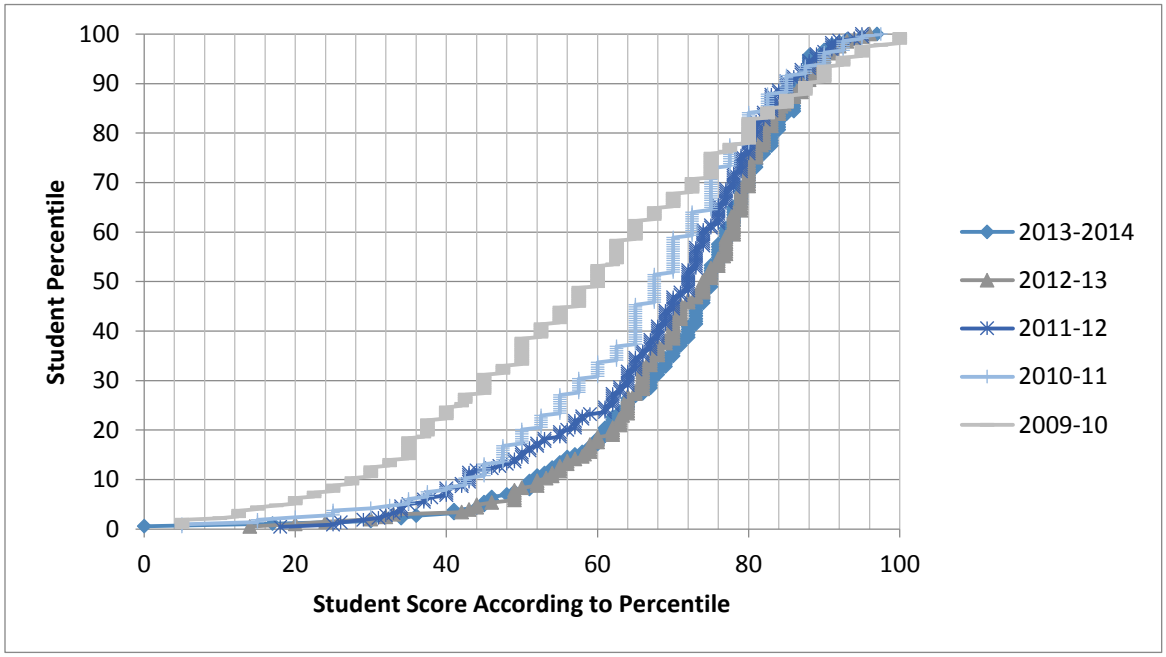


Figure 6-2: Marks by Percentile Over Years

In simple terms, in 2009-10, the lowest twenty percent of students were scoring just under 37.5% for the final assignment, however, four years later, it was 62.5%.

In detailed terms the following is found.

Table 6-3:Scores by Iteration and Decile

	Academic Years				
Percentile	09-10	10-11	11-12	12-13	13-14
10	28	43.25	43	53.52	52
20	37.5	51.5	56.8	62	61
30	45	57.5	64	66	67
40	52.5	65	68	70	72
50	60	67.5	72	74	75
60	65	72.5	74	78	77
70	72.5	75	78	80	80
80	80	80	81	82	83
90	87.5	85	85	87	87
100	100	97.5	95	96	97

For the top 20% of students, the various effects described here do not make much difference. However, for the lowest 40% of students there is a dramatic difference of approximately 20% in each case. Moreover, whereas the improvements caused by the interventions in the penultimate iteration are small in most categories, they continue to be significant for the lowest 10% of students.

A more dramatic visualisations of the changes in the marks distribution can be seen in a box and whisker diagram and a violin plot below. The former displays a box containing the first and third quartiles of the data, whilst the latter visualises the probability density of the data at different values. In this we see that the really significant change occurs during the very first iteration of the EVS technique. This took place in a highly experimental way, at a time when there was not full provision of clickers to the whole cohort (causing the students to have to work together in groups to participate in the marking event), and when the summative marking event was undertaken using Questionmark Perception. Even in this context, the median mark went up by 14%. In the violin plots, a blue line has been added, representing 35%, i.e. the pass/fail boundary for first year undergraduate students at the time. It can thus be seen how the probability of scoring in the fail range becomes vanishingly small after the second iteration of the course using these techniques.

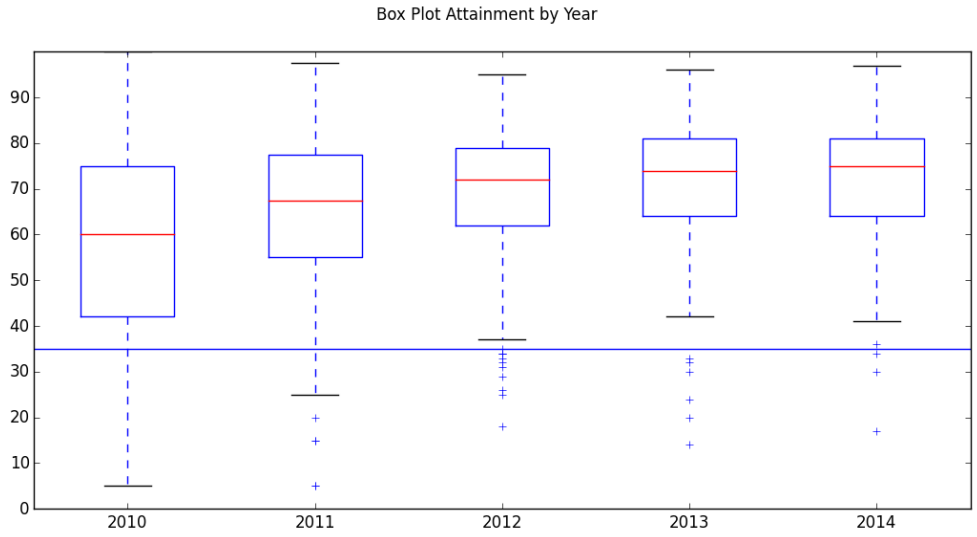


Figure 6-3: Box Plot of Performance in Final Assignment Over Years

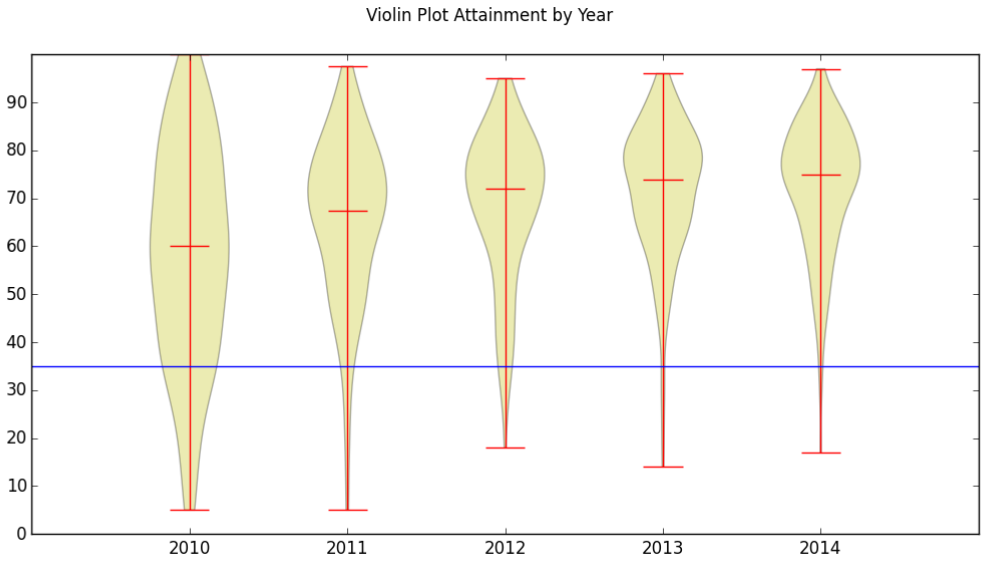


Figure 6-4: Violin Plot of Performance over Years

As can be observed, dramatic changes take place in the first two iterations using live feed forward techniques (2011/12), whereby just by adding the collective evaluation sessions, the median has dramatically improved. The fact that it has had this effect on students in the lower quartiles may be due to the compulsory attendance at the evaluation events. Over the years, attendance at "voluntary" formative evaluation events was usually a third down on attendance where the exercise was summatively graded. As previously mentioned, unfortunately, there were technical problems during the

2012-13 rehearsal event, so there is an absence of data for that, but for the other three years the data are comparable.

Table 6-4:Attendance by Event

Year	Rehearsal	Final	% of students attending rehearsal
2010-11	118	203	58.13%
2011-12	138	214	64.49%
2012-13	No data	195	
2013-14	102	181	56.35%

This indicates that without some summative component, participation goes down by at least a third. At this point, the scores for the final assignment as marked by a tutor, are considered in more detail. That is to say, the marks awarded by that tutor when assessing the students' final assignment is the focus and this not related to the way students voted when applying the rubrics in their own work.

6.1 The Marking of the Final Assignment

Whilst a number of different rubrics were used and experimented with in the student sessions, in terms of the marking of the final artefacts by the tutor, there were only two in operation (although in the final year [2011-12] of the first rubric, the marker was able to respond with a continuous scale, rather than the 100%/50%/0% option for each criterion, which was in operation during the 2009/10 and the 2010/11 iterations).

Table 6-5: Rubrics 09/10/11 vs 12/13

2009/10/11 Rubric (2009/10 only yes/no/maybe responses) (2011 continuous scale responses)	2012/2013 Marking Rubric
1. Publish an SWF file and upload it to Studynet (5) 2. Has correct number of pages with correct headings on each (5) 3. Correct background colour (5) 4. Correct width and height of the Flash file (5) 5. Correct number of buttons with correct colours for them (5) 6. Make buttons navigate to correct frames using simple action script (5) 7. Contains at least two images of you (5) 8. Small file-size (5) 9. Motion Tweening Of Position of Things in the animation welcome screen OR Motion	1. Publish an SWF file of under 150k and upload it to Studynet (4) 2. Has correct number of screens with correct headings on each (4) 3. Appropriate choice of screen colour – providing good contrast (4) 4. Correct width and height of the Flash file (4) 5. Correct number of buttons with good colour selection (4) 6. All buttons navigate to the correct frame script (4) 7. Contains at least two images of you (4) 8. Good spelling and use of language (4)

Tweening of Visibility of Things (for fade ins and fade outs) in the animation welcome screen (5) 10. Correct and nice positioning of buttons and content (5) 11. Good easy on the eye content (text and image) not too little not too much and all relevant (10) 12. Button clicks have small sounds associated with them (10) 13. Some very clever and visually elegant animation (using Shape Tweens or Motion Guides or Masking) in the animation welcome screen (10) 14. Extremely well positioned and pleasant looking buttons (5) 15. Extremely well judged content (5) 16. An immediately visible and functioning background music toggle (10)	9. An animation in the welcome screen (4) 10. Aligned and uniform sized buttons (4) 11. Text content is relevant and expressive and compact (5) 12. Buttons have appropriate sounds on click event (5) 13. Images of self show high production values (5) 14. Text and image presented well on the screen (5) 15. Animation demonstrates originality and visual appeal (10) 16. Background music is appropriate and is controllable by user (10) 17. The CV is suitable for being viewed by someone who might give you a job (10) 18. The CV's design expresses a kind of brand identity of you as a person (10)

Of the iterations of the course, probably the most illuminating comparison we could make is of the 2009-10 and 2010-11 years since exactly the same rubric was used and in each case each criterion had only attainment levels: 100%, 50% or 0%. In the table below, the average score per criterion is presented.

Table 6-6: 2009-10 and 2010-11 Average Score Per Criterion for the CV Assignment (scores awarded by tutor)

Criterion	Ave 2009	Ave 2010	Difference	Difference as Percentage
Q1 Publish an SWF file and upload it to Studynet (5)	4.64	4.84	0.20	4.00%
Q2 Has correct number of pages with correct headings on each (5)	4.38	4.72	0.34	6.80%
Q3 Correct background colour (5)	4.50	4.48	-0.02	-0.40%
Q4 Correct width and height of the Flash file (5)	3.70	4.24	0.54	10.80%
Q5 Correct number of buttons with correct colours for them (5)	3.61	4.52	0.91	18.20%
Q6 Make buttons navigate to correct frames using simple action script (5)	3.69	4.52	0.83	16.60%
Q7 Contains at least two images of you (5)	4.41	4.40	-0.01	-0.20%
Q8 Small file-size (5)	4.14	4.38	0.24	4.80%
Q9 Motion Tweening Of Position of Things in the animation welcome screen OR Motion Tweening of Visibility of Things (for fade ins and fade outs) in the animation welcome screen (5)	4.04	4.39	0.35	7.00%

Q10 Correct and nice Positioning of buttons and content (5)	2.61	3.25	0.64	12.80%
Q11 Good easy on the eye content (text and image) not too little not too much and all relevant (10)	4.47	5.74	1.26	25.20%
Q12 Button clicks have small sounds associated with them (10)	5.65	7.28	1.63	16.30%
Q13 Some very clever and visually elegant animation (using Shape Tweens or Motion Guides or Masking) in the animation welcome screen (10)	1.97	1.82	-0.15	-1.50%
Q14 Extremely well positioned and pleasant looking buttons (5)	0.87	0.36	-0.51	-10.20%
Q15 Extremely well judged content (5)	1.06	0.33	-0.73	-14.60%
Q16 An immediately visible and functioning background music toggle (10)	3.95	5.25	1.30	13.00%

What is absolutely fascinating here is the fact that it is in the lower order criteria and typically, the technical criteria where improvement is most clearly registered. And in fact, in a number of the higher-order criteria, the effect is, paradoxically, to reduce the score.

The seven criteria contributing the most to the increase in the students' average score are:

- Q11 Good easy on the eye content (text and image) not too little not too much and all relevant (10);
- Q5 Correct number of buttons with correct colours for them (5);
- Q6 Make buttons navigate to correct frames using simple action script (5);
- Q12 Button clicks have small sounds associated with them (10);
- Q16 An immediately visible and functioning background music toggle (10);
- Q10 Correct and nice positioning of buttons and content (5);
- Q4 Correct width and height of the Flash file (5).

While the background music toggle is technically the most difficult it is however, a technical exercise and one not concerned with overall holistic quality. The only criterion among these to do that is that which did register the highest overall increase: "good easy on the eye content". The other criteria to register large improvements were in fact the basic ones: button sounds, number of buttons and correctly navigating buttons. That is, much of the improvement came from attending to fairly basic criteria – though the fact that "good easy on the eye content", a very global and subjective judgement, achieved

the highest overall increase, indicates that potentially there was a level of seriousness being adopted by the students'

Interestingly, the two criteria that were developed to reward highest levels of achievement:

- Q14 Extremely well positioned and pleasant looking buttons (5);
- Q15 Extremely well judged content (5);

both actually went down in terms of average score, which underlines the fact that the improvement that took place over the first two years was primarily the weaker students catching up – with very little effect on the highest percentiles of students.

If the percentages of Yes, Maybe and No for each criterion in the graphs below are considered in more detail (where blue represents the 2009 cohort and the orange the 2010 cohort) the number of Nos is almost always less in 2010 than it is in 2009. Below are the proportions of Nos, Maybes and Yeses in years 09-10 and 10-11. In the lower order criteria the biggest change is the reduction in number of Nos

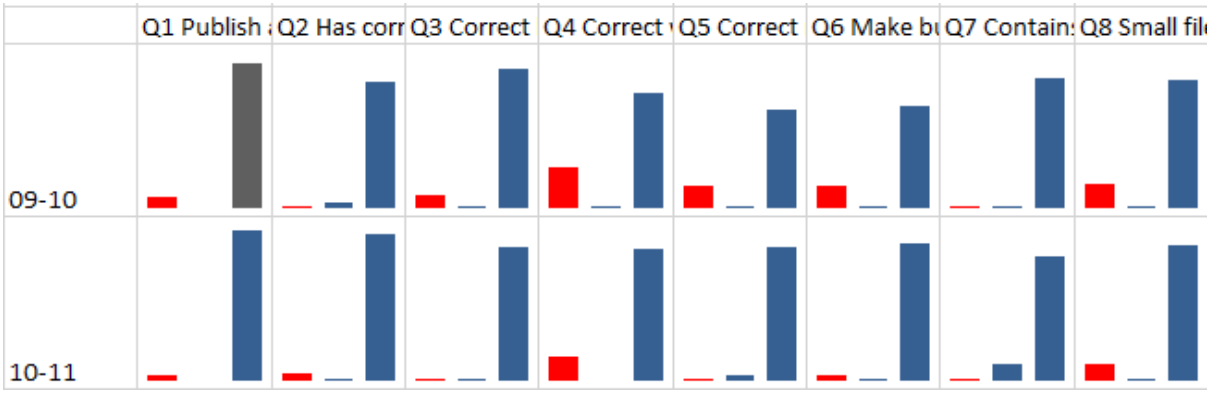


Figure 6-5:Proportion of No/Maybe/Yes in First Eight Criteria in Years 09-10 and 10-11

This pattern is also true in the later (more higher order) questions, except for the aforementioned criteria relating to higher order achievement in terms of buttons and content.

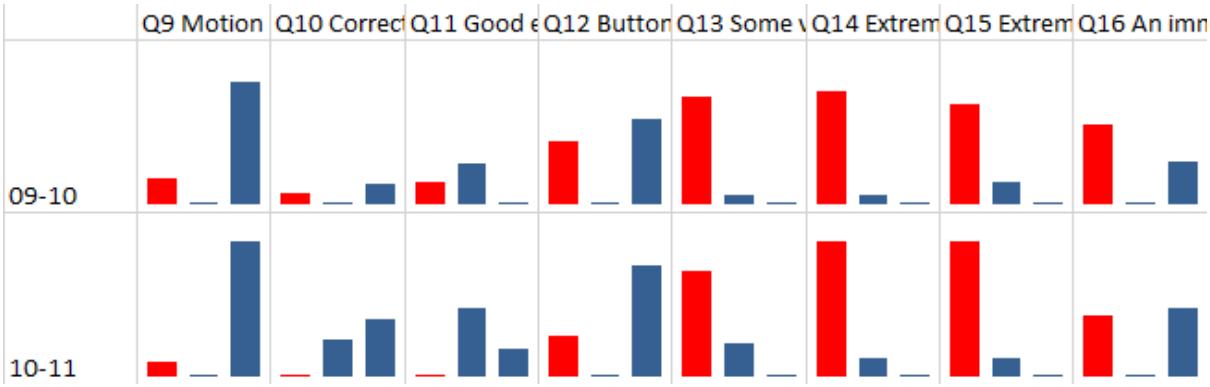


Figure 6-6:Proportion of No/Maybe/Yes in Final Eight Criteria in Years 09-10 and 10-11

Clearly, whilst the intention was to help the students develop higher order skills, in practice, what the exemplar marking seems to have done is to make them concentrate on the basics, i.e. to cross the is and dot the ts; to not lose marks unnecessarily by failing to comply with the instructions in the assignment. Moreover, it seems to have contributed to a sense of seriousness, with the students avoiding clearly inappropriate content, which had been visible in previous years. For the sake of completeness, here is the comparison between the marks in 2010 and 2011. However, the latter uses a continuous scale in each marking category and therefore, variations in scores between criteria might be down to a more nuanced way of compensating less successful work.

Table 6-7: 2010-11 and 2011-12 Average Score Per Criterion in CV Assignment (scores awarded by the tutor)

Criterion	2010 Average	2011 Average	Difference	Difference as Percentage
Q1 Publish an SWF file and upload it to Studynet (5)	4.84	4.73	-0.11	-2.2%
Q2 Has correct number of pages with correct headings on each (5)	4.72	4.78	0.05	1%
Q3 Correct background colour (5)	4.48	4.48	-0.01	-0.2%
Q4 Correct width and height of the Flash file (5)	4.24	4.23	-0.01	-0.2%
Q5 Correct number of buttons with correct colours for them (5)	4.52	4.57	0.06	1.2%
Q6 Make buttons navigate to correct frames using simple action script (5)	4.52	4.60	0.08	1.6%
Q7 Contains at least two images of you (5)	4.40	4.76	0.35	7%
Q8 Small file-size (5)	4.38	3.84	-0.54	-10.8%
Q9 Motion Tweening Of Position of Things in the animation welcome screen OR Motion Tweening of Visibility of Things (for fade ins and fade outs) in the animation welcome screen (5)	4.39	4.59	0.20	4%

Q10 Correct and nice positioning of buttons and content (5)	3.25	3.89	0.64	12.8%
Q11 Good easy on the eye content (text and image) not too little not to much and all relevant (10)	5.74	5.82	0.09	0.9%
Q12 Button clicks have small sounds associated with them (10)	7.28	7.16	-0.12	-1.2%
Q13 Some very clever and visually elegant animation (using Shape Tweens or Motion Guides or Masking) in the animation welcome screen (10)	1.82	2.36	0.54	5.4%
Q14 Extremely well positioned and pleasant looking buttons (5)	0.36	1.19	0.83	16.6%
Q15 Extremely well judged content (5)	0.33	1.12	0.78	15.6%
Q16 An immediately visible and functioning background music toggle (10)	5.25	5.88	0.63	6.3%

In many of the categories, there is not much difference in marks, however this iteration seems to have recovered slightly regarding the high order criteria (Q13/Q14/Q15), which witnessed the fall in 2010. However, as mentioned above, because the marks per criterion were awarded along a continuous scale. rather than the yes/no/maybe quantisation of the previous marking scheme, these improvements may have come from greater discrimination in attainment levels. rather than genuine improvement in student artefacts. The years 2012/13 and 2013/14 had new criteria and used continuous scales so again represent comparable marks. In tabular form, the two years look as follows.

Table 6-8: 2012-13 and 2013-14 change of average score by criterion (as marked by tutor)

	Criterion	2012	2013	Difference	As Percentage
1	Publish an SWF file of under 150k and upload it to Studynet (4)	3.54	3.65	0.11	2.75%
2	Has correct number of screens with correct headings on each (4)	3.90	3.87	-0.04	-1%
3	Appropriate choice of screen colour – providing good contrast (4)	3.68	3.77	0.10	2.5%
4	Correct width and height of the Flash file (4)	3.70	3.55	-0.15	-3.75%
5	Correct number of buttons with good colour selection (4)	3.65	3.79	0.14	3.55%
6	All buttons navigate to the correct frame script (4)	3.78	3.84	0.07	1.75%
7	Contains at least two images of you (4)	3.54	3.78	0.25	6.25%
8	Good spelling and use of language (4)	3.21	3.29	0.08	2%

9	An animation in the welcome screen (4)	3.48	3.33	-0.15	-3.75%
10	Aligned and uniform sized buttons (4)	3.45	3.38	-0.07	-1.75%
11	Text content is relevant and expressive and compact (5)	3.41	3.74	0.33	6.6%
12	Buttons have appropriate sounds on click event (5)	3.64	3.84	0.21	4.2%
13	Images of self show high production values (5)	2.71	2.95	0.24	4.8%
14	Text and image presented well on the screen (5)	3.31	3.47	0.16	3.2%
15	Animation demonstrates originality and visual appeal (10)	3.60	3.92	0.32	3.2%
16	Background music is appropriate and is controllable by user (10)	5.94	5.19	-0.75	-7.5%
17	The CV is suitable for being viewed by someone who might give you a job (10)	6.02	6.43	0.41	4.1%
18	The CV's design expresses a kind of brand identity of you as a person (10)	6.04	6.38	0.34	3.4%

If the breakdown of marks on a criterion by criterion basis is considered, it can be seen that for the first nine criteria, this is as follows.

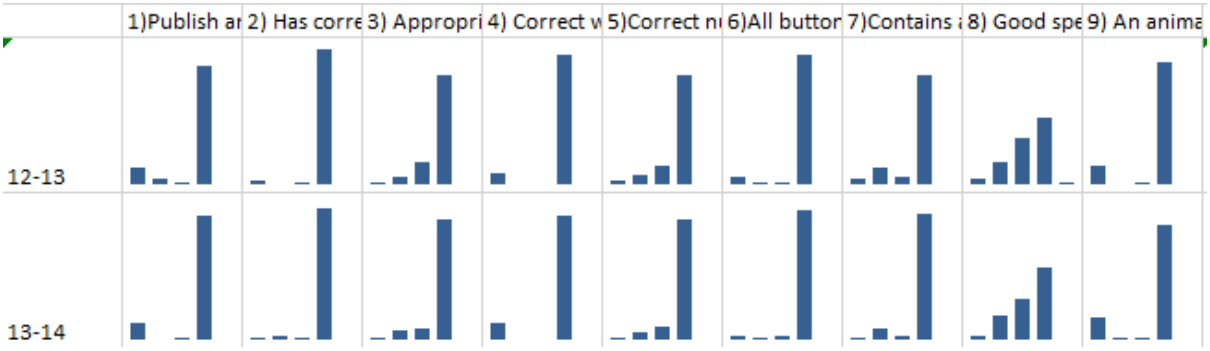


Figure 6-7: Marks Distribution First Nine Criteria in Years 12-13 and 13-14

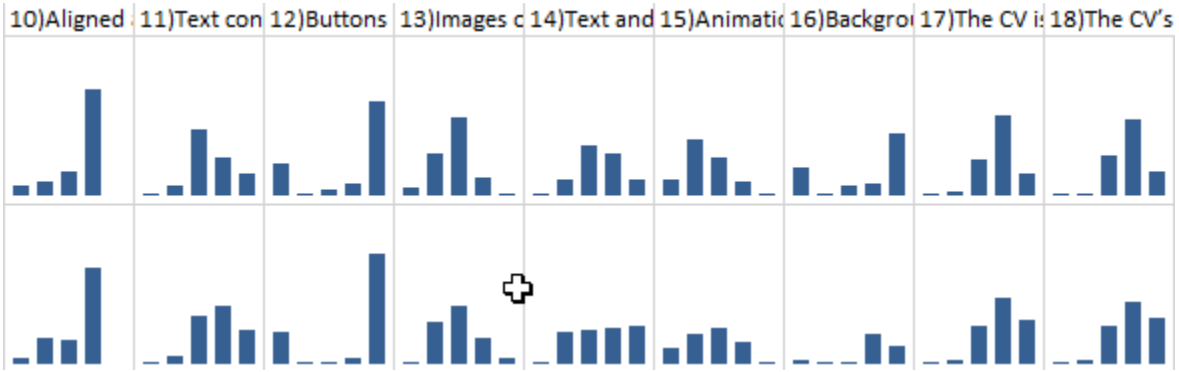


Figure 6-8: 2012-13 and 2013-14 Final Nine Criteria in Years 12-13 and 13-14

Clearly, there is a very similar distribution of marks across all the criteria. Moreover, if the average marks per criterion are compared in the form of two line plots, an almost identical pattern emerges.

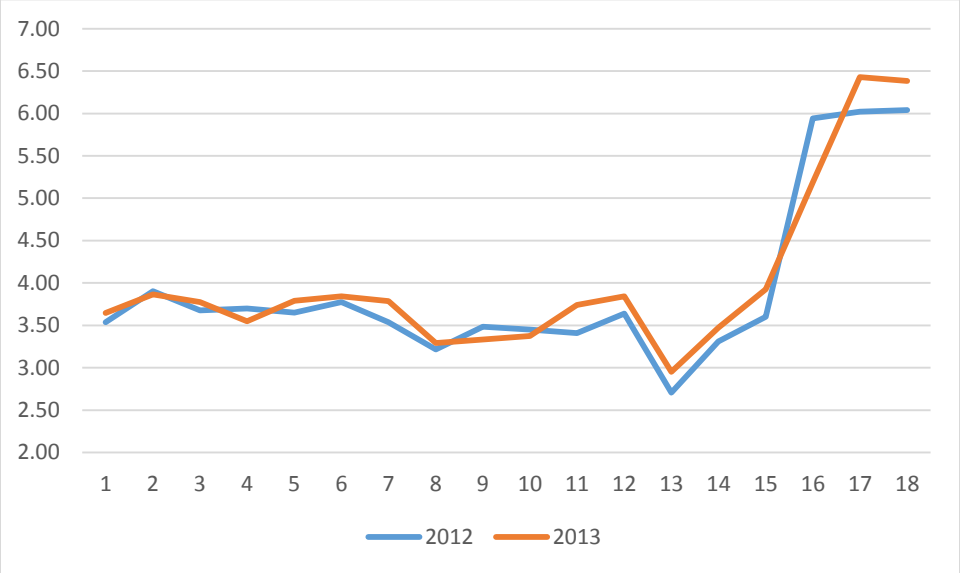


Figure 6-9: Line-plot marks by criterion in final assignment

What this clearly represents is a stabilisation of student attainment, such that between two very large cohorts, the maximum difference in attainment per criterion is 7.5% (background music) over the 18 criteria total. Given these small differences, this suggests there may be a ceiling effect here. It is difficult to be certain what changed, or what features of a successful multimedia CV were more attended to, to explain the improvements. In the beginning, the clear indication is that this was in the lower order criteria, which explains the improvement resulting in fewer students failing. Because of the change in criteria between the 2011/12 and 2012/13 years, it is not possible to trace the process on a criterion by criterion basis. However, the level of consistency between the results for the final year suggests an equilibrium was reached and that thus is unlikely that any further improvement will occur. Since the evolution of rubrics has been a part of this story, it might be worth at this point considering the literature behind rubric design and how it relates to the sequence of results elicited. Might the mere use of a rubric in itself account for some of the improvement – or is it the engagement with it?

6.2 The Nature and Effects of Rubrics

So far, the evolution in the rubrics used for marking by the tutor, from the highly objectively phrased ones used in 2009/2010/2011 to the more nuanced, compact and subjective ones used in the final two iterations of the course has been described. The motivation for the new rubric in the final two iterations was that, having raised the basic average of the students, what was sought to improve the quality of work being produced, in other words to raise expectations. The areas where this was most clearly manifested were the final two criteria:

- The CV is suitable for being viewed by someone who might give you a job (10);
- The CV's design expresses a kind of brand identity of you as a person (10);

These criteria are less straightforward. The suitability of a CV to someone (who must be imagined by the marker) capable of giving a job – requires a rich effort of evaluation. The criterion "a kind of brand identity of you as a person", requires some mental construction of the idea of "branding". As such, they deliver a much more open-ended understanding of quality than in previous formulations.

In this we see two of the central issues in academic measurement, namely: reliability (the ability of a measuring instrument to deliver reproducible marks between different raters over the same assessee, or the same rater over different assessees) and validity (to make sure the measuring instrument truly marks that which the course is aiming to teach). In this case, the aim is to teach the students to have the ability to produce an engaging multimedia artefact, suitable for its audience, with an understanding of media and the capacity to use a multimedia development tool effectively. The earlier rubric, it might be argued, was potentially more reliable, with many closed questions, limited response range and more likely to deliver agreement among markers. However, it may have lacked validity: by taking a highly atomised view of the kinds of techniques needed to be mastered, it may have miss out on the most fundamental technique of all, the ability of the student to combine the various elements into a satisfactory whole.

Another set of dimensions in which to consider rubrics are whether they use "analytical" or "holistic" criteria. By this is typically meant, whether a mark is arrived at through the totaling of scores on a number of specific dimensions (analytical) or whether it is obtained through some sense of an overall impression (holistic). This, however, may be too broad a contrast. Each criterion in an "analytical" rubric might be extremely narrow, but could also be broad. The two rubrics used by the final marker on the E-Media Design course had 16 and 18 criteria, respectively, while some of them were extremely

specific (correct width and height of file) both also contained those were highly synthesised, demanding a holistic response.

6.2.1 Research Relating to Rubric Design

These two dimensions (reliability vs validity) and (analytical vs holistic) come from two central papers in this field: Jonsson and Svingby's "The use of scoring rubrics: Reliability, validity and educational consequences" (2007) and the other being Sadler's "Indeterminacy in the use of preset criteria for assessment and grading" (2009).

Jonnson and Svingby begin by setting themselves three questions to answer:

1. Does the use of rubrics enhance the reliability of scoring?
2. Can rubrics facilitate valid judgment of performance assessments?
3. Does the use of rubrics promote learning and/or improve instruction?

In broad terms, the answer to each of these questions is a qualified yes, they do appear to support the above claims, but only when there has been a real effort actively to engage students in using the rubric. In the area of true peer assessment (where students give grades to their peers), the use of rubrics has some value in delivering greater validity of the marks - see Panadero, Romero, and Strijbos (2013). Moreover, Sadler and Good (P. M. Sadler & Good, 2006) (a different Sadler in this case) describe the negative effects when students have to grade to a too open-ended form of assessment:

> When students were simply handed a rubric and asked to use it voluntarily, but were given no training in its use, they ignored the rubric (Fairbrother & Black). When given the opportunity to self-assess using open-ended, qualitative statements instead of formal guidelines, students were terse or obscure to the point of uselessness.

These are if you like the practical advantages of using a rubric. However, the philosophical objections to using one is most compellingly expressed by Sadler in the paper "Indeterminacy in the use of preset criteria for assessment and grading" (2009) . He mentions a number of flaws with an "analytical rubric":

- you cannot revisit the work being marked for each criterion – it would take too long – therefore while viewing the work you are simultaneously holding a variety of criteria in your head at the same time and "noticing" things in their regard;

- sometimes the aggregate of the local marks and what one might like to give as a global mark does not agree'
- sometimes criteria overlap'
- selecting any particular criterion is by definition to the exclusion of others;
- if tutors allow their global understanding to influence their local scores, then the presumed feedback value of a rubric (identifying and scoring specific parts) is negated;
- intangibility – for some criteria it is simply not possible to express them economically (in a compact way with a minimum of words).

While all of these potential pitfalls are very relevant for the academic attempting to produce a grade for a student, in practice, they are probably not that important for helping the student develop judgement. Despite Wimshurst and Manning having successfully used a holistic rubric for their peer assessment, in most other studies, students have been constrained by fairly fixed criteria.

In terms of recommendations for rubrics, one of the most common is the requirement that it uses simple language and describes quality in terms that would be recognisable to the student. Andrade (2001) writes:

> The overarching principle here is that a rubric which reflects and reveals problems that students commonly experience provides more informative feedback than one that either describes mistakes they do not recognize or that defines levels of quality so vaguely as to be meaningless (e.g., "poorly organized" or "boring").

Regarding an experiment where two schools separated students into a control and a treatment group (where the treatment group had rubrics and the control group did not), she reports that the first time the experiment was run, it produced no discernible effects, writing:

> A second reason for the lack of an effect of the treatment on the first essay may be that the rubric itself was not written in particularly student-friendly terms. The second and third rubrics were written in more accessible language. (Andrade, 2001)

However, as mentioned before, all of the truly effective interventions with rubrics in the literature make some provision for students actively to engage with them. One way of doing this is the co-creation of criteria and attainment standards, giving students some ownership of the terms used. The other way, as described in the literature survey, is through the use of exemplar marking, where students have to mark

previous students work. Perhaps the main value of this, according to Moskal and Leydens, is that it brings transparency to the process:

> Sometimes during the scoring process, teachers realize that they hold implicit criteria that are not stated in the scoring rubric. Whenever possible, the scoring rubric should be shared with the students in advance in order to allow students the opportunity to construct the response with the intention of providing convincing evidence that they have met the criteria. If the scoring rubric is shared with the students prior to the evaluation, students should not be held accountable for the unstated criteria. Identifying implicit criteria can help the teacher refine the scoring rubric for future assessments. (Moskal & Leydens, 2000)

To actually mark with the students means to surface some of these "implicit" criteria by looking at their work and then justifying the mark awarded, which means they can see clearly not only what the rules are, but also, the kind of mind-set according to which the rules or criteria are applied. That is to say, they see the kinds of concrete examples to which the abstract terms must be applied. Moreover, while Moskal and Leydens believe that the surfacing of the implicit criteria can help in the refinement of the rubric, Sadler's point about the impossibility of expressing all such implicit criteria in an economical way remains and therefore, the point is not to attempt the perfect rubric but rather, to find the most serviceable one, the applicability of which can be easily demonstrated and whose overarching goal can be communicated in simple terms.

Certainly, when the principal evolution of the marking criteria is considered, it is clear that the criteria used in the final two iterations are more synthetic and compact than those in the earlier ones, and moreover, the terms used make the individual elements less "stand-alone". Note the use of "appropriate" which appears in the second rubric, but not the first. To a certain extent this is a term that requires some explanation: what might be an "appropriate" colour, sound or background music to a multimedia CV cannot be specified in absolute terms for all cultures, times and environments. However, everyone knows that a certain formality is required during job interviews and in printed CVs and so, when demonstrating using previous work, the task of pointing out what is appropriate versus what is not is not necessarily that difficult.

Other words that might be opaque in the absence of demonstration are more common in the later rubric "expressive, compact, high production values, presented well" – these too would need to be demonstrated to manifest their meaning. The earlier rubric relied much more on highly "objective"

statements such as “correct”. It sometimes brought in technical criteria, but without any association with the overall effect: “button clicks have small sounds”; “an immediately visible and functioning background music toggle”, whereas the later rubric talks about “appropriate sounds” for button clicks and “appropriate and controllable”, much more abstract and all-encompassing terms. to describe the background music facility. The earlier rubric also had some ungainly writing: “Correct and Nice Positioning”, “Easy on the Eye Content”. In these cases, associating a colloquial style but with an authoritarian judgement “Correct” “Not too little not too much” – as opposed to “relevant expressive and compact”. Clearly, there has been an effort to improve compactness of the criterion as well as the expansiveness of scope in the words contained in it - none of are fully comprehensible away from some demonstration of the application of that criterion. It is only through the demonstration of the rubric that these terms can be made concrete and instances of their application become recognisable.

Certainly, the final rubric expects more of the students: beyond just mastering the technical skills required, but also doing so in a way which embodies a concern for the whole rather than just the mastering of the parts. It was introduced after the second iteration using rubrics, by which time it seemed the problem of the underachievement of the three lowest deciles had been solved and at this point we wished to explicitly raise expectations. Most dramatically of all were the final two criteria in the rubric, which required suitability for the audience and then, finally the creation of a “brand identity” of the student as a person.

As the literature in the preceding section has shown, merely producing an ambitious rubric will not necessarily mean students will attempt to fulfil it. It is the act of *engaging* them with marking criteria that is the difficult thing, but I believe, from the results shown above, that this has been achieved through the approach of using collective marking of exemplars using clickers. In the next chapter, how students marked, given the different conditions and rubrics they were asked to do so, is investigated.

Chapter 7. Analysis of Voting in the BSc Course over the Years

So far, we have seen how the course has evolved in terms of the technology used, the rubrics used, the training set of exemplars and the conditions under which the evaluation sessions took place have been covered. Summed up briefly they are as follows.

Table 7-1: Summary of Conditions

Year	Technology	Credit	Exemplars	Rubric
10-11	Clickers (rehearsal) but QMP (final)	25%	From 09-10 cohort – anonymized by cartoons	2010S Yes/No/Maybe
11-12	Clickers (rehearsal + final)	10%	"	2011S More Attainment Descriptors
12-13	Clickers (rehearsal + final)	10%	"	Rubric 2012S (truncated)
13-14	Clickers (rehearsal + final)	0%	From 12-13 cohort – anonymised by photographs	Rubric 2013S

In this chapter, the voting history on the course during the four years is analysed to see what inferences can made about which were the crucial factors determining increased student achievement.

7.1 Limitations of the Data

The greatest difference between this course and the previous (MSc) course examined is that regarding the latter, the format was largely unchanged apart from the removal of credit for participation in the last two iterations. This course, however, represents a much more dynamic process, where different approaches were employed over the four iterations under consideration, arising from discussions between the tutors as well as interactions with the students and consequently, the data between iterations are not so easily comparable.

The other substantial difference is the type of rubric used. The rubric used with the MSc course comprised, in the case of the prototype assignment, two holistic criteria, and in the case of the final artefact, three holistic criteria. In the case of E-Media Design, the initial rubric had 15 dimensions, whilst in the final iteration there were 18. These rubrics were written in such a way that the first half of the criteria were simply about satisfying the basic requirements of the assignment: among which were, that

the multimedia CV did indeed have two images, that the buttons did indeed work and clicking on them sent the user to the right page. The second half in some cases duplicated these criteria, but to a higher standard; the distinction between the levels being illustrated in the table below.

Table 7-2:Criteria by Level

4. Correct number of buttons with correct colours for them	(lower criteria)
13. Extremely well positioned and pleasant looking buttons	(higher criteria)
10. Good easy on the eye content (text and image) not too little not too much and all relevant	(lower criteria)
14. Extremely well judged content	(higher criteria)
8. Motion Tweening Of Position of Things in the animation welcome screen.....	(lower criteria)
12. Some very clever and visually elegant animation	(higher criteria)

Clearly, in the case of many of these criteria, there is little scope for argument or difference of opinion. There is an animation or there isn't; there are two images of the student in the page or there aren't. Using one set of marking event statistics as an example (the first item to be marked in the "final" evaluation in 2011-12), it can be seen that eight of the 15 criteria have a massive preponderance of one value, indicating almost unanimity among the students' evaluation.

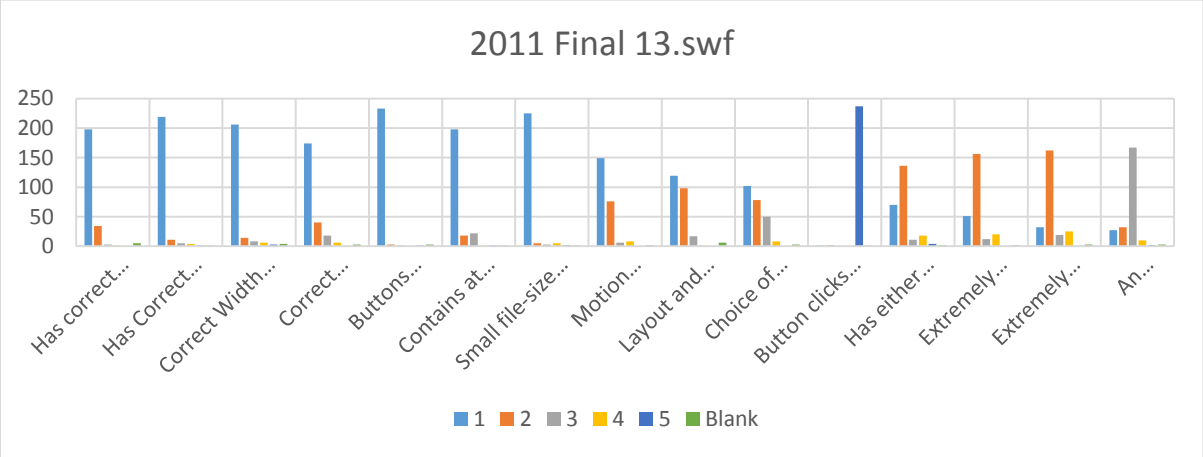


Figure 7-1:Numbers of Students Per Answer Per Criterion

The correlation coefficient between tutor and student marks, therefore, does not make much sense when (in 2010) a number of options can only be answered by a choice of Yes/No/Maybe. This was also the case in subsequent years when the lower-order criteria often delivered almost unanimous agreement. In the second iteration, whilst some of the criteria had more attainment levels, a large

number of them still only had three. Consequently, other measures, such as the median pairwise correlation are also not that helpful. In the marks given to students based on their participation in the evaluation events, closeness to the tutor was measured by levels of agreement with the tutor marks. This was undertaken only during the first three years, according to the following procedures.

Table 7-3: Method for awarding marks based on agreement with tutors over the years.

Year	Questions	Condition	Score	Condition	Score	Condition	Score
10-11	All	=	1	>=1 difference	-1	<1 difference	0
11-12	1-8	<1 difference	1	<=2 difference	0	>2 difference	-1
	9-16	<=2 difference	1	<=3 difference	0	>3 difference	-1
12-13	All	=	1	<=1 difference	.5	Difference >=2	0
13-14	All	NOT USED					

In the year 2012-13, a temporary rubric was created for students to use, owing to technical problems in the rehearsal event. (The rubric there reduced the number of questions to answer – in that year the reference mark from which students were given marks owing to their divergence was taken from a rounding of the mean score). In 2013-14 no reference mark was used.

As a consequence, it is difficult, though not impossible, to make meaningful use of correlation between tutor and student marks from the reference marks. As will be seen later, some limited use can be made of them for the 2011-12 and 2013-14 cohorts.

7.2 Comparison of Inter-Rater Reliability using Krippendorf's Alpha over the course

As in the case of the multimedia course, there is one measure that can assess the level of homogeneity or inter-rater agreement with the type of data collected (without reference to tutors' marks) and that is Krippendorf's alpha. In those cases where clickers were used, and thus, raters can be associated with ratees over whole sessions, an overall score for the session as a whole can be obtained. Because the first iteration of the course using feed-forward techniques took place in a very different manner (students logging into computers and performing their evaluations independently of the rest of the cohort and without any sense of how others voted, nor synchronized on a criterion by criterion basis) it has thus been excluded from the overall comparison of Krippendorf's Alpha on a session by session basis (Table 7 4:Krippendorf's Alpha by Event). However, data relating to individual artefacts in that year have been included for comparison in the succeeding table (Table 7-5:Krippendorf's Alpha by Item Evaluated) where levels of agreement per artefact are examined.

Table 7-4:Krippendorf's Alpha by Event

Year	Event	N raters	N artefacts rated	Grade Brearing	Kalpha	% Unanswered Prompts
11-12	Rehearsal	139	4	No	0.503	41.9%
	Final	241	4	Yes	0.695	1.7%
12-13	Rehearsal	No data collected because of technical issues				
	Final	196	3	Yes	0.631	0.8%
13-14	Rehearsal	102	2	No	0.604	31.3%
	Final	181	2	No	0.406	36.2%

The Kalpha value represents the level of agreement between those who actually voted. In each of these events, each artefact was voted per criterion, however, not all the students necessarily voted on each prompt. The unanswered prompts column represents the percentage of non-clicks as a proportion of all the possible student responses to prompts. It is very clear from the table, that where there is no grade given for the quality of the students' marking, some students did not vote at all and just watched, whilst others vote only intermittently. Next, the Kalpha per item is evaluated (and here it is possible, with qualifications to include the 2010 data – the qualification being that raters had the opportunity to rerate a piece of work if they wished and thus, it is not necessarily 1 judgement per rater in the data collected).

*Table 7-5:Krippendorf's Alpha by Item Evaluated (*no record of which files evaluated during 12-13 iteration)*

			CV1	CV2	CV3	CV4
10-11	Filename		1.swf	2.swf	3.swf	4.swf
Rehearsal	Kalpha		0.223	0.349	0.290	0.400
	%Unanswered		1.7	0.4	2.6	1.4
10-11	Filename		10.swf	11.swf	12.swf`	13.swf
Final	Kalpha		0.327	0.362	0.275	0.267
	%Unanswered		3.6	3.7	4.2	3.3
11-12	Filename		38.swf	31.swf	21.swf	
Rehearsal	Kalpha		0.451	0.070	0.536	
	%Unanswered		31.5	36.6	57.7	
11-12	Filename		13.swf	22.swf	32.swf	
Final	Kalpha		0.664	0.660	0.724	
	%Unanswered		1.1	1.9	2	

12-13*	Filename		?	?	?	
Final	Kalpha		0.653	0.619	0.416	
	%Unanswered		1.2	0.7	0.3	
	Filename		Cv4.swf	Cv2.swf		
	Kalpha		0.659	0.286		
	%Unanswered		20.6	41.9		
13-14						
	Filename		Cv1.swf	cv3.swf		
	Final		0.471	0.285		
	%Unanswered		19.1	53.4		

It can be observed that the level of inter-rater agreement on all occasions during the first cohort was fairly low. This might be explained by the fact it was the first time the technique was used, but it is also worth reflecting on how very different the modality was. In this year, students individually evaluated items in isolation from others by virtue of making the evaluations on isolated QuestionMark based survey screens via a web browser. In all the other events, which were undertaken with clickers live in a lecture theatre, immediately after making a judgement they would see the class average on screen, which would clearly put into relief any discrepancy between the scores they awarded individually and those awarded by the class as a whole. In the multimedia focus group, it was found that individual markers could be influenced by their peers. This might constitute evidence of the way social marking can lead to a kind of collective sensibility emerging. However, it might also reflect the fact that the choice of the exemplars to rate was made without any premeditation or any thought as to their suitability.

In the second year, the Kalpha value was much higher, except for one outlier (the second evaluated cv 31.swf during the rehearsal). Otherwise in this year, the level of inter-rather agreement for the final event was very high and in 2012-13 high levels of agreement were again recorded. However, in all these years, the students did receive reward for marking near to the tutors (or what they believed to be near to the tutors) and so this might have been the cause of the increased homogeneity. In 2013-14, however, within the evaluation events, while the level of agreement and participation on the first rated item was respectable, the second was much lower, on both occasions. This was the first time when there was no mark for “the way you graded” during either evaluation event, which might explain why the rehearsal event had higher inter-rater agreement scores (potentially being an elective event and so, only the most committed students attended).

Interestingly, the statistics for those events reveal that a large proportion of the students did not bother carrying out the marking for the second artefact, presumably believing they had learned all they needed to know about grading from the marking event. As can be seen above, the evaluation events being graded or not has a huge impact on the level of participation, however, this has an inconsistent impact on the level of agreement. Another area where it seems to have a large effect is the time taken actually to make a click or a judgement.

7.3 Time Taken to Make Judgements Between Grade Bearing and Non Grade Bearing Evaluation Sessions

One very interesting feature of the Turning Point clickers we used, is that they not only measure choices made by the people using them, but also the time elapsed between exposure to a question and the response given to it. Over time, this results in interesting data. Before considering these, the contextual factors that influence the amount of time taken to answer questions need to be discussed. These rehearsal and summative sessions take an hour, with the amount of time taken to respond to any particular criterion not entirely being at the students' own discretion: they will often by hurried along by the tutor to answer the questions (for instance, towards the end of the session). Very often the tutor would count slowly to 10 to get the remainder of the students to vote, which has particular relevance for those criteria at the end, where time pressure may be greater. The time taken to assess the first presented criterion tends to be longer, because during its presentation, the artefact as a whole is also presented to the class, thus enabling the students to have a "global" understanding of it.

Time taken per response could not be collected in 2010-11 using Questionmark Perception, because the time taken per answer is not measured on that platform. In the following three years using the Turning Point software, a record of time taken per clicker answer after the prompt was recorded. However, as noted before, during the 2012-13 there were technical issues during the rehearsal event meaning no date was able to be collected. Therefore, I will limit the comparison to the 2011-12 and 2013-14 iterations, where we have both the time per response in both the rehearsal and the the "final" sessions, and we are also using the same platform.

The main difference between these two iterations as far as marking is concerned, is that during 2011-12, students received 10% of the course marks for how similar their marking of exemplars was to the tutors. In 2013-14, there is no such premium. One of the most striking things about the 2011-12 iteration in

terms of marking times, is how much longer students took to mark artefacts in the "final" (summatively marked) event as compared to the "rehearsal" (formatively marked) one.

Table 7-6:Time Taken to Mark Each Criterion 2011-2012 Year in both "Rehearsal" and "Final" events

Criterion	Order	Rehearsal (Average Time + Ranking in Times)		Final (Average Time + Ranking in Times)	
Has correct number of screens with correct headings on each	1	25.58	1	74.95	1
Has correct background colour	2	8.45	14	36.31	7
Correct number of buttons with correct colours for them	3	12.93	10	19.33	12
Make buttons navigate to correct frames using simple action script	4	12.00	11	29.44	9
Contains at least two images of you	5	16.22	8	29.11	10
Small file-size	6	16.58	7	36.48	6
Motion Tweening Of Position/Visibility in the welcome screen	7	21.41	2	42.03	5
Layout and Positioning of Buttons and Text	8	16.95	6	50.84	4
Choice of material, text and tone appropriate for a CV (text and image)	9	19.26	4	54.22	2
Button clicks have small sounds associated with them	10	14.41	9	13.28	14
Has either very clever or visually elegant animation	11	11.58	13	35.95	8
Contains some very well positioned, elegant, and suitable buttons	12	11.63	12	22.41	11
The text is relevant, of appropriate length and evenly distributed across the cv, images are appropriate to cv and of a high quality	13	18.83	5	16.52	13
An immediately visible and functioning background music toggle	14	19.35	3	51.14	3

When we view this graphically the difference is obvious:

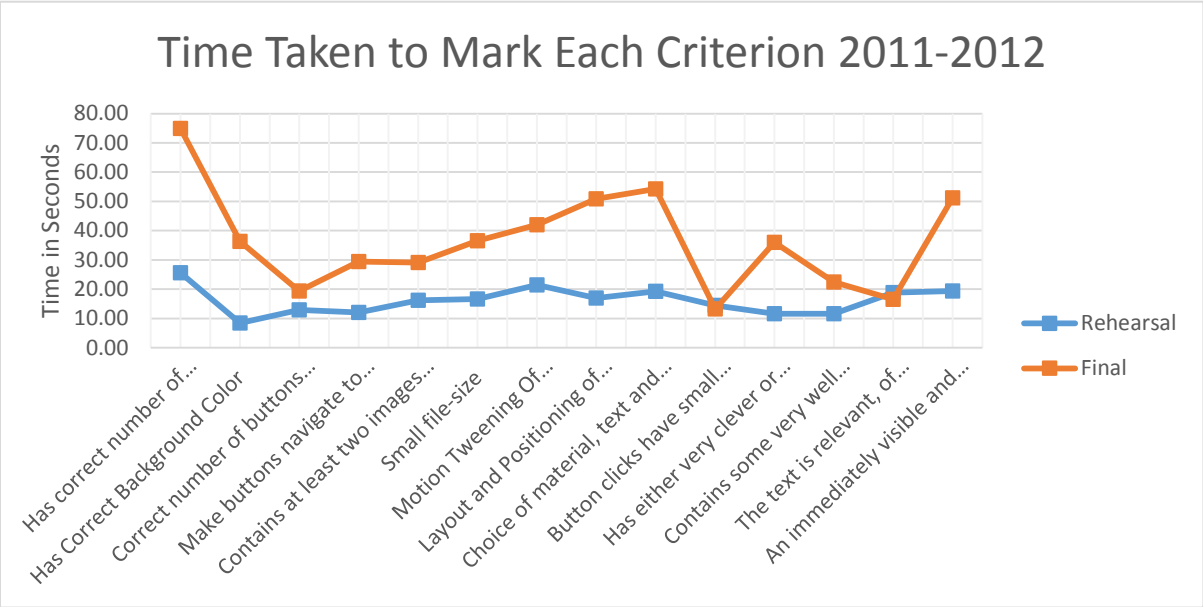


Figure 7-2:Time Taken to Mark Each Criterion 2011-20142Year in both "Rehearsal" and "Final" events

As before, the first criterion can be ignored because that is the one where the artefact is also being presented to the class. However, as can be seen, aside from two criteria (button clicks have small sounds…, text is relevant…) the amount of time taken in the "final" event, which is grade bearing dependent on similarity to tutor marks, is much greater. In some cases, in the order of 20 to 30 seconds. Hence, this could be the result of students applying more thought to something that will impact on their grades, but equally, it may be the effect of "exam conditions". Any event like this requires that students do not communicate and therefore, the act of making judgements under conditions of silence might also impart solemnity to the event, which is not the case under the more raucous conditions of the rehearsal, where there is discussion between the students and contestation with the presenter.

Now let us look at the difference between time taken in rehearsal and final events in 2013-14 (when both were formative).

Table 7-7:Time Taken to Mark Each Criterion 2013-2014 Year in both "Rehearsal" and "Final" events

Criterion	Order	Rehearsal (Average Time + Ranking in Times)		Final (Average Time + Ranking in Times)	
Publish an SWF file of under 250k and upload it to Studynet	1	26.96	16	71.12	17

Has correct number of screens with correct headings on each	2	17.22	10	19.16	13
Appropriate choice of screen colour – providing good contrast	3	17.15	9	15.52	8
Correct number of buttons with good colour selection	4	16.15	7	11.55	4
All buttons navigate correctly	5	11.25	2	9.36	2
Contains at least two images of you	6	12.06	3	10.66	3
Good spelling and use of language	7	25.74	15	28.14	15
An animation in the welcome screen	8	9.36	1	6.57	1
Aligned and uniform sized buttons	9	19.45	13	12.14	5
Text content is relevant and expressive and compact	10	16.66	8	30.69	16
Buttons have appropriate sounds on click event	11	12.97	6	12.29	6
Images of self show high production values	12	17.55	12	18.38	12
Text and image presented well on the screen	13	12.65	4	17.32	9
Animation demonstrates originality and visual appeal	14	17.26	11	18.35	11
Background music is appropriate and is controllable by user	15	36.41	17	18.15	10
The CV is suitable for being viewed by someone who might give you a job	16	12.7	5	13.55	7
The CV's design expresses a kind of brand identity of you as a person	17	22.61	14	28.06	14

When considering this graphically the level of similarity of time taken per marking for the rehearsal and the "final" events becomes clear, notwishstanding two instances of disparity for particular criteria.

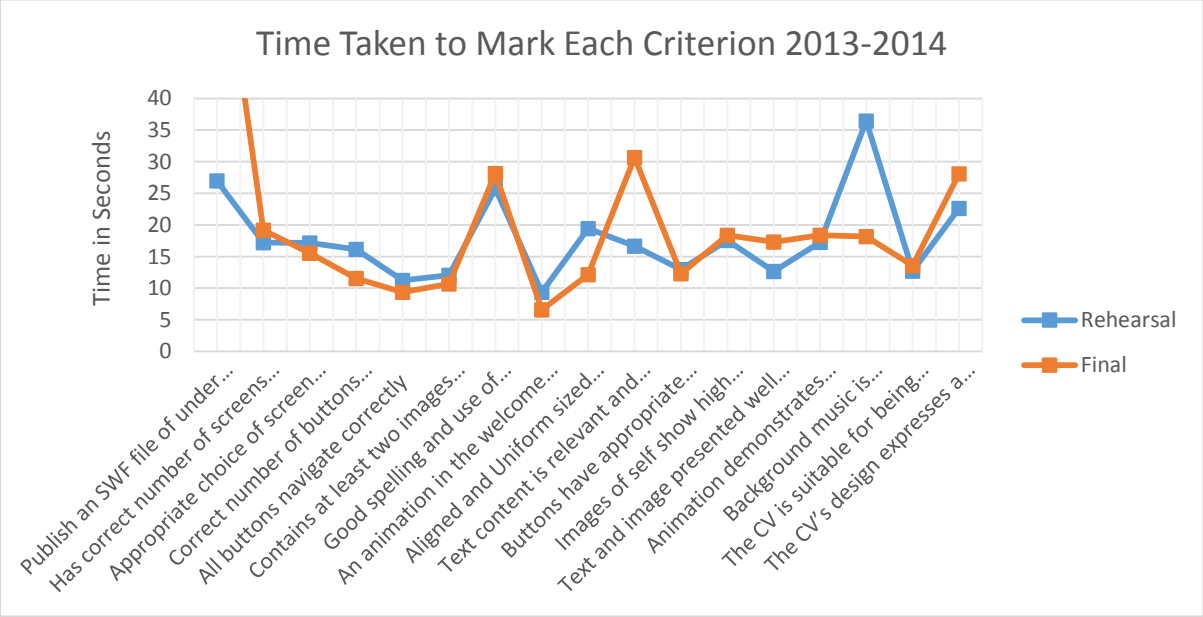


Figure 7-3:Time Taken to Mark Each Criterion 2013-2014 Year in both "Rehearsal" and "Final" events

Clearly there are a few "spikes" where various criteria appear to take different amounts of time between the events, but the overall trend is fairly even.

Another interesting comparison is the amount of "non-clicking" by students – that is to say, cases where a prompt is made, but no response is recorded for a student.

Table 7-8:Percentage of Non-Clicking Students By Year Modality and Artefact

Year	Cv1	Cv2	Cv3
11-12 Rehearsal	31.5%	36.6%	57.7%
11-12 Final	1.1%	1.9%	2.0%
13-14 Rehearsal	20.6%	41.9%	
13-14 Final	19.1%	53.4%	

Two things appear obvious here. Firstly full participation seems only to take place under summative conditions. Under formative conditions, participation is always down, but also, always deteriorates over the session.

It is difficult to know what might motivate this: the difficulty of concentration over time, or perhaps the fact that everything they wished to find out was answered in the first act of marking. Certainly, there does not seem to be a straightforward relationship between the level of engagement at the marking events and the quality of work students produce at the end of the course. It has been shown how the average final mark increases through the four iterations, and yet it is the 2011-12 cohort who demonstrate the greatest level of engagement with the marking events. Equally, it must be pointed out that these evaluation events were merely one hour sessions and yet, the work the students subsequently undertook for their multimedia CVs would take much longer and moreover, would have been submitted three weeks later. For this reason, it worth looking into how these marking events were subsequently understood and framed by the students as well as what effect it had on their subsequent practice. This is undertaken in the next chapter when the attitudes of students in a focus group held during the final iteration of the course are reported on.

As mentioned above, the dynamic nature of the course during the use of clickers and the different approaches adopted and amended make a straightforward narrative of the whole course difficult to achieve. However, It is possible understand much about the process of how collective marking influenced student achievement by concentrating in detail on the 2011-12 and 2013-14 iterations. Both

these iterations were marked by innovation at the level of the training set presented. In 2011-12, the tutors undertook an exercise of remarking the 40 examples from the training set to find the best artefacts to use as exemplars, since it was clear that certain examples could generate much more divergent responses than others. Asking students to participate in an exercise where some grading awards would come from the level of congruence with tutor marks, required at least some artefacts where gradings would not be so controversial. In 2013-14, a new set of training examples was used where the anonymisation would not be through using cartoon characters, but by comparable stock photos. Also, these two iterations represent a good comparison between grade bearing evaluation and non-grade bearing and so illustrate clearly the consequences of both approaches. Here we will attempt to relate how the students marked to the artefacts they were marking.

7.4 2011-12 Iteration

7.4.1 Details of Student and Tutor Marking

In 2011, there was a rehearsal event where three flash artefacts were evaluated and a final (summative) event where a further three artefacts were evaluated. The exemplars were of a much less polished kind and all relied on a yellow/purple color scheme, which perhaps was not ideal, but formed part of the last minute customisation challenge used in 2009-10. In that year, part of the customisation challenge at the end was that students had to modify their submission in one hour such that buttons were white text on a purple background and the background of the page was yellow. (These customisation challenges were used to discourage plagiarism and contract cheating – forcing students to make last minute modifications meant they had to know Flash really well).

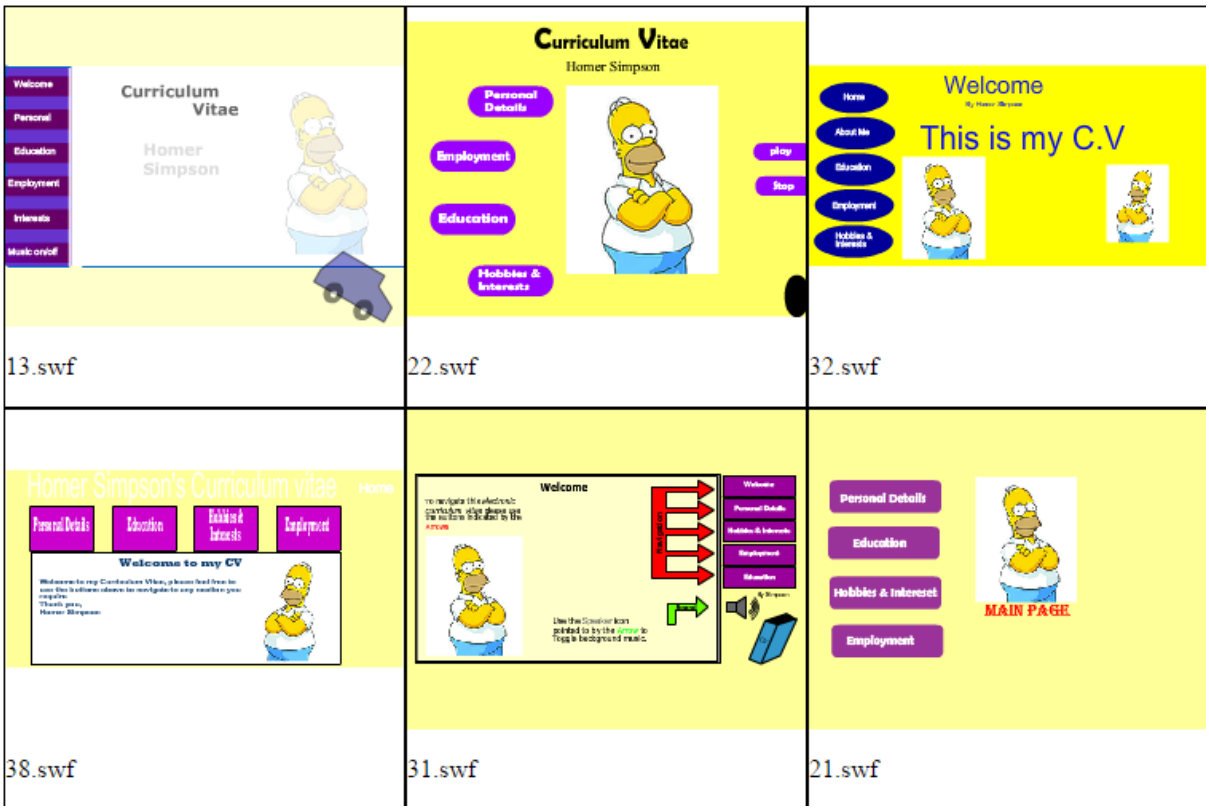


Figure 7-4:Exemplar Set used in 2011-12

Whilst for the 2013-14 iteration, the comparison between tutor and student marks over the exemplars was obtained through the use of how the tutor marked the originally submitted work (prior to anonymisation), in the 2011-12 iteration, two points of comparison were made: the original mark (when marked with a very *yes/no* rubric, offering very little granularity of achievement per criterion) and from an online exercise the tutors undertook, which involved marking many of the anonymised pieces of work in order to find which might be the most suitable as a training set. Unfortunately, because there were some errors in reproducing the sounds in the anonymised set (some anonymisations resulted in sounds not playing), this resulted in two criteria being invalid for comparison (background music toggle + buttons have small click sounds). This, together with the fact that the images used were fundamentally different from what was originally submitted (cartoons rather than photographs), means comparison with the original mark is now insufficiently robust to represent useful data.

For this reason, the data from the tutor exercise is used as the point of comparison. The tutor exercise contained 15 criteria, however, for reasons of time, during the rehearsal event, only 13 of these criteria were used, but in the final event, 15 were used (See table below).

Table 7-9: Criteria Used, Rehearsal vs Final Event

Criterion	R	F
Q1 Uploaded to Studynet	N	N
Q2 Has correct number of pages with correct headings on each (5)	Y	Y
Q3 Correct background colour (5)	Y	Y
Q4 Correct width and height of the Flash file (5)	N	Y
Q5 Correct number of buttons with correct colours for them (5)	Y	Y
Q6 Make buttons navigate to correct frames using simple action script (5)	Y	Y
Q7 Contains at least two images of you (5)	Y	Y
Q8 Small file-size (5)	N	Y
Q9 Motion Tweening Of Position of Things in the animation welcome screen OR Motion Tweening of Visibility of Things (for fade ins and fade outs) in the animation welcome screen (5)	Y	Y
Q10 Correct and nice positioning of buttons and content (5)	Y	Y
Q11 Good easy on the eye content (text and image) not too little not too much and all relevant (10)	Y	Y
Q12 Button clicks have small sounds associated with them (10)	Y	Y
Q13 Some very clever and visually elegant animation (using Shape Tweens or Motion Guides or Masking) in the animation welcome screen (10)	Y	Y
Q14 Extremely well positioned and pleasant looking buttons (5)	Y	Y
Q15 Extremely well judged content (5)	Y	Y
Q16 An immediately visible and functioning background music toggle (10)	Y	Y

Consequently, given that three criteria (Q1, Q4, Q8 – greyed out in the table) are not being used, for purposes of overall score comparison, the sum of all the other criteria is out of 85. In the cases of the tutor exercise, the student rehearsal and the final student evaluations, the choices available were on a scale of 1-5 – though the marks awarded by tutors for the final assignment would be out of 10 (having a higher weighting). Therefore, to calculate this final score the values (input by tutors) for Q11, Q12, Q13 and Q16 are halved so as to give a total of 5. However, in order to calculate the level of correlation between the tutor and student marks only the undivided values for all of the remaining 13 criteria used, which provides the following data.

Table 7-10:Correlation of Marks Between Tutor Exercise and Student Voting (2011-12)

			Q2 Has correct number of pages with correct headings on each (5)	Q3 Correct background colour (5)	Q4 Correct number of buttons with correct colours for them (5)	Q6 Make buttons navigate to correct frames (5)	Q7 Contains at least two images of you (5)	Q9 Motion Tweening Of Position of Things (5)	Q10 Correct and nice positioning of buttons and Content (5)	Q11 Good easy on the eye content (10)	Q12 Button clicks have small sounds associated with them (10)	Q13 Some very clever and visually elegant animation (10)	Q14 Extremely well positioned and pleasant looking buttons (5)	Q15 Extremely well judged content (5)	Q16 An immediately visible and functioning background music toggle (10)	Score	Correlation
Final	13.swf	Tutors	5.00	5.00	5.00	5.00	5.00	5.00	5.00	4.50	1.00	4.50	4.50	4.00	4.00	84.12	0.98
		Students	4.81	4.83	4.60	4.96	4.71	4.53	4.42	4.16	1.04	4.05	3.98	3.85	3.31	77.44	
	22.swf	Tutors	5.00	5.00	5.00	4.00	5.00	3.50	4.00	2.50	1.00	1.50	3.00	1.50	4.50	64.71	0.88
		Students	4.39	4.64	4.65	4.94	4.87	2.86	3.74	3.43	1.05	2.06	3.53	3.24	4.36	69.01	
	32.swf	Tutors	5.00	5.00	4.00	5.00	4.50	2.50	4.00	3.50	1.00	1.00	1.50	2.50	1.00	55.29	0.87
		Students	4.53	4.58	3.88	5.00	4.85	3.18	4.10	3.93	1.02	2.39	3.75	3.74	1.03	63.94	
Rehearsal	38.swf	Tutors	5.00	5.00	5.00	5.00	3.00	3.00	3.50	4.00	1.00	3.00	3.50	3.50	1.00	64.12	0.58
		Students	3.38	1.00	4.34	4.79	2.95	2.84	3.34	2.57	1.44	2.09	2.50	3.05	1.35	50.71	
	31.swf	Tutors	5.00	5.00	5.00	5.00	5.00	5.00	5.00	4.50	5.00	4.50	4.50	4.50	4.50	95.29	0.73
		Students	4.70	4.53	4.62	4.74	4.04	4.81	4.59	4.18	4.43	4.03	4.38	3.89	4.04	86.65	
	21.swf	Tutors	4.50	5.00	4.00	4.50	5.00	1.00	3.00	2.50	3.50	1.00	1.00	2.00	1.00	54.12	0.89
		Students	4.26	4.81	4.75	4.53	3.73	1.54	3.72	2.92	3.56	1.63	3.00	3.07	1.19	61.17	

What is interesting in the 2011 data, is that the correlation between tutor and student marking appears to increase incrementally after each evaluation is made. However, this has to be weighed against other evidence, whereby in the rehearsal session, much as was the case in 2013, students appeared to lose interest as the session progressed. Below, is a graph of the number of students participating for each criterion being evaluated during the rehearsal event.

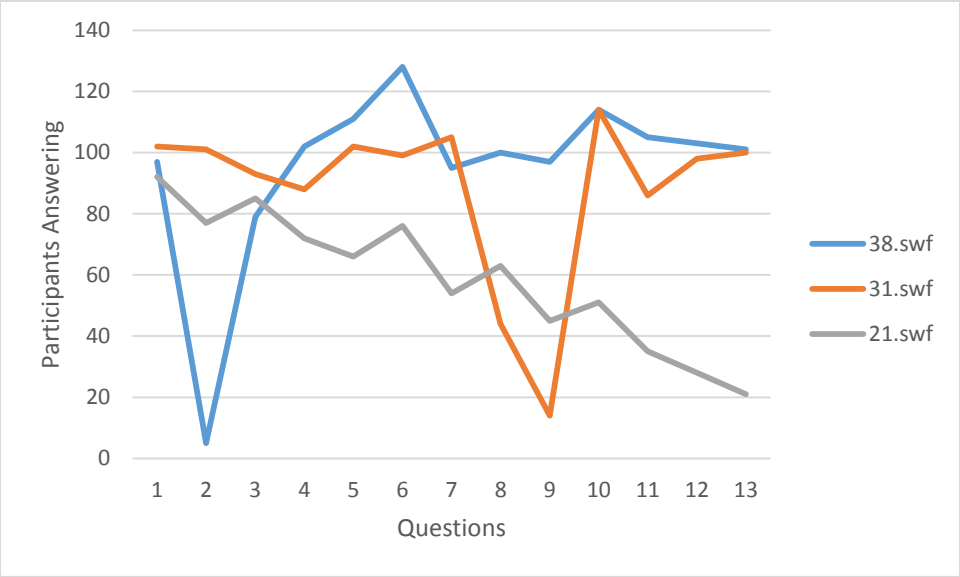


Figure 7-5:Participation by Criterion by Artefact- 2011-12 Rehearsal Event

As can be seen, the final artefact to be evaluated (21.swf) during the rehearsal event has a smaller number of students participating and declines rapidly during the evaluation. However, in the final exercise, because the students were being graded on how similar their marks were to those of the tutor, the same falling off is not observed: (15 questions appear here, because whilst two were not considered for the purposes of correlation, they were nonetheless asked).

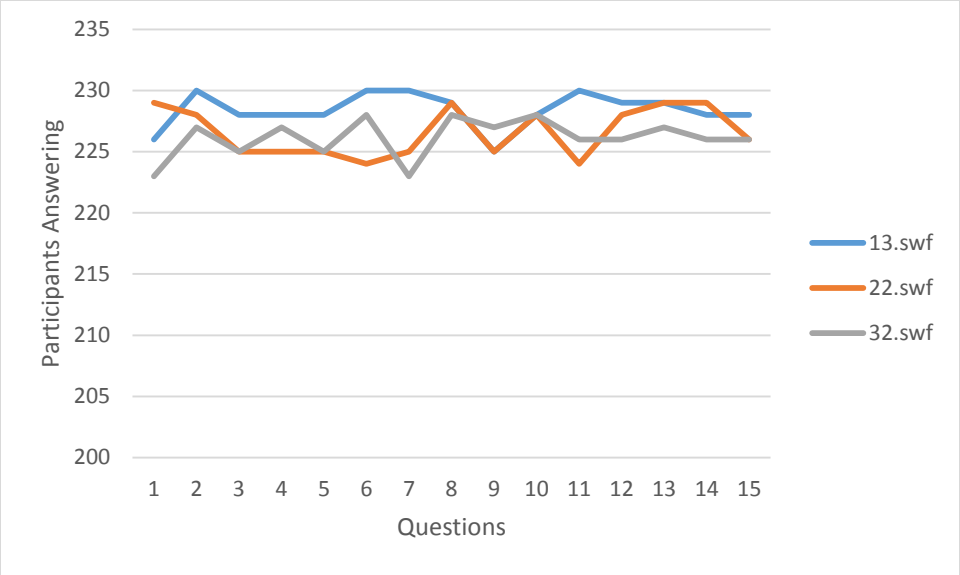


Figure 7-6:Participation by Criterion by Artefact- 2011-12 Final Event

The fact that the comparability with the tutors marking remains so high here is satisfying. However, it must be said, the way the exercise was set up, to mark like the tutors was precisely the goal.

7.5 More In Depth Look at Voting Patterns in the Final Year

As has been explained, the E-Media course involved a continuous evolution where different approaches were tried, and based on the statistics already seen, it could be said to have reached maturity in its final iteration, which was characterised by:

- A more polished and well thought out training set;
- A more complete rubric;
- Elimination of credit for grading similarly to the tutors.

I propose therefore to focus on the statistics of this final iteration to address the question of how the techniques of feed-forward evaluation enabled the students to produce better work.

7.5.1 The Training Set

Whereas in previous years, the training sets came from a random selection of 40 submissions during the 2008-2009 academic year, in this year, the training set comprised six submissions during the 2012-13 academic year. Moreover, the previous training sets involved anonymising students by replacing any pictures with cartoon characters, whilst in this case, the original students' images were replaced by stock photograph images of students with the same gender and ethnicity. As in previous years, there was a rehearsal session and a final session. Because of the length of the rubric, in practice, we found we

could only carry out two evaluations per session. Below are screen captures of the six flash files that were used.

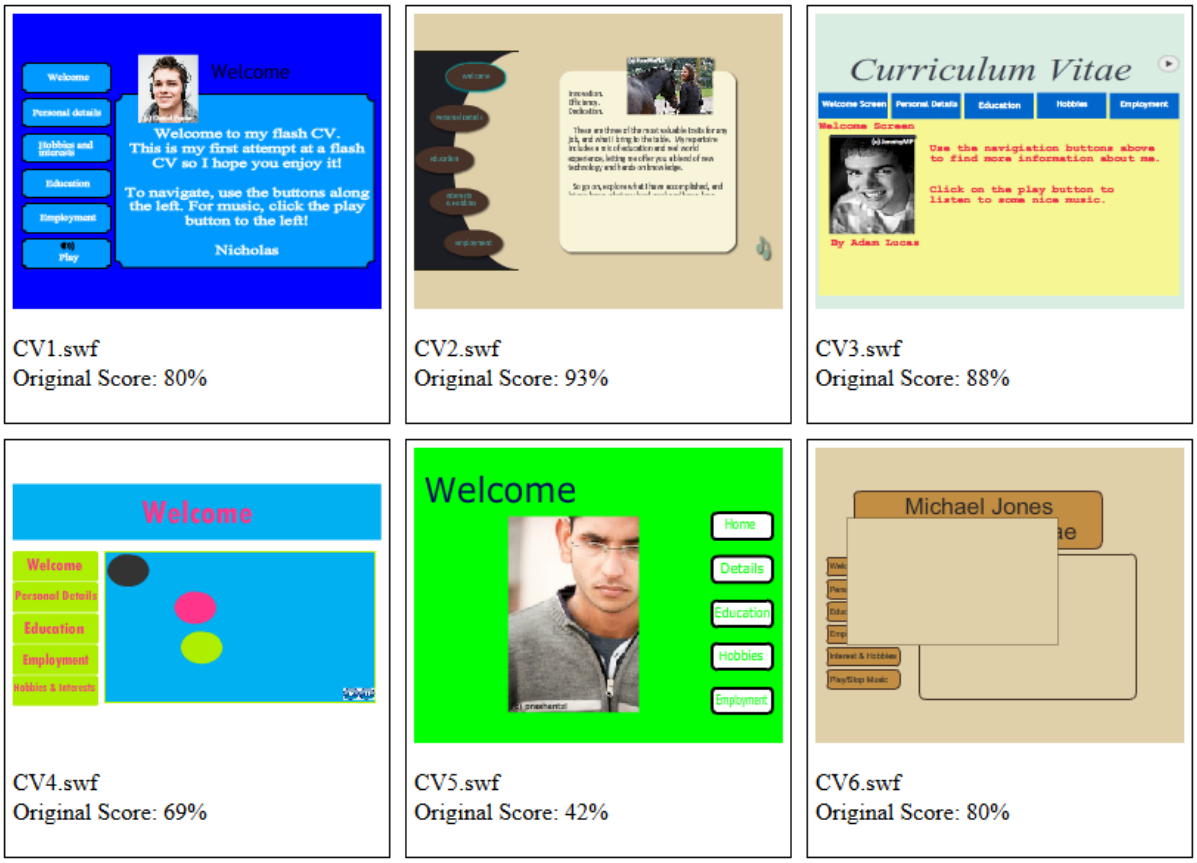


Figure 7-7:Exemplar Set Used in 2013-14

The training set essentially contained four first class level pieces of work, one 2:1 level (CV4.swf) and one 3rd (CV5.swf). In the feed-forward sessions we only had time to cover CV4.swf and CV2.swf in the rehearsal session and CV1.swf and CV3.swf in the final one. The marks presented here are the scores given to the student in the previous year, before the work was anonymised. Hence, certain marks given (for example the quality of the photographs, and some of the text regarding personal details) might not be valid after anonymisation, however, the mark at least points to some basic sense of the quality of each artefact.

7.5.2 Levels of Participation and Agreement in The Rehearsal Session

In the rehearsal session, as aforementioned, cv4.swf and cv2.swf were evaluated. Unlike in previous years, there was no credit given for level of agreement with the scores of the tutors and therefore, it was not required that the students voted on every occasion.

The blue line here shows the percentage of students present during the rehearsal, who voted for each criterion for the first demonstration and the orange line shows this percentage for the second. There does not seem to be any trend of diminishing attention within presentations, but it is clear that during parts of the second presentation, 20% fewer of the students were actually engaged or clicking.

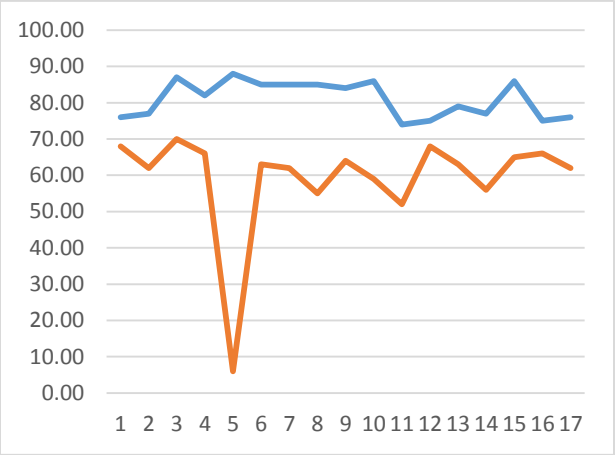


Figure 7-8: Voting Participation (%) by Artefact and Criterion 2013-2014 Rehearsal Event.

This is interesting in that the second presentation was by a long way better than the first. It could have been the case that the students who were no longer following the second presentation had learned all they needed to learn by the time it started. The steep drop for question number 5 (All buttons navigate correctly) was caused by an error on the part of the lecturer moving too fast to the subsequent slide.

Turning to the marks awarded by the students compared to those given by the tutor, these were as follows.

Table 7-11:Scores Given by Tutors vs Scores Given By Students 2013 Final Event

	CW4.swf	CV4.swf	CV2.swf	CV2.swf
	TUTOR	STUDENTS	TUTOR	STUDENTS
Publish an SWF file of under 250k and upload it to Studynet	4	3.95	4	4.00
Has correct number of screens with correct headings on each	4	4.00	4	3.92
Appropriate choice of screen colour – providing good contrast	4	2.15	4	3.21
Correct number of buttons with good colour selection	4	2.42	4	3.26
All buttons navigate correctly	4	4.00	4	4.00

Contains at least two images of you	2	3.67	4	3.62
Good spelling and use of language	4	2.95	4	3.53
An animation in the welcome screen	4	3.76	4	3.67
Aligned and uniform sized buttons	2	2.87	4	3.85
Text content is relevant and expressive and compact	4	3.19	5	4.37
Buttons have appropriate sounds on click event	5	1.44	5	4.90
Images of self, show high production values	2	3.37	3	3.57
Text and image presented well on the screen	3	2.81	5	4.11
Animation demonstrates originality and visual appeal	3	3.69	7	6.46
Background music is appropriate and is controllable by user	4	5.79	10	8.74
The CV is suitable for being viewed by someone who might give you a job	6	4.51	9	8.73
The CV's design expresses a kind of brand identity of you as a person	6	6.29	9	7.74

While the scores given are comparable the correlation between the tutor marking and the student marking is low for cv4 (0.32) but high for cv2 (0.98). However, if the four criteria most susceptible to variations owing to the anonymisation and also the nature of public performance are removed, namely:

- Publish an SWF file of under 250k and upload it to Studynet;
- Buttons have appropriate sounds on click event;
- Images of self, show high production values;
- Text and image presented well on the screen;

There is now an acceptable correlation for cv4 (0.50) and a very high one for cv2 (0.99).

So far, a largely comparable marking behaviour between the students and the tutor for these artefacts has been found, but the number of students marking the second artefact (cv2) significantly decreased during the session.

7.5.3 Levels of Participation and Agreement in the Final Session

In the final event, there were many more students (181). Again, there was greater student engagement during the first evaluation (cv1.swf) compared to the second (cv3.swf). In the graphic to the right, the blue line represents the percentage of students voting per criterion on cv1.swf, whilst the orange line is the percentage voting per criterion on cv3.swf. In addition to the visualisation of the percentages of students voting, the percentage of those voting who had attended the rehearsal event compared to those who had not can also be considered. (Blue represents the first artefact and orange the second).

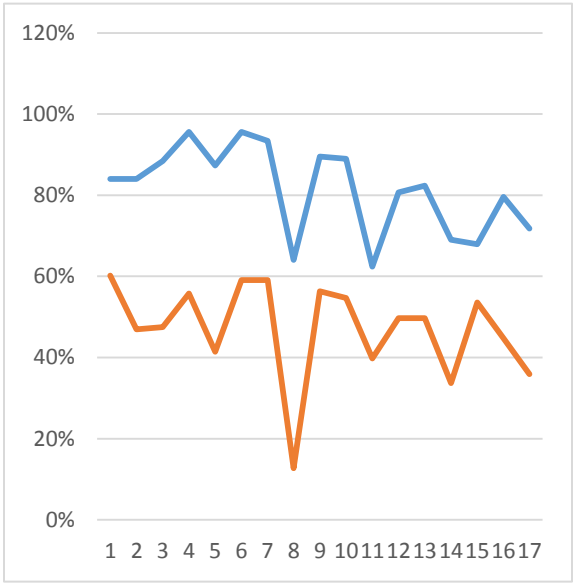


Figure 7-9:Participation by Artefact (line colour) and Criteria during Final Event 2013-14

At this point it might be worth asking what about the level of participation in the final event of those who had already participated in the rehearsal event.

In this graph below, it can be seen that the two major patterns representing the two CVs (orange and red line = cv1.swf, the light blue and blue line=cv3.swf). In each case, the lighter colour shows those students who *had* attended the rehearsal session, and the darker colour shows those who had not. Clearly those who had not attended the rehearsal session in each case were less likely to vote than those who had, but more significant than this is the trailing off of student interest for the second demonstration for both groups of students (those who had attended the rehearsal and those who had not).

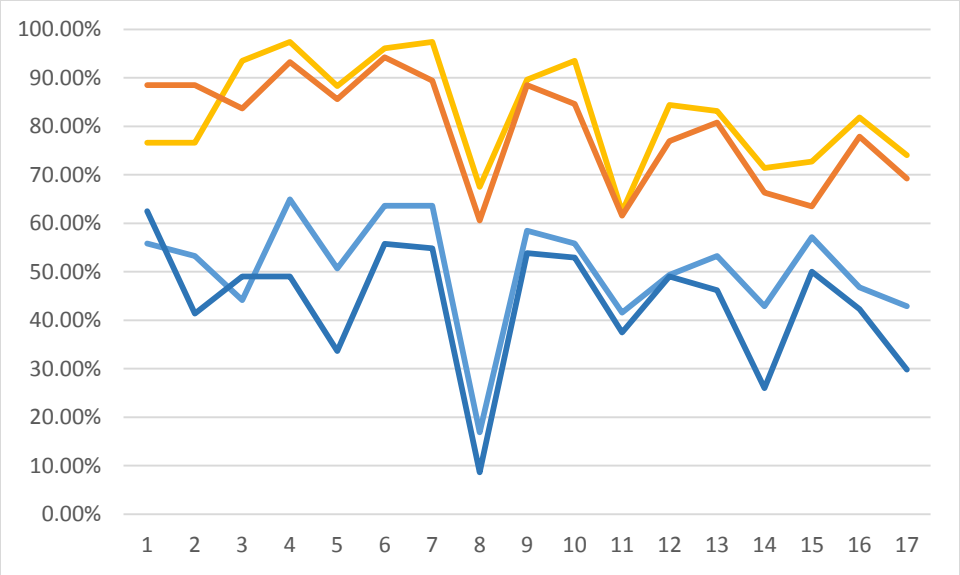


Figure 7-10:Participation by Artefact and Attendees (line colour + shade) and Criteria during the Final Event 2013-14.

The lighter colour here represents those students who had attended a rehearsal session, whilst the darker colour represents those who had not: orange and red (1st artefact), light blue and blue (2nd artefact)

When considering how people voted in the final session, a more interesting data-set emerges for two reasons. There was a larger number of participants (181) and is possible to assess whether having attended the rehearsal session had any effect on how they participated in this session.

Earlier, a number of caveats about using the correlation between tutor awarded scores and student awarded scores (the primary measure when investigating the MSc course) were raised owing to the fact that a number of the lower-order criteria resulted in something near unanimity.

Table 7-12: Marking Pattern for Students who Attended Rehearsal vs Those Who Did Not.

	CV1.SWF				CV3.SWF			
Criterion	all	Attended rehearsal		tutor	all	Attended rehearsal		tutor
		Yes	No			Yes	No	
Publish an SWF file of under 250k and upload it to Studynet	3.92	3.93	3.91	4	2.29	2.28	2.29	4
Has correct number of screens with correct headings on each	3.86	3.78	3.90	4	3.34	3.47	3.22	4
Appropriate choice of screen colour – providing good contrast	2.61	2.59	2.62	4	2.48	2.49	2.48	4
Correct number of buttons with good colour selection	3.24	3.11	3.34	4	3.15	3.20	3.12	4
All buttons navigate correctly	3.91	3.84	3.96	4	3.84	3.86	3.81	4
Contains at least two images of you	3.80	3.76	3.84	4	3.74	3.76	3.72	4
Good spelling and use of language	2.89	2.88	2.90	5	2.92	2.86	2.97	4
An animation in the welcome screen	2.14	2.19	2.10	4	4.00	4.00	4.00	4
Aligned and uniform sized buttons	3.83	3.83	3.84	4	3.71	3.73	3.69	4
Text content is relevant and expressive and compact	3.70	3.74	3.67	3	3.50	3.40	3.58	4
Buttons have appropriate sounds on click event	4.20	4.04	4.32	5	4.68	4.49	4.84	5
Images of self, show high production values	3.05	3.03	3.06	3	3.74	3.50	3.92	3
Text and image presented well on the screen	3.07	3.13	3.04	3	3.13	3.15	3.13	4
Animation demonstrates originality and visual appeal	6.52	6.33	6.67	6	6.80	6.67	6.96	7
Background music is appropriate and is controllable by user	6.39	6.14	6.61	7	6.90	6.32	7.38	10
The CV is suitable for being viewed by someone who might give you a job	4.41	4.00	4.69	6	6.95	6.83	7.05	7
The CV's design expresses a kind of brand identity of you as a person	6.81	6.49	7.06	6	7.50	7.09	7.94	8

Table 7-13: Correlation with Tutor Marks: Students who Attended Rehearsal vs those who did not

	Those Who Attended Rehearsal	Those Who Did Not Attend Rehearsal	All Students
Artefact 1	0.79	0.74	0.77
Artefact 2	0.90	0.87	0.89

As can be seen, the effect on levels of correlation between those two groups of students (attendees of rehearsal vs non-attendees of rehearsal) is negligible. That is, there is no evidence that the students

voted any differently whether they attended the rehearsal event or not. The only slight predictive power that attendance at the rehearsal event has is to say those who did not attend were less likely to actually vote during the final event. Also, in terms of the final mark awarded, the differences between the two cohorts were very small. Those who did attend the rehearsal gave an average mark over both artefacts at 71.66, whilst those who did averaged 72.4. There are two possible explanations of this. Firstly, that the rehearsal has no effect on voting in the subsequent event. However, this would be contrary to much of the literature on peer assessment, which attaches importance to training and inducting the students in the practices and values of peer assessment. Secondly, it could be that the very public way in which things are marked means that a kind of collective consciousness begins to operate and that students mark as a collective. This might take place in the following way. A student unable to attend the first event, attends the second, alongside his/her friends, some of whom did attend the first event some who did not. When the time comes to making a judgement with their clickers, they choose an option, see their immediate neighbours' choices and then see the result projected. Thus, they can compare how they and their friends voted relative to the others in the hall.

7.5.4 Which Criteria Required the Most Time

In the 2011-12 and 2013-14 iterations, it has been explained how two different rubrics were used and in each case, timing differences between the different criteria were observed. Recall that the events run and the rubrics used were as follows.

Table 7-14:List of the Marking Events

Year	Type	N Attendees	N Artefact Evaluations	Marksheet
2011-12	Rehearsal	138	3	2011S (16 criteria)
2011-12	Final	214	3	2011S (16 criteria)
2012-13	Final	195	3	2012S (9 criteria)
2013-14	Rehearsal	102	2	2013S (18 criteria)
2013-14	Final	181	2	2013S (18 criteria)

The 2012 event also involved using a hastily rewritten rubric and so, is anomalous. In 2011, however, a broadly similar rubric was used for the rehearsal and final events (small changes in wording, but without substantive change in focus), which was also the same as that used in 2010.

In 20011, the four criteria that required extra time, both in rehearsal and in the final assignment were:

- Motion Tweening Of Position/Visibility in the welcome screen;
- Layout and positioning of buttons and text;

- Choice of material, text and tone appropriate for a CV (text and image);
- An immediately visible and functioning background music toggle.

In 2013, the criteria that were proving to take more time to assess (excluding the first), both in the rehearsal and the final version were:

- Background music is appropriate and is controllable by the user;
- Good spelling and use of language;
- The CV's design expresses a kind of brand identity of you as a person;
- Text content is relevant and expressive and compact;
- Has correct number of screens with correct headings on each;
- Images of self, show high production values;
- Animation demonstrates originality and visual appeal;
- Appropriate choice of screen colour – providing good contrast.

Comparing the headline figures of both the 2011-12 and the 2013-14 iterations (and the level of correlation between the time taken per criterion in a rehearsal compared to that per criterion in a final assignment), the following emerges.

Table 7-15:Comparison of Marking Times Between Events on Per Year Basis

	Sum of Time	Pearson Correlation*	Spearman*
2011 Rehearsal	225	0.63	0.57
2011 Final	512		
2013 Rehearsal	313	0.56	0.76
2013 Final	347		

**Pearson and Spearman values represent the correlation of time taken per criterion between both events. The Pearson, is the correlation between the absolute times per criterion across the two events, whilst the Spearman pertains to the correlation of the rankings of time taken per criterion in each event.*

What can be conclude from these timing results? Certainly, it is difficult to assert categorically that higher-order criteria always take longer than lower-order ones to evaluate. For, whilst there is evidence of this in 2013, this not strong in 2011. The reason for this is the complex nature of judgement by rubric,

involving two distinct and multidimensional variables, namely, (a) the artefact itself and (b) the form of words used in the rubric. Whilst it might be easy to award a mark out of 10 for a noun like "design", deciding which of five articulated attainment descriptors maps most closely to the artefact you are evaluating involves a more complex mental operation. Also, an artefact might inconsistently exemplify the attributes desired for a particular criterion. An otherwise well designed page might have an inappropriate font for the heading, at which point, unless the case of inconsistent attainment is explicitly articulated in the attainment descriptor, the evaluator has to make a judgement beyond the options provided for in the rubric.

There are four possible reasons why giving a mark for a particular criterion might be harder than for others, these being:

- *incommensurability* (the artefact not easily mappable to an attainment descriptor);
- *inconsistency* (the artefact both satisfying, whilst also not satisfying an attainment descriptor);
- *visibility* (a feature required that may be present but that requires more effort to see);
- *coverage* (a feature or level needing to be present across a large range of the artefact, and verified repeatedly as such).

However, beyond these cases it does appear that the common sense idea, that items requiring a more holistic and synthesised kind of judgement (or evaluating features by creativity rather than compliance) require more time to make a decision.

What is also clear, when considering the rankings of which criteria took the longest times, in many cases these are common across both the rehearsal and the final. In fact, a Spearman's rho correlation of those rankings in the final event of the 2013-2014 cohort, where there was no summative element potentially distorting the time taken to mark things, registers a high figure of 0.76, which indicates fairly clearly, that those criteria requiring the most time to arrive at a conclusion in the rehearsal event, were also those needing the most in the final event. Put more succinctly, it genuinely does seem that the higher-order criteria seem to require the most time for students to give a response. However, there are also criteria which, on whatever level they are viewed, require more time. For instance, the criterion "has correct number of screens with correct headings on each" – on the surface a lower-order criterion, merely involving checking for compliance, rather than creativity – requires that the presenter to scroll through all of the screens in the CV in order to allow the student spectators to count the number of

screens and to check the headings. Whilst "aligned and uniform sized buttons" (another compliance criterion) might only require the recollection of such buttons (or the lack of a recollection of unaligned and non-uniform size buttons).

Summing up so far, a very close similarity of the marking between tutors and students has been found. We can see that some criteria require more time than others, and these tend to be determined according the whether they pertain to achievement or compliance criterion and if the latter, the level of coverage required to come up with an answer. It has also emerged that there is a tailing off of student engagement when evaluating two artefacts under formative conditions; approximately 20% fewer students participating in evaluating the second artefact compared to the first. Moreover, it has been discovered that students who did not attend the rehearsal event did not vote in any way differently to those who did, apart from being less inclined to vote and moreover, this disinclination did not have any effect in terms of how similar their marking pattern was to the tutors, nor any impact on their eventual grade for this assignment.

7.6 Critical Success Factors

From the preceding statistical analysis, insights in the way social marking works has been provided. In 2011-12, the levels of correlation between the tutor and student marking improved between the rehearsal event and the final event, in a context where the latter's grading patterns were given academic credit according to their level of similarity with the former's marks. Greater deliberation and thought given to how they were grading the exemplars in the final session was also evidenced in the amount of time they took for each grading decision. However, from the 2013-14 iteration, it can be seen that precisely the same ultimate effects, namely, the improvement in the scores for students' own subsequent artefacts occurred, even when a less deliberate and more desultory participation took place. The improvement in 2013-14 turned out to be the greatest of al and yet, in the final marking event, because there was no academic credit, a number of students present in the hall just did not vote. Moreover, there was no real change in the amount of time for those students who did participate to make their grading decisions.

One might argue that the different rubrics used could have played a part in this subsequent improvement – however, there was no change of rubric in the first three years (though the 3rd year allowed more granular attainment levels), and yet, the change between the 2009-10 academic year and the 2010-11 academic year, (when clickers and feed-forward was first used) was the most dramatic of all.

It could have been the case that the cohorts were of higher level on a year by year basis, however, the results of the first phase test of the year, in all the years of the course (except 2012-13) were roughly the same.

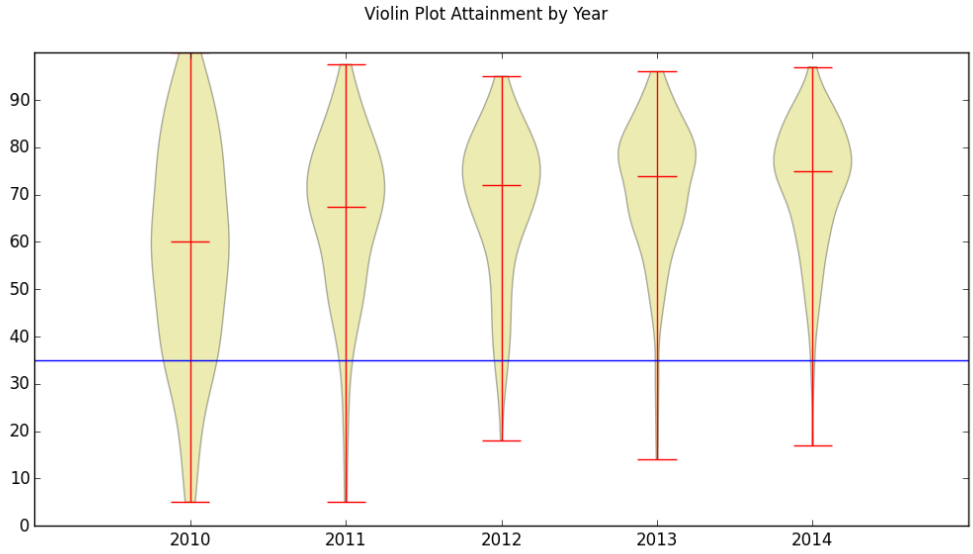


Figure 7-11:Violin plot of attainment levels over 5 years of the course for the final assignment

Improvements in the final iteration have been apparent, in the form of a better training set, fully non-credit bearing evaluation sessions and a more robust rubric, which led to a slight improvement in the marks compared to the previous years. One outcome appears to be that the mere provision of collective marking sessions, however imperfectly organised, enables students to appreciate the marking criteria and to internalise the standards expected of them in an effective way. Moreover, these are scalable, such that this process of appreciation and internalisation can take place in just two hourly sessions, with an audience, on this course, of a minimum of 180 students in the final summative sessions. Further, the subsidiary effects of better training sets and rubrics but the central experience of *applying* a rubric to a concrete piece of work (of whatever standard) seems to be most influential.

Compared to the other study, the results of LPA applied to this course are much clearer: we have a demonstrable 13% gain in student scores for a single assignment over a 4 year period. However, the insights yielded up by the voting data is messier and less transparent. What we can say in this regard is:

- Some form of mandatory participation is probably necessary to make sure that LPA benefits all students
- Some form of light requirement (e.g. positive correlation with tutors' marking) is needed to ensure clicking at a feed-forward event takes place
- A hybrid-assessment (part objective – summative, part subjective -formative) might achieve the goal of mandatory participation
- However, even in situations where students do not participate, the actual event may be sufficiently impactful to demonstrably influence the way students undertake the assignment

As has been said, these evaluation events took only 1 hour each, and yet participation in them, however desultory or engaged, seemed to exert a very strong influence over the conduct of the development of the subsequent multimedia CV. In the next chapter where I look at the internalisation of standards of quality, we will attempt to work out where that influence was exerted as students completed their final assignment, and what students may have done differently as a result.

Chapter 8. Internalisation

In the previous chapters, we have seen how the scores for the final assignment improved over four iterations of a course, where the technique of collective marking events took place. In addition, a substantial amount of statistical data around the collective marking events has been presented. The statistics for the improvement in the final assignment have shown very large improvements in the lower decile achievers, as well as some evidence of improvement for higher ones, such that the number of students getting into the first class area (in so far as that term can be on a first year course) also increased. This may be related to the more ambitious requirements of the rubric used on the final two years of the course.

However, the meaning of the various voting data we have looked at is more suggestive and not as precise. When participation in the voting events was mandatory and involved academic credit, the degree of engagement, as measured by the students who actually clicked answers when the criteria were presented in evaluation of a multimedia CV, was much higher. However, in the last year when this method was used, when academic credit for participation in the grading events no longer applied, engagement was more intermittent and variable. Nonetheless, that was the year the overall highest average score for the multimedia CV assignment was registered. This was against a baseline of a multiple choice test in the 6th week of the course (across the three years of the course where Questionmark Perception was used), which showed no improvement over the iterations of the course, thus indicating that attributing the improvement in performance in the multimedia CV to an improved cohort was not supported by the evidence.

The data regarding time taken to answer each evaluative criterion was broadly in line with what should be expected. That is, those criteria involving the greatest subjectivity and evaluating the greatest synthesis of features required more time (except for those where overall coverage over the whole application was emphasised). Generally, a good degree of inter-rater reliability was found, with scores very often above 0.6 on the Kalpha scale. It also emerged that participation in the rehearsal events did not really predict evaluation patterns in the final events, beyond a very marginally improved level of engagement. This meant either that the evaluation was straightforward or that there was some kind of social conformity pressure causing participants to rate in a highly similar way. There certainly was no effect in terms of marking patterns established between those who attended the final rehearsal event, and those who did not.

As a consequence, it would seem reasonable to conclude that these marking events either established a shared understanding of what constituted quality in the assignment particularly what each element in the marksheet meant, or potentially more dynamically, established some kind of shared framework, which structured subsequent practice and conversations about the assignment among students. If the claim is made that these collective marking events caused the students to work differently to before, it is necessary to find out from the students whether that was the case. As a result, a focus group was organized to try to find out how students did go about completing their final assignment. However, before considering the perspectives of the focus group itself, some of the theory about how professionals go about complex tasks (such as one involving the creation of a multimedia CV) is discussed. This is in order to get a conceptual frame around what happens when someone undertakes a task, such as in this case, when the students carried out their CV assignment.

8.1 Reflective Practice and Communities of Practice

It is probably Donald Schön in his books about reflective practice and particularly, the concept of "reflection-in-action", who has provided the most compelling theoretical framework in which how professionals go about their work is explained. In his writing, he uses Geoffrey Vickers' concept of an "appreciative system" to represent the inner intuitive *value set* through which professionals decide the value of any action they undertake. He regards one of the features of a professional to be the "appreciative systems they bring to problem setting, to the evaluation of inquiry, and to reflective conversation". He uses the term "move" to describe the experimental action that the professional may imagine in the mind or on paper or in actual deed, to resolve a particular problem as they progress through it and refers to the terms "framing" and "reframing" to describe how the professional contextualises the result of any "move". Regarding which he states:

> the practitioner's moves also produce unintended changes which give the situation new meanings. The situation talks back, the practitioner listens, and as he appreciates what he hears, he reframes the situation once again. (Schön, 1983)

This concept of a conversation between the "situation" and the "practitioner" can be applied to the multimedia CV assignment, for here also, new elements were added, causing unintended changes, which then required reframing by the student.

However, this cycle of "move" and "reflection" and the parameters in which it takes place, is not some innate capability of human beings, but rather, is the result of some formal or informal process of

apprenticeship within a community of practice. While Schön (1983)writes in great detail about how the values and appreciative system of a profession can be transmitted in one on one relationships of experienced and less experienced professionals, the writing of Etienne Wenger shows how this is more a process of community enculturation, rather than one-on-one transmission. Wenger (1999) talks particularly regarding how a community establishes the vocabulary with which it understands problems, a kind of collective memory of informally held case-studies, which form points of reference with which to understand novel problems. This is comparable with the "repertoire of similar situations", which in Schön's understanding of professional practice, the individual brings to any situation that he or she is working on.

More recently, D Royce Sadler has applied this kind of framework to the understanding of academic quality, in two papers: "Making Competent Judgements of Competence" (Sadler, 2013a) and "Opening up Feedback" (Sadler, 2013b). In these papers, he describes how a skilled practitioner in the process of working on something will "sense" problems, perhaps without even being able to verbalise them, and then seek remedies or "tactics from their repertoires to change them". He also acknowledges a social dimension to this saying that "competent judgements" in any such situation are something that will be shared by a guild of like-purposed professionals, writing:

> "For a given set of phenomena or objects, the meaning and significance of evidence are shared, as is what is deemed to count as evidence. In short, given the same stimuli, the people making the judgments would react or respond similarly and judge similarly" (Sadler, 2013a)

8.2 Applying a Reflective Practice Framework to the Multimedia CV

This high level abstract description can be applied to students undertaking their interactive CV, with the given set of phenomena or objects being: screens, text, buttons, backgrounds and music toggles. There are judgements students make about their interaction, in which some understanding about "what counts as evidence" needs to be marshalled in making them. It is towards these minute, everyday decisions, framings and reframings, management of unintended consequences of the introduction of new elements, judgements on what is lacking and what is fine – it is there that ultimately the rise in student achievement will be enacted.

In more detail, there is the CV assignment with its rubric and subsequently, there is the activity that involves the creation of the CV with the five screens, involving decisions about colour, layout, text content, the provision of background music toggle, the need to keep the file size down to a minimum and finally, the necessity to design it in such a way as to permit rapid redesign in the final 50 minute test. At each point there are decisions to make, or in Schön's terminology the "moves", including: how to design the background music toggle to make it in keeping with the other buttons, despite the function being very different and no precise wording being recommended; and how to make the title "Hobbies and Interests" fit into a button shape, even though its text is longer than that of the other buttons, without resizing the font which would then make the text inconsistent on the page. Attempting either of these things will then produce a new "situation", which needs to be evaluated, and might impact on other aspects (as per the example here, reducing the size of a font to fit in a button shape will then have effects on the appearance other buttons). This situation might need to be reframed, e.g rather than put in the whole title "Hobbies and Interests" maybe just use "Interests" instead, which would potentially violate the assignment spec, but for the greater good of the overall harmony of the page.

In pursuing these "moves", the appreciative system might alter. In taking a particular approach to a design decision, this might cause one to call into question one's previous assumptions. Thus, the "appreciative system" is not some kind of fixed world-view, which subsequently interacts with world and judges it on a case by case basis, but rather a dialectical relationship, a conversation between the creator of something and the thing he or she is trying to create. That is to say, an ability to appreciate a situation, make a "move", that is to say do something in it, or sketch out something in it, evaluate the result (and any of its unintended consequences and ulterior effects) and decide whether it meets the terms of the appreciative system. However, the appreciative system can also change through the moves that are made and the results they engender, this appreciative system itself can be interrogated.

8.3 Focus Group

The focus group was run approximately two months after the end of the course. Three students participated, two male and one female, with each being given a £10 Amazon gift voucher for participation. All students were UK nationals and went direct to university after secondary education. This focus group was structured around the ongoing practice of producing a multimedia CV and the impact the evaluation sessions had on it. The focus group began with typical ice-breakers comparing

sixth form to university education and then went on to discuss the evaluation sessions and the work students had completed. The three students who participated came 40th, 58th and 100th in a group of 186 students, representing the 79th, 69th and and 47th percentiles.

8.3.1 The Centrality of the Rubric

Throughout the discussion, a clear theme was how central the rubric was to evaluating their practice. In the first exchange, where we asked the students whether the evaluation sessions helped them, the following exchange occurred:

> Tutor 1: Do you think that influenced you in any way?
> Student 1: It helped us to think *what we were supposed to look out for* in the actual test so it was really helpful.
> Tutor 1: So, what is interesting is that it told you to know what to look out for and so when you did come to do your own CVs what were the things that you were primarily looking out for?
> Student 3: The layouts, colours, performance, sizes, are they readable? Or they shouldn't be too bright colours, so that other people can't read it and the background colour has to match well with the text.
> Tutor 1: Okay, interesting. [name] you wanted to say something?
> Student 2: *The sheet that you get basically gave us a list and markings of what you guys expected*, *so it kind of told us exactly what to do*.
> Tutor 1: If it told you exactly what to do, while you were writing out your CV you were kind of ticking bits off?
> Student 2: Well, what I did was I read through all of it, then did my CV, then read through it all again and checked everything and compared it with my CV and then changed things around it.
> Tutor 1: That is really interesting. Did you two do the same?
> Student 3: Yes, pretty much. I just followed the requirements sheet.
> (My emphases)

This is interesting in that the students talked, to some extent mechanistically, about just following the rubric, but at the same time they said how they realised that "the background has to match more with the text". Whilst there is requirement of good colour contrast in the rubric, there is no explicit requirement to "match" colours. Nonetheless, the students extrapolated a holistic criterion from a set of more specific ones. Student 1, when talking about completing the assignment in the final 50 minute

test (where students have to resize their CV to new arbitrary dimensions) said “I took about 20 to 30 minutes and most of that time was mainly to do with checking the list; ticking off everything”.

In another exchange, the students were asked if they were even able to predict their marks based on their compliance with the rubric:

> Tutor 1: When you were looking at the brief and you were ticking things off, did you kind of work out what your mark might be?
> Student 3: Definitely.
> Tutor 1: You did? And do you think you got roughly what you thought?
> Student 3: More or less.
> Tutor 1: What about you?
> Student 2: Yes.

One student talked about the marks for the alignment of fonts:

> Student 1: I am not too sure if I remember, but maybe on the brief it said make sure fonts are aligned and then - as we were marking or did you see that there is a mark as well, you don’t want to lose that mark, that is pretty simple. So, I am sure everyone would try and do it to make sure that it is correct and then submit it,

When asked about what problems students had managed to solve during the making of the CV and the same student said:

> Student 1: I would say my animations, because at first I thought it was just text, but I was thinking about it, because when I looked at it at home it was fine, but when I got in and I went through the brief again, I started having doubts again, thinking, is that really acceptable? It is quite a cheap animation just having it fade in, but in… anyone would love fade out I am sure, it feels like a book, you turn the pages and the text just comes in, so I thought that should be okay.

The degree to which the marks sheet structured students’ evaluation of their own work appears quite unusual, compared to other assignments during my career, where students sometimes appear to have had very little sense that the rubric is a meaningful document. It seems the fact that two sessions took place where the students completely marked work according to a rubric raised its importance in the eyes of the students.

> Student 1: It is when you look up at the marks sheet and what you learn in the lectures, they say the same thing. So, when you attend the lecture and you do that with the clickers then you know that immediately when you get to those sheets that that is what the requirements are.

It seems, therefore, that one big impact of collective marking is to raise the prestige of the rubric as an authoritative guide to how marks are obtained. By students marking using the same rubric as the teacher, its value and meaningfulness is affirmed. This relates very strongly to the duality of reification and participation described by Etienne Wenger in "Communities of Practice"

> "On the one hand, we engage directly in activities, conversations, reflections, and other forms of personal participation in social life. On the other hand, we produce physical and conceptual artefacts—words, tools, concepts, methods, stories, documents, links to resources, and other forms of reification that reflect our shared experience and around which we organize our participation. (Literally, reification means "making into an object."). Meaningful learning in social contexts requires both participation and reification to be in interplay. Artefacts without participation do not carry their own meaning; and participation without artefacts is fleeting, unanchored, and uncoordinated. But participation and reification are not locked into each other. At each moment of engagement in the world, we bring them together anew to negotiate and renegotiate the meaning of our experience."(Wenger, 1999)

Thinking of how we might translate this into understanding of the practices of the course and education generally, it is contended that a rubric, without any understanding of its application, will remain a remote and distant object, making such rubrics unable to "carry their own meaning". That is, towards the popular understanding of the word "reified" – namely reduced to an object, made lifeless. In the case of producing a multimedia CV, without some explicit exemplars together with statements of quality (as one would find in rubrics), this would make the activity "fleeting, unanchored and uncoordinated". Hence, the application of the rubric in the marking events ensures that its structuring power is present after the event has concluded. And that structuring power is essential for the ongoing decisions that students make as they produce their multimedia CVs to be effective.

8.3.2 The Impact of the Collective Marking Experience

It has been shown how the alignment between the rubric used for marking the final assignment and the experience of undertaking the marking collectively meant greater concentration on that rubric. Now, the focus turns to the impact of the collective marking experience itself.

The first and most obvious point is the greater attention students paid when there was the use of clickers.

> Student 2: I think the clickers forced you to listen, because then if you don't listen then you miss the question and then you won't answer it so the clickers, whereas most people would fall asleep in that lecture or go on Facebook or whatever. When you have the clickers there everyone starts paying attention.

Clearly, from the statistics about participation in the marking events for which no summative credit was given, a number of students probably did "fall asleep" or "go on Facebook". However, as this student says, it nonetheless, among students who did participate, provided a more compelling experience.

In a number of cases, the students experienced recognition of levels of quality when looking through the CVs. One student said of the exemplars: "I think each one of those has always had some major problem that stood out and pretty much everyone noticed it". However, being asked to apply a rubric meant that not everyone agreed

> Student 1: I would say I agree with most of the opinions that we said about the CVs, because certain CVs you could tell that they're rubbish and certain ones you could tell that they are good, but then when it comes to saying that you strongly agree with or do you agree with …[inaudible], everyone will just have different taste and so someone might say this bottle of coke may look nice, but I might say well this bottle of coke looks okay. It depends on personal preference.

However, even in the cases of clear and agreed recognition, students appreciated the closer analysis given by the tutors in the marking sessions.

> Student 2: I think we could have all just looked at the CVs and gone that is a good one, that is a bad one, but when you guys went through it and explained why that is a good one and why that is a bad one and pointed out the bad things and the good things, then everyone got a better idea of what you expected of us. So like you said, the basic problems didn't come up because you guys told us what the basic problems were, so we all avoided them.

One student was in fact dissuaded from making the more ambitious animation that they had originally intended from the experience of the collective marking.

> Student 2: I didn't want to risk it too much, because some of the example CVs that you guys showed, you said that some of the animation was too much or maybe not suited to the CV, so I didn't want to risk it, I wanted to play it safe.

Clearly, here, the actual practice of marking in the lecture theatre is being remembered during the decision making around whether to add an animation or not. That is, the example of the overblown animation that was shown becomes part of a "personal repertoire" that informs judgement. Much of the experience in the lecture theatre relates to what Royce-Sadler in the two above mentioned recent articles (Sadler, 2013a, 2013b) has referred to as "noticings" or "knowing-to". Regarding which, the professional in carrying out their work, at any point is capable of recognising salient problems, that is to say, capable of "noticing" and subsequently "knowing-to" do something about it. As quoted above, one student said the major benefit of the rehearsal sessions was that it "helped us to think what we were supposed to look out for in the actual test so really helpful". This appears to be the value of these sessions: the "pointings out" and what to "look out for". Moreover, it begins to establish a "repertoire" in the students' minds of similar situations, or similar tasks which can be used to guide practice.

8.3.3 Solving Problems

During the focus group, the students were asked to describe in detail some of the decisions they took in the development of their multimedia CV and how they set about realising their goals. What was interesting about the discussion that took place was not only how clearly they related their practice to the requirements of the assignment, but also, the amount of self-knowledge they brought to the table. Having seen what other students had achieved previously they then applied it to their own situation, deciding what was capable of emulation and what was not. For instance, the student who decided "not to risk it" with the animation, described their own animation as:

> Student 2: The shape of the sheet, that guide thing, said that you should have animation on your front page and text coming in is animation so I just kind of went with that.

Another student, however, during the 50 minute session at the end decided their animation was insufficient:

> Student 1: I created an animation, but when I got into the class I looked at it again, I felt like it is not really finalised.

And so they decided to change it in that 50 minute session.

> Student 1: I had lots of practice before, when I was at home, so when I got into the class I knew what to change, but because it didn't feel right to me when I bring it to the class I had to just quickly change it.

In both cases, there is the use of the word "feel": "I felt like it is not really finalised" and "it didn't feel right". This is I believe is exactly what Sadler means when he uses the term "knowing to":

> "It constitutes a distinct form of knowing, 'knowing to', which involves detecting, 'sensing' without necessarily using the standard five senses. It begins with a feeling of unease or discomfort that something is not as it could be, and that a change would improve correctness, efficiency, flow or elegance."

At this point it is worth quoting this student's alterations to their animation in greater length.

> Tutor1: I see. What would you say were the major problems you did solve in doing the CV in – this is not so much in the 50 minute session, but in the preparations?
>
> Student1: I would say my animations, because at first I thought it was just text, but I was thinking, about it because when I looked at it at home it was fine, but when I got in and I went through the brief again I started making doubts again, thinking, is that really acceptable? It is quite a cheap animation just having it fade in but… anyone would love fade out I am sure. It feels like a book, you turn the pages and the text just comes in, so I thought that should be okay.
>
> Tutor1: It was a good, it was very nice, each of those fades when you click on it, that was really – the timing on it was great and I thought it was great. Sorry, I only briefly looked at your CV beforehand, when you talked about the positioning thing, what was it that was the positioning? What was the animation? You said about the positioning of things being near the text and at the end with the animation?
>
> Student1: That was the one with the sword that actually goes through.
>
> Tutor1: Oh that, I remember, yes.
>
> Student1: When I was designing it I was thinking should I put it closed without a tray, put it sort of closer to the text and then when I actually played the animation it didn't look right, because from when the sword starts until it finishes, the starting point was okay, but where it ended it

> looked off. So, I had to just adjust it make it go up, to create more space and the sword and the handle is more closer to the text and the blade was closer to the text.

This is a fascinating exchange because it brings together a number of the themes that have already established, namely: the centrality of the rubric, which was re-consulted during the 50 minute session at the end when the student asked "is that really acceptable". Then, there comes the addition of the sword animation, but again structured through a quite sophisticated sense of "knowing to" – starting with the initial feelings of dissatisfaction "it didn't look right" and "where it ended it looked off" – and ultimately the solution, "I had to adjust it to make it go up". Clearly this is not expressed optimally – but that is precisely the point of "knowing-to" –it doesn't need to be. In Sadler's formulation it is almost pre-verbal:

> "This type of knowledge cannot necessarily be made explicit, that is, expressed in words. It nevertheless exists, even when a person cannot define it in concrete terms or otherwise explain it."

In that session, the student solved the problem they posed to themselves, however subsequently inelegantly recounted. This raises one of the paradoxes of the use of collective marking events on this course. The original goal was to get students to perform "higher order thinking", however, when student practice is considered and how they themselves made sense of their practice, it is not so much "thinking" as "feeling" which seemed to be in operation. A more sensitive antenna to issues and quality in their work and a more resourceful power of "knowing-to" in order to resolve them.

The other two students in the focus group also described their own attempts at experimentation. Another exchange with another student is worth quoting at length. It begins with the moderator asking the students how they made their colour choices:

> Student 3: Well, I am not really arty, so I just kept changing colours for the background, because that was a starting point and I just picked the colour that I liked, but then I moved on to the buttons and to the box where you put text and information and I just tried to fit that to the background colour and that was it, except maybe those boarder lines, they were the same thing really – just tried to match it with the background.
>
> Tutor 1: Was there any reason for choosing those boarder lines? Where did you get that idea from?

> Student 3: I just – it was actually just random and accidental… I just… it was just something new and I just kept discovering it. I think it was an option and I found it and I just chose it randomly. I think there…
>
> Tutor 2: You got a feeling of like, you just liked it?
>
> Student 3: Yes, because I am not a very arty person and I am not into art at all.

As before, we see the student bring some self-knowledge into their decision-making and also a kind of instinctual processing of the design decisions (colour matching and line thickness) along with some personal understanding: "I am not really arty" and "I am not into art at all". The process this student described again encodes some real "knowing-to": "I just kept changing colours for the background, because that was a starting point". Then after this starting point, he describes a sequence of moves and their consequences.

Another area in which the students in this focus group demonstrated an extended sense of "knowing-to" was in the time planning of work. That is, they recognised when they had reached an impasse where they would need help from others. For instance:

> Student 3: I used to skip certain bits and come back to them eventually but I tend to do the easiest bit or the bits that I am not struggling with.
>
> Tutor 1: So, you did the easiest bits and sort of ticked them off and then, you came back to the harder bits later. What were the harder bits?
>
> Student 2: The ones that involved more – like putting a background colour in grey is easy. It just involves doing something in stuff but making things move around or something that involves putting more work into it, so you would come back to that.

During the focus group, the students talked about receiving and giving help. One student said "We had a couple of people on Facebook asking questions in a group chat". When asked about what was discussed in this Facebook group, the student said:

> Student 1: People were asking questions. They were asking more to do with animation and sound, less than the colours, because they were pretty easy, because you just get a pallet and you just pick whichever one. It's trial and error really; you don't like a colour or it doesn't match, you just change it and that is pretty easy.

Another student said:

> Student 2: I think there was a Facebook group but I didn't use it. I had to ask a friend for help on one thing and my friend asked me on one thing too, so I guess we just helped each other.
>
> Tutor 1: So, what were those things?
>
> Student 2: My friend asked me if I could help her on the sound, because they couldn't get the sound buttons to pick up the music and I had forgotten how to make music like a package - shorter.
>
> Tutor 1: How to select the full bars?
>
> Student: Yes. So, I needed someone to show me how to do it.

One student found a problem with controlling animations

> Student 3: I think it was mostly animation related. I was quite fine with everyone else and a bit of coding to do with buttons… when you add information in animation and the buttons just scope on and then, the players just keep on playing everything so there isn't a stop bit or whatever it is called. I didn't struggle with anything else.

This comment is interesting since it highlights a technical problem that needs to be solved and is, unlike many of the other decisions described in this chapter, a self-contained one, namely, how to make the animation stop. This, interestingly, is also the student who said he was "not arty at all". In this focus group, there remained among these two participants, the desire to separate the technical and "arty". This may be a feature of the discipline of computer science and the construction of its culture by first year students. This student said:

> Student 3: because it is what is being assessed, your flash skills - because not everyone is creative.

Another student when asked what would be their recommendation for future students undertaking the course, said:

> Student 2: Just to follow the guidelines that you give and try and show some Flash skills really. I understand that creativity is probably a big part of it, but like you guys say, not everyone is creative so the best thing you can do is just try and show off your skill.

Clearly, in past years, this demarcation of the "arty" and the "technical" had negative consequences when students added plainly irrelevant and unrelated content to their artefacts merely to demonstrate the mastery of some skill. Here, in this final year of the course, I believe that through the rubric and its marking events, the students obtained a holistic view of quality, where things inter-related, and where the effects of novel elements were more keenly examined regarding how they affected other elements. Consequently, more unified and coherent artefacts were delivered as a result.

8.4 Conclusion

At this point, what the effects of the collective marking feed-forward events are have become somewhat clearer. It is evident that these are more than just sudden moments of illumination, for it has been shown, they appear to structure subsequent student practice and thinking a long time after the events. Moreover, this is not just about individual practice and thinking: it also structures the kinds of conversations students have when seeking help, firstly, in terms of knowing what help to ask for and subsequently, regarding of the moves they take and tactics that they adopt. Schematically, from the evidence of this focus group, I would put the effects of these events under three headings:

1. Raising the profile of the marking rubric
2. Establishing a repertoire of examples
3. Modelling the "noticing" of salient features

By using collective marking, where the students follow a rubric largely the same as that of the tutor, they both see how marks are obtained, in an authoritative way as well seeing the rubric as a living document, whose requirements can indeed be ticked off. The rubric is no longer the "small print" or the "ts & cs" of the course, but is rather the core document used in the marking of their work. Moreover, the rubric necessarily is expressed in abstract terms, which to the student, must of necessity prove difficult to grasp, in so far as it encodes tutor-understood characterisations of quality, speaking to a shared experience of marking. However, by giving concrete instances of abstract quality descriptors, those descriptors become more meaningful and tangible to the students.

As Schön and Sadler have pointed out, the activity of the competent professional is achieved through being able to take an individual situation and connecting it to a range of prior mentally catalogued case studies; classifying it into ranges of comparable situations with comparable trade-offs and dilemmas. This canon of prior cases, with their associated tactics and moves, constitutes the framework with which the professional understands a novel situation. Of course, what the novice lacks is precisely this

framework of prior cases and therefore, by exposing students to previous work, the tutor is beginning to "seed" this repertoire. But more than that, it is being seeded by work that can be said to have some exemplary value: a very fine piece of work, an uneven piece of work or a poor piece of work and therefore, these exemplars are contextualised as representing classes of quality. Consequently, despite the repertoire being small, at least the examples in it are given the authority of the tutors, and moreover, their levels are being communicated.

In addition, tutors in presenting such exemplars, are isolating parts of them for either praise or censure, or maybe for commenting on their relationships and in so doing, are relating how "features" of the artefact relate to its overall quality. They can point out anomalies, they can begin to articulate in words how an overall feeling of dissatisfaction with an artefact can be traced down to the individual collection of features which constitute it. Sadler, recommending such evaluation events, writes:

> Much more than we give credit for, students can recognize, or learn to recognize, both big picture quality and individual features that contribute to or detract from it. They can decompose judgements and provide (generally) sound reasons for them. That is the foundation platform for learning from an assessment event, not the assumption that students learn best from being told. They need to learn to discover what quality looks and feels like situationally. They need to understand what constitutes quality generally, and specifically for particular works. Equally, students need to be able to detect aspects that affect overall quality, whether large or small, and understand how and why they interact.

A fourth element, which though not explicitly mentioned during this focus group, but was mentioned in the masters group, was the influence of other students on students' own judgements during the marking events. This was found to exert a strong influence among peer graders worried they had been too harsh or too lenient on other students (see chapter 4). Moreover, this was witnessed on the BSc course through the evidence of there being frequent homogeneity among markers during the marking events over the course. It seems part of the persuasive power of the marking events is this ability of individual students to compare their own marking decisions with those more broadly made by the class as a whole.

One final element that did not come out during this focus group which in some ways transcends all of what was discussed during it is "inspiration". Typically, in any field, when being presented with high quality work created by practitioners in the same field, there comes the desire to emulate what is being

presented or in some way, take what is being presented and incorporate it into what one is doing oneself. Only one student in the focus group used the word “inspiration”, but when doing so, did not relate it to the work demonstrated by fellow students, but instead looked across the World Wide Web, saying “there is loads of inspiration around, a lot of people on the web so you can research it before you go in”. However, it may be possible to trace the existence of inspiration by seeing if any of the features in the exemplars presented to the students in 2013-14 also began to figure in the artefacts developed by the cohort that year, which is the focus of the next chapter.

Chapter 9. Inspiration

Up to this point, the artefacts delivered after the experience of collective feed-forward evaluation have been probed in a black box way. The inputs (rubrics, marking events), some of the behavior in marking events (time taken, level of correlation with tutors) and the improvement in the final grade for the multimedia CV assignment over the years have been investigated. The improvement in that score has been attributed to greater explicit compliance with the assignment criteria and a greater understanding of them. However, engagement with the marking criteria does not mean merely a greater familiarity with the published text of an assignment. Rather it appears to lead to a greater engagement with the concept of quality itself, in terms of what it consists of and the details of its manifestation, such that it affects the sensibility of the developer, enabling him or her to be able to routinely adjust what they are working on in line with this greater awareness.

Over the years when live peer assessment was being used on the course, in addition to seeing scores in the lowest percentiles of students increase by up to 20%, there was also an improvement at the 60th percentile, where the already good were getting even better. Hence, there may be something operating more akin to "inspiration" rather than merely better understanding of the assessment criteria. That is to say, not only a greater ability to sense and diagnose defective things, but also, a greater ambition to produce improved work.

Did the exemplars presented to the final cohort engender such an aspiration? Or put another way, can any connection be made between the exemplars presented in one iteration of the course, and any of the artefacts generated during that iteration? Do the exemplars presented influence the practice of those students who see and evaluate them? In this chapter, these matter are investigated by looking at the broad range of artefacts produced by the cohort. Since the exemplars used in the 2013-2014 session came from works produced in the 2012-13 session and moreover, represented the most authoritative exemplars yet produced, the aim is to elicit through a detailed examination whether there is any evidence of influence of the exemplars on the works produced by the 2013—2014 cohort.

9.1 Comparison of Use of Colour by 12-13 and 13-14 Cohorts

To begin the process of comparing the use of color across the multimedia CVs in these two cohorts, the Flash files produced during the 2012/13 and 2013/14 iterations were converted from that format to a single graphic, which enabled all the artefacts produced to be viewed side by side. The conversion was carried out using the jpex flash utility (Petřík, 2011-2017), which has a feature for capturing screens from

particular frames. Because there were over 400 conversions to take place, this was undertaken in an automated fashion (by choosing five frames, typically, 1, 10, 20, 30, 50) to obtain a representative screen in that artefact. Because the requirement for these artefacts involved an animation at the beginning, sometimes those frames would not produce a representative screen (they might for instance, just show a partial screen during the initial animation and before the full interface is established). In those cases, the same process was run and later screens were captured. For any remaining artefacts where a suitable capture was not able to be found, screen captures were obtained manually by running the files and then using an ALT PRINT SCREEN command. This resulted in a large number of files for viewing in thumbnail format.

The first thing to seek out was whether there was any influence of the exemplars on the colours chosen by students in the 2013-14 cohort. In order to do that, I used the "color extract" php tool (Gelotte, 2011), which is a php script that can find the five most common colours in an image. This script was run over all the graphics obtained from the method previously described and then saved as a .csv file of rgb values. The colours were then sorted according to the methods found in Alan Zucconi's blog on sorting colour (Zucconi, 2015). As he explains, sorting by colour is an extremely complex thing to do, because it has so many different dimensions. When considering the two cohorts I decided to present the colours used through three separate sorting techniques :1) by hsv values; 2) by ordering the colours in terms of their score when the red component is subtracted from the blue component (establishing an ordering essentially of from blue to red); and 3) using a formula given for greenness, as follows.

$$greenDifference = greenChannel - (redChannel + blueChannel)/2$$

This technique comes from a posting by John D'Errico on the Matlab site (D'Errico, 2014).

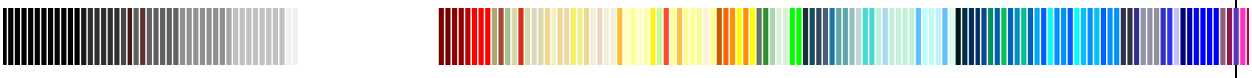


Figure 9-1:Spread of principal colors of 2012-13 artefacts sorted by hsv values

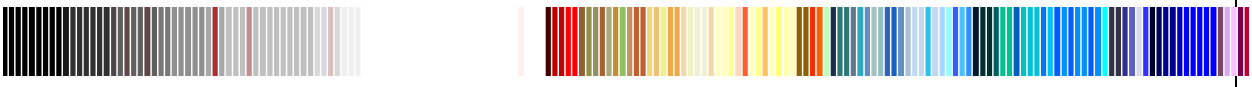


Figure 9-2:Spread of principal colors of 2013-14 artefacts sorted by hsv values

Figure 9-3:Spread of principal colors of 2012-13 artefacts sorted by red minus blue values

Figure 9-4:Spread of principal colors of 2013-14 artefacts sorted by red minus blue values

Figure 9-5:Spread of principal colors of 2012-13 artefacts sorted by greenness values

Figure 9-6:Spread of principal colors of 2013-14 artefacts sorted by greenness values

From these comparisons, it appears that during the 2013-14 iteration, a number of artefacts had a high contrast interface with either white or black to grey backgrounds. Certainly the amount of white background goes up a little. It is difficult to work out which colours appear less; potentially, there is less on the "greeny" side of the spectrum, but it is difficult to make a claim beyond that. However, aside from the greater preponderance of white and gray, there is actually a quite similar distribution of colours between the cohorts.

The six principal colors in the exemplars shown were the following:

Figure 9-7:6 Principal Colours in the Exemplars

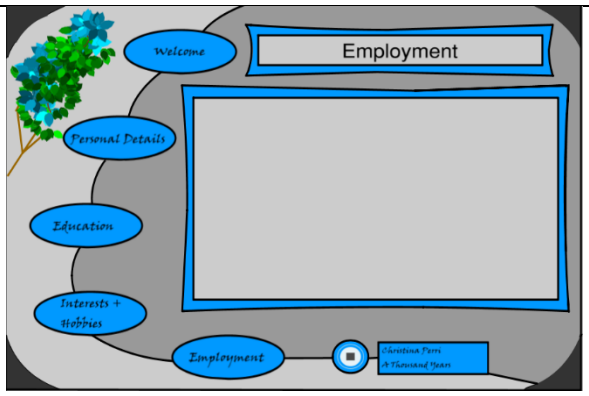
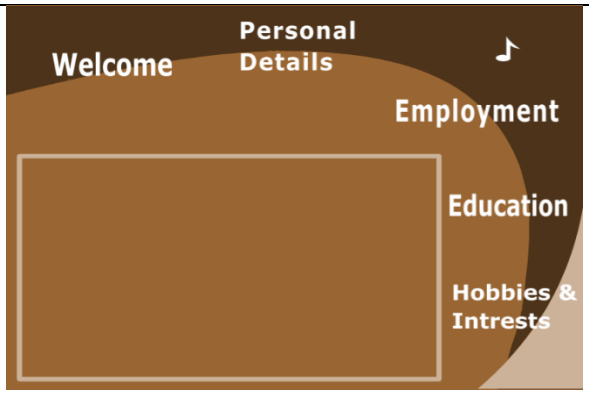
Which, suggests that the focal exemplars had little influence on the colours chosen by the 2013-14 cohort.

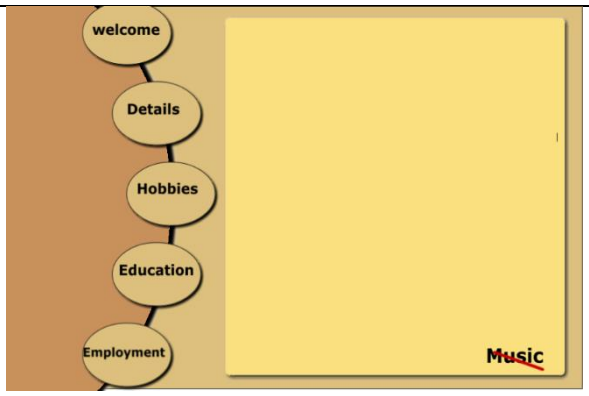
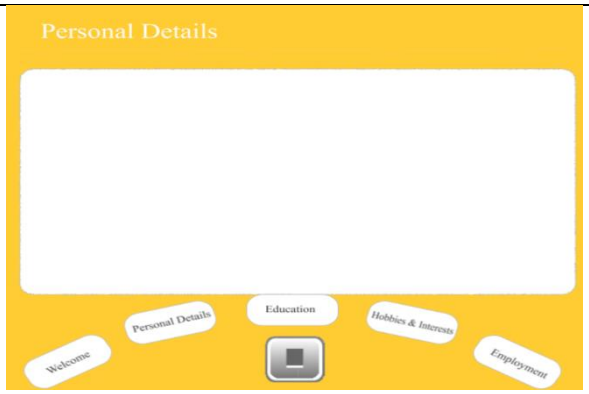
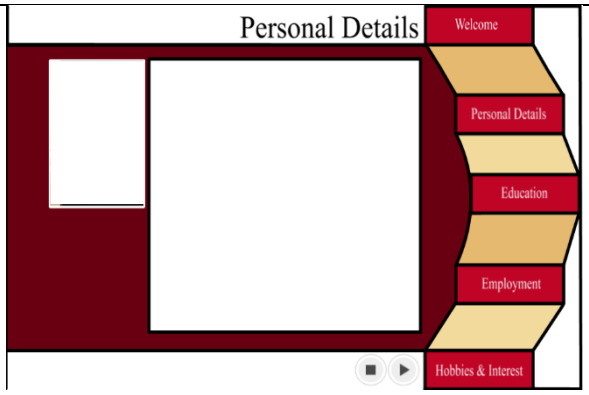
However, one area an influence from the exemplars to the final year cohort is apparent was one very particular case of a *curved navigation panel*. This was probably the most impressive exemplar demonstrated to the students – and it was the only one in the entire (2012-2013) cohort to have a curved navigation bar, where the buttons in it were not vertically (along the left or right of the screen) nor horizontally aligned (along the top or more rarely the

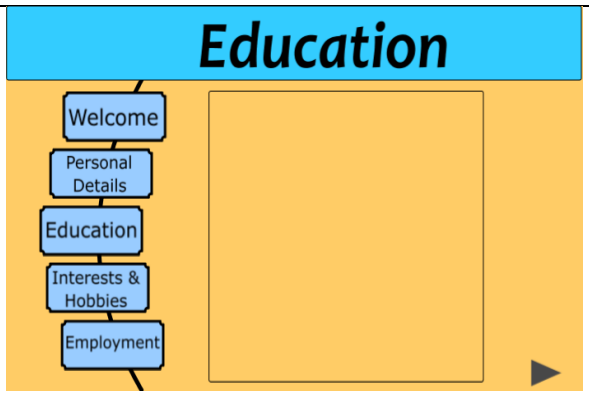
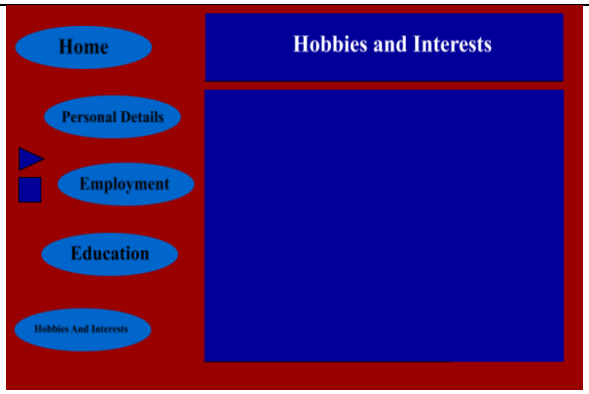
Figure 9-8: cv2.swf Exemplar

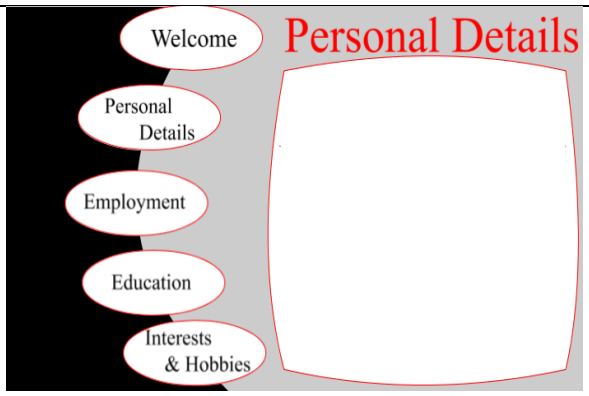
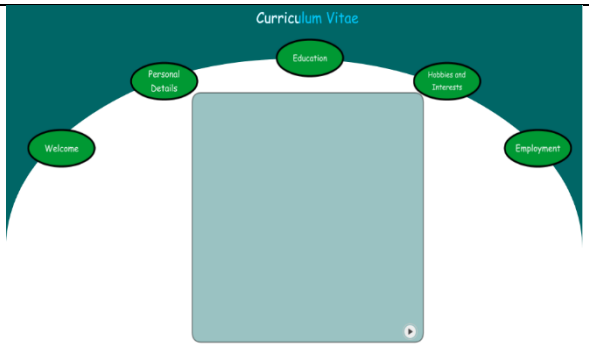
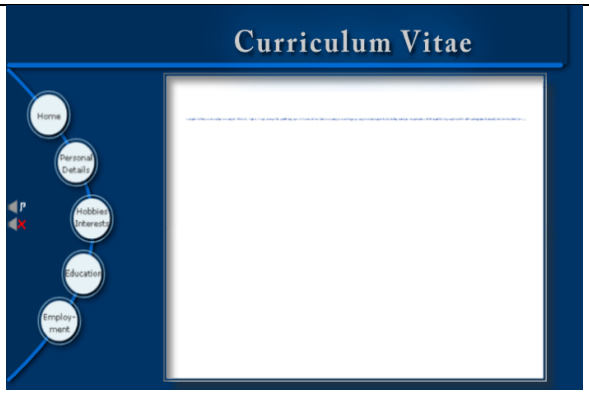
bottom of the screen), but were instead centered on a curved line running from one end of the screen to the other. Here also, the buttons themselves are of oval shape. In the later (2013-14) iteration of the course this particular idea was implemented by 12 students. This implementation took place in items of varying levels of achievement and in all but one case the idea had been integrated into a quite different aesthetic.

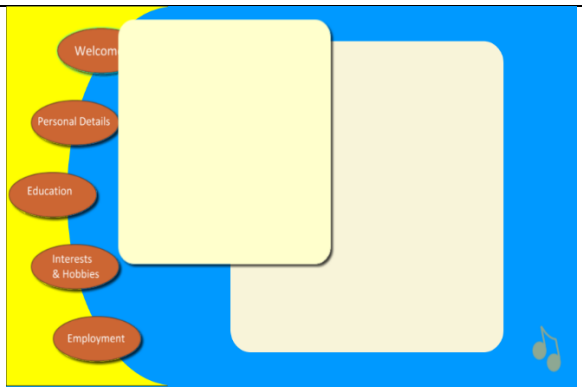
9.2 Examples of a Curved Navigation Bar

This uses a blue on grey rather than the autumnal colours of the exemplar. The music is simple piano playing rather than the classical flute in the exemplar. The panel containing content does not have fly-ins, unlike the original. The use of curved frames around the central panel is fairly unusual, whilst the music player is stylistically a little jarring and unlike the simple double quaver image in the exemplar. For all the other buttons, the oval shapes of the exemplar are followed.	 *Figure 9-9: Student Curved Navigation Bar 13-14 (1)*
In this example, the colour scheme is autumnal, however, the buttons are not enclosed in clickable shapes and rather, it is the words themselves that are centred on the curved line. This also does not have fly-ins for the change of content in the central panel. While stylistically elegant, there were bugs which resulted in the central panel not showing fixed content, but instead, cycling through content randomly.	 *Figure 9-10: Student Curved Navigation Bar 13-14 (2)*

This example has a brighter palette than the exemplar and whilst the buttons, like it, are oval in nature, the font is of a higher contrast. The music toggle is implemented just as a word, with a line through it when it is in a "playing" state, thus requiring the user to click on it again to stop it playing.	 *Figure 9-11: Student Curved Navigation Bar 13-14 (3)*
This embodies an interesting mixture of the curved navigation panel, however, transported to the bottom of the screen, but it is marred by stylistically very basic buttons – together with a play/stop button taken directly from the Adobe Flash built-in libraries. It is also an example of a prominent white background for the information panel, which is an interesting feature surprisingly common in the 2013/14 cohort.	 *Figure 9-12: Student Curved Navigation Bar 13-14 (4)*
This is one of the most creative re-purposings of the exemplar. A curved navigation panel in terms of the locating of its buttons, which are assembled in a 3d concertina arrangement. White has also been used for its information panel, however, like most of the others, there are no fly-ins for the information. Also, "library" buttons have been used for the music player, and though slightly stylistically out of keeping with the rest of the UI, not terribly so	 *Figure 9-13: Student Curved Navigation Bar 13-14 (5)*

This example has a relatively naïve interface, but has an interesting reveal on the "welcome" screen. The play/pause button stylistically works well, however, it does not have an alternative mouse cursor when the mouse goes over it. Another interesting thing about this exemplar is that it incorporates an anomalous feature found in another of the exemplars, namely, an inverted rounded corner for the buttons (not visible at this size).	 *Figure 9-14: Student Curved Navigation Bar 13-14 (6)*
This is possibly the most dramatic reworking. The curvature of the buttons is subtler, the colour scheme very monochromatic, and the text is written at a 25 degree angle (as per the "WELCOME" in the image), which continues down the page and floats around the passport style picture frame. This design scheme is kept up on all screens. The music pause/play button is a bit incongruous in terms of style and positioning.	 *Figure 9-15: Student Curved Navigation Bar 13-14 (7)*
Like the exemplar this uses oval buttons and like it also does not have high contrast between the text and the background in those buttons. The music player is reasonably well executed and there are animations around some of the pictures, which can be a bit distracting. The text is white and so there is high contrast against the background of the information area. The bright text in a serif font against such a dark background does not quite work.	 *Figure 9-16: Student Curved Navigation Bar 13-14 (8)*

Another example of an interface with a lot of white. It follows the exemplar with oval buttons centred on a curve and the information area is also is curved (curving out). It does not have a music button and is also slightly buggy in that returning to the welcome screen causes a second music player to start up, out of sync with the first one.	 *Figure 9-17: Student Curved Navigation Bar 13-14 (9)*
The use of green colour schemes in this year was less than in the previous one, however, this example uses it, with an arrangement of oval buttons around a curve, this time at the top of the screen. Like other examples in this year, there is a large amount of white. It has superb fly-in animation on the welcome screen and the music toggle is tastefully located at the bottom of the information screen. No fly-ins, but wonderful transitions between information screens.	 *Figure 9-18: Student Curved Navigation Bar 13-14 (10)*
Begins brilliantly with the buttons animating along the curve line on the left to their ultimate positions. The text is high contrast in the information screens, but with a filter applied over it, which makes it difficult to read. The buttons themselves are adaptations of buttons in the Flash library and are therefore of a lesser achievement than the animation that brings them on screen.	 *Figure 9-19: Student Curved Navigation Bar 13-14 (11)*

This example is the only one that really follows the exemplar in a very direct fashion. Aside from the brighter interface, the choice of background music is the same, the same double quaver button is used for the music toggle and fly-ins of panels have rounded corners. The only element that is different is the sound of button clicks, which is garish and not really suitable for the artefact.	 *Figure 9-20: Student Curved Navigation Bar 13-14 (12)*

Only one example shown above is an uncreative emulation of the exemplar (the last one) – all the others involved attempting to integrate this new element (the curved navigation bar) into very different artefacts. Moreover, it's not a case of simply high achieving students who opted to use this technique. Two of the examples scored in the 40s and one in the 60s, whilst all the others registered 74 and above.

9.3 Use of Inverted Rounded Rectangles for Corners

Another case of a unique feature among the exemplars getting repeated in a number of subsequent artefacts, was the use of inverted rounded rectangles. That is, when the radius of the corner of a rectangle is set to a negative value. For instance:

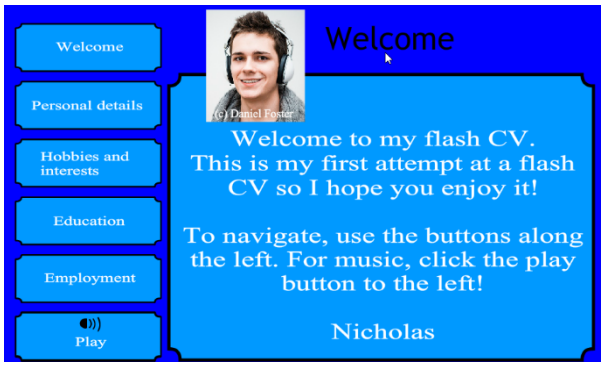


Figure 9-21: Exemplar cv1.swf 12-13

Figure 9-22: Inverted Rounded Rectangle Corners in Button

the corner of which, close up, looks something like  . As a design feature this is fairly unusual, and in fact this student during the 2012-2013 iteration was the only student to use it. This was one of the two exemplars used in the "final" feed-forward evaluation session in 2013-14, seven students used this device.

A very elegant artefact which uses a gradient background and an innovative button shape. The creator has used inverted cornered rectangles, but unlike the exemplar, has not used them for buttons, but instead, for the central information panel and the title area. The music toggle deploys a library button and is not really in keeping with the rest of the screen, but it is small enough not to disturb things.

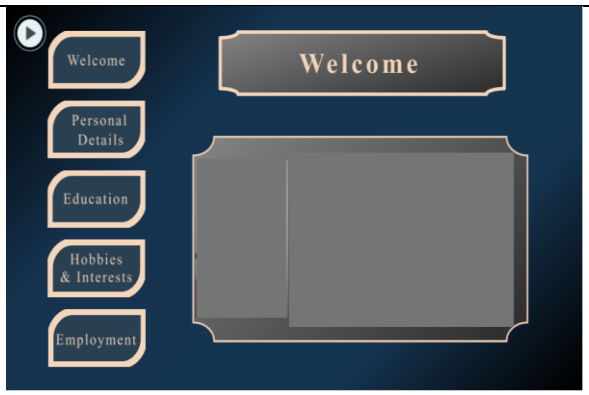


Figure 9-23: Student Inverted Rounded Rectangles 13-14 (1)

This was also seen in the example above – the inverted corners are very small on the buttons, but visible to the user. This is an unusual artefact in that it directly incorporates features from more than one exemplar.

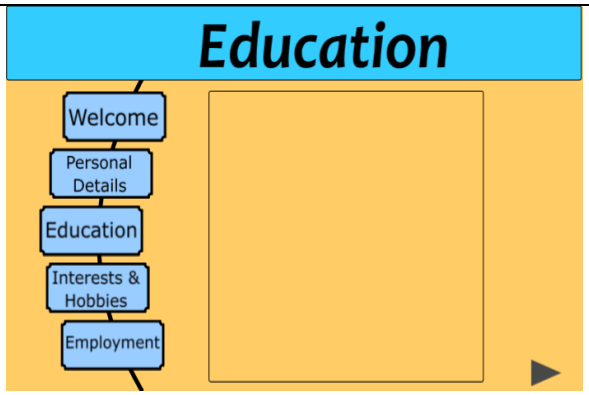


Figure 9-24: Student Inverted Rounded Rectangles 13-14 (2)

Not particularly easy to make out given the lack of contrast between the button and the background (especially as the button uses a black line style of only slightly different colour from the background itself). On screen, however, this is a very elegant looking artefact, although it lacks an introductory animation.

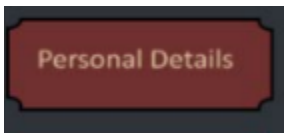


Figure 9-25: Student Inverted Rounded Rectangles 13-14 (3)

Another example of quite small inverted corners. The overall Flash file has a striking opening animation, but a more naïve look in the rest of the artefact, with large serif text. Noticeably, the buttons themselves have text of varying font size.

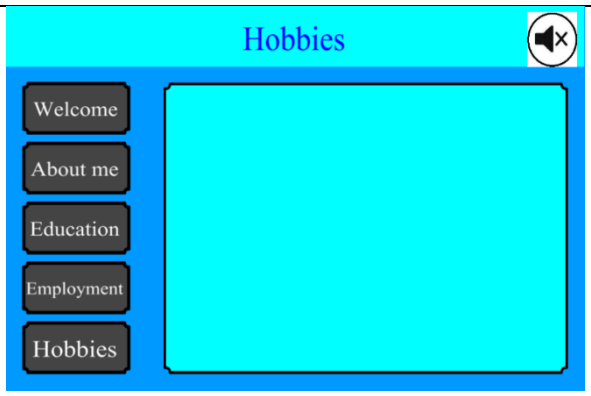


Figure 9-26: Student Inverted Rounded Rectangles 13-14 (4)

An innovative and more extreme use of the inverted curved corner. So much that a vertical side to these buttons can hardly be said to exist at all. The rest of the artefact is similarly extreme, having constant colour changing over the bottom six red rounded rectangles as well as the letters in the "WELCOME" text changing every half a second.

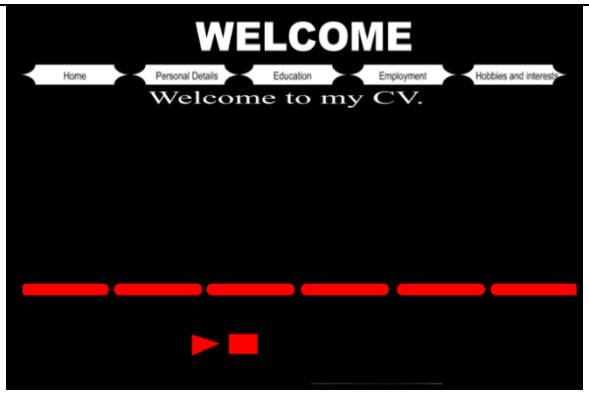


Figure 9-27: Student Inverted Rounded Rectangles 13-14 (5)

This example uses the inverted curved corners on the central panel as well as the buttons, but not all of them. The home button has normal curved corners, whilst all the other buttons have the inverted version. There is no background music toggle in this example.

Figure 9-28: Student Inverted Rounded Rectangles 13-14 (6)

This example is the most faithful to the original, however, it has a number of divergences, such as a textured background. It is innovative in its placement of the music toggle. It also presents a superb opening animation of raindrops to coincide with the cloud-texture used behind the blue panels (I have put in a non-gradient blue over certain areas for anonymisation purposes, whilst the real artefact uses texture everywhere.	 *Figure 9-29:Student Inverted Rounded Rectangles 13-14 (7)*

9.4 Conclusion

The purpose of looking at these examples is to see the ways in which features of the exemplars become utilised, but refigured in the work of the later cohort. And in doing so, they are going beyond previous conceptions of understanding the marking criteria, or understanding some abstract notion of "quality". That is, they embody a spirit of synthesis and transcendence, in which new things can be incorporated, and yet this incorporation itself is done with some effort of integration into a harmonious and coherent whole.

Looking at these artefacts is important in terms of the debate around the effects of so-called "feedforward" exemplar based teaching approaches. Hendry, White, and Herbert (2016) in a very recent paper covered this in the context of an assignment requiring students to produce a critical review of a scientific article, saying:

> anecdotally, some colleagues have expressed concern about students being exposed to examples of peers' good quality written work because they think students may be tempted to 'closely model their own work' on the good example(s) and, as a result, produce a formulaic product that lacks independent thought and/or creativity. (Hendry et al., 2016)

One student in that study reported that exposure to exemplars helped them "to mimic (for the lack of a better word) writing style, language, review structure and utilise that in my own assignment for a superior result'. The authors were worried that potentially students might become adept at simply 'mimicking' the genre of the critical review assignment without acquiring critical writing skills in depth.

Certainly, in the examples shown in this chapter, there has been one case where the model was used in an extremely direct way. However, in all the other cases, what they did was to take a feature of one or other exemplar, and recombine it in a new artefact often with a very different aesthetic from the original. Hendry et al.'s paper also questioned whether some students after exposure to exemplars, might not have sufficient writing skills to be able to benefit from them when carrying out their own critical analyses. In the current case, this issue pertains to technical skills rather than writing ones. Certainly, in the examples presented, sometimes students have maybe attempted to overreach themselves, demonstrating an ability to implement the curved navigation bar for instance, but then using much poorer sound or transition effects than present in the original exemplars.

Concluding, in this chapter, we have seen that the effect of presenting exemplars goes beyond students complying with assessment criteria, or understanding them. For as well as developing a greater sensitivity towards problems or defects in their work, exemplar marking with live peer assessment constitutes also source of inspiration, a place from where they can be stimulated to seek approaches to try out some of the things they have seen demonstrated.

Chapter 10. Summary and Discussion

What has been demonstrated so far, is that the use of live peer evaluation using clickers can have transformative effects in terms of student engagement and student achievement. Moreover, it can do so extremely cost effectively. All the improvements observed in the scores on the E-Media course were the result of just two hours of contact time for each iteration of the course. Those two hours, the first a rehearsal session for evaluating multimedia CVs and the second, an official one, were the only real changes to students' timetables compared to prior iterations of that course. In sum, the final sessions for the BSc course the lecture theatre accommodated at least 180 students, thus demonstrating that this method is extremely scalable.

For the masters' course, the student enrolment was smaller, and given the more complex artefacts being developed, the evaluation sessions took longer. In these cases, four hours for assignment 2 was necessary (two hours rehearsal and two hours final), and similarly for assignment 4. That is, on this course, where real summative peer-evaluation took place, the amount of time dedicated to it was greater, but still in single figures: eight hours of contact time per iteration. Of course, much more time was needed in terms of preparation, but if a process like this is embedded, then this should progressively become less. The recommendations I would make in this regard are set out below.

10.1 Logistical Recommendations: Preparation

In terms of preparation for these sessions, there are a number of steps that need to be taken. Firstly, there is the selection of exemplars. Ultimately, there is not a scientific method to identify the most appropriate, but a number of considerations have to be borne in mind.

1. Try to find examples that produce non-divergent responses. This can be done by tutor marking exercises, where one can see which examples produce the greatest agreement between tutors. In addition, in the case of a repeating course, look at the previous years' voting on either submissions or exemplars.
2. Following on from this, some thought needs to be given to what the teacher wants to convey in a feed-forward exercise, or perhaps more appropriately, what is it the tutor wants the students to *mull over*. In the current case, it was the sense of what constituted quality in a multimedia CV

(on the BSc course) and in a general multimedia presentational artefact (on the MSc course). To do that, probably what you need are equal numbers of very high quality work, that is to say almost unimpeachably high quality work together with an equal number of not-quite high quality work – where the distinction between the two can be most clearly made, and the idea of quality made most tangibly manifest.

3. If only one person is marking the work, they should consider having an "ease of marking" category (not necessarily shared with the student) to fill in while marking. This refers to a record of how straightforward the marking was on any particular item and if it is, then this will make it a candidate for inclusion in subsequent feed-forward exercises.
4. Anonymise artefacts that are being represented to the class, but try to keep the exemplars as similar to the original submissions as possible. Essentially, all types of artefacts, even those that are not optimally anonymised, as was the case with the E-Media Design course, will result in learning gains, however, the more credible the exemplar the more effective the intervention.

Another thing to be borne in mind are the logistics of clickers and authentication. For all except the first iteration of E-Media Design, the clickers were given out at the start of the year to all the students. The roll call of students and their clicker IDs were built into the University of Hertfordshire LMS Studynet, which made the collation of large data sets very easy. For the masters course, an ad hoc set of clickers was used, which thus involved a long process of manual handing-out and collecting-back of clickers with a form needing to be filled in associating a clicker id with a student id.

10.2 Anonymity vs Accountability: Libertarian vs Authoritarian

A consequence of this is also the question of whether participation in evaluations sessions should be anonymous or not. In all the sessions described in this thesis, all except those that took place in the final iteration of multimedia specification took place with the tutors knowing who clicked what. However, in the final iteration this happened under conditions of anonymity, which resulted in much more generous and less rigorous marking appearing to take place. The value of a longitudinal study like this also allows for the examination of cases where things did not go to plan, thus being able to identify any potential flaws in the process. In addition, it results in better understanding of what findings were repeated and consistent as well as those that happened only occasionally. Examples of things not going quite to plan included: students "copying" the tutors' marking in assignment 4 2011-12, the over generosity of

marking under conditions of anonymity in assignment 4 2013-14, the evidence of non-engagement of a sizeable minority of students during the uncredited marking events for e-media design. This perhaps leads to the most complex of all the decisions to make when using peer evaluation, how to encourage participation and engagement. Moreover, should participation be mandatory, incentivised, graded, or at the other extreme, purely formative and anonymous?

It seems neither of the two extremes appear to work perfectly. The first is the more authoritarian option of awarding credit for participation and engagement by comparing the marks awarded by individual students with those of the tutor. The danger of this is that it represents value judgements as being akin to judgements of fact and thus, might engender a servile attitude, whereby the students try to judge exactly the same as the tutor(s). This can foster its own distortions, as was apparent on the MSc course in 2011, when students just copied the tutor. It clearly also results in students taking longer to make a decision. However, is that longer amount of time truly a reflection of higher-order thinking or merely the result of the stress caused by knowing that the score one awards might result in a lower or higher mark for the student assessor?

The other option, which might be called the libertarian option, refers to peer evaluation under conditions of anonymity or when offered in purely formative mode and also has its pitfalls. In this case, marking appeared to be more generous with high effect-size (Cohen's D) scores being reported. Also, on occasions when the act of evaluation was not credited on the BSc course, this resulted in non-engagement. Clearly, the two outcomes were directly related to the way the peer-evaluation was operationalised. For the MSc course, given that it was in everyone's interest to score highly, there was no reduction in the level of engagement (the amount of students actually clicking to any particular prompt), it was just that the scores were higher. On the BSc course, though not taking place under anonymity, but with such a large course and no incentives for actually doing the marking, then a not insubstantial minority of students just did not actually click to the prompts given.

Given the above concerns some kind of middle way between these two polarised stances is recommended. For instance, make the students 1) attend and 2) respond to at least 80% of the questions and 3) have a positive correlation (say > 0.2) with the tutor marks. This lattermost condition, though it might permit quite low correlations, would also communicate clearly that marking patterns will be scrutinized. Also, it would be reasonably straightforward to explain (student marking should have

at least some relationship to tutor marking, however small). Performing these exercises in a rehearsal, where at the end of the session the students could be emailed with the correlation value they obtained, would help to inspire confidence.

10.3 Debriefing

At the other end, there needs to be some kind of debrief for the tutors. Some of the statistical measures that can be used to evaluate the success of a peer evaluation session have been discussed in the thesis, these being:

1. *Correlation between average student mark and average tutor mark*
   (This enables observation as to whether student marking is more or less OK and is important when peer evaluation is applied summatively to the ratees)
2. *Effect size*
   (This will show if students are marking over generously or too strictly)
3. *Median correlation of individual students' marking correlated with that of tutors*
   This helps in understanding whether, overall, students have in some way taken on board the "appreciative system" being projected by the tutors. The higher the median value, the correlation between the average student marks and the average tutor mark can be considered as not merely an effect of an averaging process
4. *Magin's Reciprocity Matrix*
   (This will inform the tutor as to whether there is any kind of reciprocal over marking at play, where individuals or groups agree to award higher marks to others in return for the same)
5. *Krippendorf's Alpha*
   This gives very interesting results, but they are relative to the conditions under which peer evaluation is taking place. In the E-Media Design course, on a number of occasions a high Kalpha was witnessed, in situations where students were making many different judgements regarding the same artefact, many of which were almost binary (Yes/No/Maybe) However, on the masters course, where many different artefacts were being presented, with few criteria per artefact and those criteria being highly subjective, the score was typically lower, but nonetheless, relative to one another, they did highlight the more successful interventions compared to the less successful ones)
6. *Agreement and Overmarking Count*
   (Agreement always has to be defined by the tutor and is not really comparable across different

studies. However, if a reasonable definition of agreement is reached, then it is possible to count the number of times students agree as a proportion of all judgements. With that done, it will also be possible to verify whether over generous marking is taking place. Moreover, it will highlight very quickly if there are issues like copying of tutor marks.)

7. *Skewness and Kurtosis of the distribution of individual student vs tutor correlations*
   High values here is a very clear indicator that things have gone wrong and that the variation of student levels of correlation with the tutor is not distributed normally. This could be for a number of reasons. In my case, during one iteration, I found a number of students copying the tutor's marking – but other occasions where a subset of students engage in aberrant marking practices can easily be envisaged.

It is not always the highest value that is desirable in some of these categories. A correlation between average student mark and average tutor mark in the 0.7s is a good outcome, but one much higher than that, might indicate either problems arising from students acting in concert, or copying the tutor or slavishly seeking to reproduce the exact type of thinking as the tutor. What is wanted is evidence of responsible independent judgement. The effect size would ideally be around 0, but would only become a cause for concern when it goes above 0.4. We also don't want the median correlation between individual students marking patterns and that of the tutors to be too high, nor Magin's reciprocity matrix to be too low, since they too would represent an excessive purity which might be at odds with developing independent judgement. Clearly at present, to perform all the tests requires a lot of tweaking and Excel tinkering. Some LMS's do have rich peer-assessment functionality (for instance Moodle's Workshop), however, they do not provide the rich digest of statistical data, which is important for evaluating the success of a live peer evaluation event.

10.4 Benefits to the Student

In the course of this research, it has been demonstrated how the act of collective evaluation of other students' submissions, whether that of a prior cohort, or of peers, leads to students having a better understanding of the marking criteria and subsequently, producing better work. I believe this is down to six factors:

1. It enhances the authority of the rubric as a living and foundational document
2. It makes its (abstract) terms meaningful by reference to real examples

3. It establishes a repertoire of examples
4. It models the "noticing" of salient features
5. It helps students calibrate their own responses to examples by comparison with the class average
6. It inspires though high quality exemplars created by comparable peers

It is the potency of these things together that can account for the dramatic influence it has had on the E-Media Design course. These benefits would probably apply to other subject areas, but I believe there are specific reasons why it is particularly applicable to computer science. However, if these things are the answer, what is the question? Or in other words, what is the lack or deficit to which this technique provides some solution?

10.5 Surface vs Deep Learning

If any article can be said to most clearly impact on pedagogical thinking in Higher Education over the last 20 years, then it is that of John Biggs' "What the Student Does: Teaching for Quality Learning" (Biggs, 1999). He popularised a number of terms that have become central to pedagogy in HE, including "surface learning", "deep learning" and "constructive alignment". Writing at a time on the cusp of the massification of higher education he describes two stereotypical students: traditional Susan and non-traditional Robert. Susan is diligent, learns for its own sake, reflects on what she does and she is what is called a deep learner. Robert merely wants a qualification and just wants to pass; he takes a surface approach. Biggs' goal is to set out the methods of teaching that will effectively force Robert to change his approach and adopt a deep learning mind-set, because the challenges posed mean no other approach would be successful. In terms of taking lecture notes, Biggs suggests "Susan is relating, applying, possibly theorizing, while Robert is taking notes and memorising". Biggs declares: "Good teaching is getting most students to use the higher cognitive level processes that the more academic students use spontaneously". To this end, he recommends a *constructivist* pedagogy, under which it is taken that student learning occurs not through *direct instruction*, but through *learning activities*.

To make this happen, Biggs' argues for more explicit criterion referenced learning outcomes and then, the design of activities to bring these to fruition, a practice he calls "constructive alignment". He goes on "It is a fully criterion-referenced system, where the objectives define what we should be teaching; how we should be teaching it; and how we could know how well students have learned it". In saying this, Biggs challenges a number of previously held assumptions about "spreads of marks", contending that if outcomes are criterion referenced, the spread of marks, and discrimination itself, should not be a

goal: “teachers shouldn’t want a ‘good spread’ in grade distributions. Good teaching should *reduce* the gap between Robert and Susan, not widen it”.

Biggs’ prescriptions have proved extremely persuasive and underpin a lot of training of academics in Higher Education. However, in some ways it does not tell the whole story and some of its emphases do not easily map to the kinds of things students had to do on the courses described in this thesis.

Before the use of collective evaluation using clickers, the course had involved using explicit criteria and there had always been plenty of practical work related to the CV assignment. However, the notion that the learning activity necessarily resulted in a consequent learning outcome was not borne out. Instead, it was found that not only does the learning outcome (as embedded in an assignment rubric) need to be clear and the activities relating to them aligned, but also, the students *need to understand* the learning outcome. Moreover, they need to see embodiments of those learning outcomes in order to have a tangible understanding.

Again describing the difference between academic Susan and non-traditional Robert, Biggs writes:

“He is less committed than Susan, and has a less *developed background of relevant knowledge*; he comes to the lecture with no question to ask. He wants only to put in sufficient effort to pass. *Robert hears the lecturer say the same words as Susan heard, but he doesn’t see a keystone, just another brick* to be recorded in his lecture notes. He believes that if he can record enough of these bricks, and can remember them on cue, he’ll keep out of trouble come exam time”. (my emphasis)

The emphasis I have applied here is not foregrounded Biggs’ writing, but to my mind allude to a much greater defining fact of Robert and all such non-traditional students. He has a “less developed background of relevant knowledge”, which corresponds to the “repertoire” written about by Sadler. This is the library of past examples in the practitioner’s head that guides the various “moves” he/she attempts in order to improve or perfect the artefact on which he/she is currently working. This lack of background is not a minor issue but rather, a crucial limiting factor when the student attempts tasks for which he/she has no context unlike his more knowledgeable peers. Moreover, this lack of background also prevents the student from distinguishing the salient from the commonplace: “Robert hears the lecturer say the same words as Susan heard, but he doesn’t see a keystone, just another brick”. Because the use of collective evaluation involves the lecturer pointing out the salient, thereby seeding students’ minds with some basic repertoire of examples, at least some “background of relevant knowledge” is added and moreover, the lecturer is capable of pointing out what is really important as opposed to what

is not. The fact that the student has to vote in such events, and therefore commit to certain judgements about things, means their participation is already more active that would be the case if the tutor merely declared what is good work and what is not good work.

In Biggs' formulation, the deep and surface approaches are not representative of character (e.g. shallow vs profound), but merely of an approach to learning. Also, it has been shown in this thesis how much very small interventions can have unintended consequences in terms of the deep and surface approaches. For instance, anonymity in the 2013 iteration of the MSc course allowed the students to grade over generously, clearly nudging them towards the surface mind-set, i.e. the course is merely something to pass. The copying of the lecturer's marking in 2011 also demonstrated a surface approach where students, even though the benefit to them in terms of marks was negligible, nonetheless, preferred to accumulate the marks rather than develop independent judgement.

However, when undertaken well, live marking can be a very powerful tool for encouraging students to adopt *deep* approaches to learning. As recommended above, the conditions under which this takes place need to be judiciously engineered to be neither too authoritarian or too libertarian so as to allow students to achieve both responsibility and independence. On the problematic occasions where this technique was used (copying the tutor/grade inflation), we can see that use of surface and deep approaches by students are not a purely individual response, but will also be determined by the emergent culture of the course – with a tendency to surface or depth being established. However, it has also been elicited that the influence of such feed-forward events does not just present in the session itself, but structures the subsequent conversations students have about their assignments, and those conversations, as emerged from the focus group discussion and thus demonstrates that they can engender a deep learning approach.

10.6 Further Research

The two areas where the kind of findings in this thesis might be further interrogated would be (1) social equity in computer science education and (2) potential for the use of these techniques in MOOCs.

10.6.1 Social Equity in Computer Science Education

We saw a striking increase in student attainment in the E-Media Design course, but it is unclear which demographic it affected the most, beyond that of those who typically scored very lowly previously. The research undertaken for this thesis was focused on the student body as a whole rather than on any identifiable subsets within it.

However, further examination related to differential effects on different groups could be a very interesting line of enquiry. Computer science, as a subject, is in the unenviable position of having witnessed a precipitous decline in female enrolment since the mid-eighties. Thomas Misa, writing in 2011 states:

> "The most recent NSF figures suggest that women may account for just one in seven undergraduate computing students, or around 15%: a catastrophic drop from the peak of 37%." (Misa, 2011)

In 2012, the Council of Professors and Heads of Computing's "CS Graduate Unemployment Report" (2012) noted "massive difference in female representation against other subject areas (16.8% v. 57.1%)" (Mellors-Bourne, 2012).

In terms of race the picture is much more complicated. In the United States, books like Margolis et al.'s Stuck in the Shallow End (Margolis, Estrella, Goode, Holme, & Nao, 2010) show significant disincentives to enrolment among black and Latino students. In the UK, the picture is more nuanced, with BME students actually increasing in computer science. The CPHC report also states ""CS has been extremely successful in attracting a higher proportion of Black and Minority Ethnic (BME) students to study, to the extent that we have significantly higher proportions of BME Students than the average across all subjects" (Mellors-Bourne, 2012). However, there is an attainment gap with white students and a greater concentration of BME students is found in the post-92 institutions, with the consequence of a higher graduate unemployment rate.

Historically, computing has been a profession less credentialed than many others. Many great icons of the PC age, such as Bill Gates, Steve Jobs and Mark Zuckerberg were students who dropped out of college. The computing industry has been built around auto-didacticism, but though the industry has been an enabler of a certain kind of social mobility it may be true that computing at universities has less claim to that title.

However, it is possibly the auto-didactic affordances of computing that potentially result in its unattractiveness to unprepared students. Margolis' Stuck in the Shallow End describes a project involving classroom observation of a mixture of schools in the Los Angeles area. Describing the observation of one class she writes:

> "One female student explained how some students would purposefully use code that went beyond what was taught to tackle assignments, simply to show the rest of the class the extent of their "advanced" knowledge. All this perceived "showing off" resulted in an atmosphere wherein some students felt reluctant to ask for help…On more than one occasion, our researchers observed the most tech-savvy students rolling their eyes during class discussions, or making snide remarks about what they construed to be their teacher's or classmates' lack of knowledge. This disrespect was thinly veiled."(Margolis et al., 2010)

Clearly, in all subjects there will be some students who have a head start, however, there are probably few where it can be as tangibly manifest as computer science.

The association of computer science with solitary learning, and stereotypes like "nerds" and "hackers", means that potentially, among any first year cohort of students, there might be hugely varying levels of computing and programming skills. This wide disparity of level means that nurturing a genuine "community of practice" can be difficult – instead there might emerge atomised cliques of students who, defined by their ability level, create their own mini-cohorts. Lewis et al. have pointed out how stereotypes of computer science students are problematic as "students measure their fit with CS in terms of the amount they see themselves as expressing the traits of singular focus, asocialness, competition, and maleness" (Lewis, Anderson, & Yasuhara, 2016).

For a course that presumes no prior programming experience, running introductory programming classes is potentially akin to running an adult literacy class, where some students are learning vowels, whilst others are writing sonnets. However, in order to achieve any kind of social equity, such that the "preparatory privilege" of certain students (e.g. through parental connection to IT networks, or school facilities) is not just consolidated in HE, such introductory classes do need to be run. In some senses, the disparity is insoluble, however, by making introductory computing more relevant to students lives, and encouraging a culture of collaboration, some advances to address this could be made. Potentially, the experience of the large lecture theatre, using peer evaluation or feed-forward, where the opinions of all students, not merely the most vocal, are aggregated together and where value judgements about student submissions are publicly negotiated, means a greater sense of belonging can occur. Further research could involve investigating how techniques of live peer evaluation might differentially benefit less prepared students.

10.6.2 Potential for Use in MOOCs

One area that has recently seen a massive interest in terms of peer assessment is MOOCs (Massive Online Open Courses). Launched as a way of offering high quality education, largely through the use of expert instruction and various interactive exercises, they have been criticised for their transmission model of learning and in many cases, suffer from very low completion rates. However, a number of studies have shown extraordinary levels of educational innovation, especially in the area of peer assessment. Given a strong motivation for these courses is economic, that is, they are a way to deliver courses to vast number of students with minimal teacher intervention, their primary interest in peer-assessment is to facilitate some form of reliable scoring for assignments. For, when dealing with students in such large numbers, there can be little specific moderation, supervision or guidance.

Unlike in the cases in this, where the whole cohort, or half the cohort evaluated each submission, the MOOC model of peer assessment is undertaken asynchronously. This might also be a consequence of the artefact being assessed: only something demonstrable within 5 minutes can be reliably peer assessed in a live situation, as anything longer would require prior exposure by assessors, which might not be reliable (that is to say, participants may attend a marking event declaring they had examined the items under consideration when that might not be true). For this reason, a small subset of students has to mark each assessed item. In order for a score be given to the student who did the work, some mathematical treatment has to take place to increase the likely reliability. As a consequence, a number of models have been developed to improve reliability. Piech et al. (2013) have a model that factors in grader bias and grader reliability in order to arrive at a final score. Interestingly, they posit the truth of a grade to be the student average as opposed to the tutor's and moreover, believe that any repeated discrepancy between the two is a function of a poorly drafted rubric. For instance, in a cohort of 1000, where each student marks five artefacts, then each artefact should receive five grades. However, it might be possible to arrange things in such a way that a particular subset of artefacts is graded much more often (which they call "super-graded"). From these examples, a comparison with the tutor awarded grade can be made as well as between the student rater and the average for any super-graded artefact to estimate the grader's reliability. Subsequently, the grader's own score in assignments is used as a factor in their reliability: the higher scoring the grader is, the higher reliability in the grades they award. Whilst the full significance of the figures are difficult to fully grasp, in the reported statistics, 97% of assignments being graded to within 10% of the "true" grade is impressive. However, one flaw in this model is the treatment of the reliability of a grader as something almost fixed; a function of the amount of time they spend grading and their own level of grader bias.

A study involving a much greater sense (like the current one) of the act of grading as being in itself a transformative process, is Balfour's "Assessing Writing in MOOCs"(2013) a model called "Calibrated Peer Assessment". Explaining this procedure, Balfour writes:

> First, students write an essay which is scored by taking the weighted average of ratings given by three peer reviewers. Second, the students calibrate to the instructor's expectations by rating three essays provided by the instructor on a multiple-choice rubric. The instructor assigns a correct answer to each item on the rubric for each calibration essay and the students are scored by how well they match their instructor's answers. At the end of this task, students are assigned a Reviewer Competency Index (RCI) which functions as a weighting multiplier on the scores they have assigned to other students. Very low RCIs result in a 0 weight. Third, each student reviews three of their peers' essays with the rubric. The peer review task is scored by how well the individual reviewer's rating of the essay matches the weighted rating of the essay. Finally, students complete a self-evaluation of their own essay which is scored by how well they match their peers' weighted review scores.(Balfour, 2013)

Here, there is the explicit assumption that the student should learn to grade like the tutor, and that students would receive credit to the extent that they do. Then, this level of correspondence is factored into any evaluation score they subsequently give, with a weighting defined precisely by the closeness of the marking pattern. In a situation such as a MOOC, it is perhaps the most practical way of attempting to influence the way students' grade. However, as has been elicited in the current study, when the tutor's rating is taken as being the true rating to which students aspire, there is the danger that they will interpret the exercise as mind reading or extended second guessing, rather than assimilating that way of evaluating into their own mind-sets.

Alongside these optimistic reports about peer-evaluation in MOOCs, there have also been a lot of less optimistic reports of very low completion rates and very differing levels of engagement among students. Balfour himself states:

> Thus, if a MOOC has 100,000 students in it, and 10% finish the course (10,000), a 10% problem rate in CPR would translate to 1,000 essays with potential scoring problems. …. Most MOOCs do not reach the 100,000 student level; a 2,000 student MOOC may only have 20 problem essays. (Balfour, 2013)

Whilst the final number does seem manageable, those 20 problem essays would also represent 20 examples of injustice. Suan (2014) reviewing different papers on peer assessment in MOOCS describes a number of possible sources of rater error and then adds: "one remaining problem with peer assessment in MOOCs is the probability of an assignment being rated by all poor raters." Some possible solutions he entertains includes:

> facilitate online discussion forums by putting more weight on opinions of student raters whose judgments of peer performances are close to that of the instructor's. Another potential use is to use student raters' performance-as-raters to supplement final summative evaluations of each student for the purpose of credentialing. (Suen, 2014)

But again, this use of the instructor as the gold standard, and this epistemological absolutism, where statements of opinion become reified as statements of fact, and even more absolutely, become embodied in credentials, opens the door to distorted outcomes, where students are mimicking the tutor's judgement, rather than truly assimilating it into a new world view.

Live peer marking has been demonstrated in this thesis as scaling very well to large audiences in lecture theatres and with the availability of new mobile rating apps, such as Poll Everywhere ("Poll Everywhere Home Page," 2017) and Turning Point's ResponseWare, this will increase its applicability. Combining these with video conferencing platforms mean the kinds of experiences described in the classroom in this thesis can also be performed online, with the same scalability. The only difference is the inability of students to "speak to their neighbour" in the same way as might in the lecture theatre. As has been demonstrated, the power of live peer marking comes from the ability of students to compare their own voting/grading with that of their peers. Moreover, the discussion around this can help students establish their own appreciative system, which will be experienced as unique and a product of the integration of themselves and the habitus of the field, not something which is an imperfect replica of the tutor. Moreover, when run by a competent presenter, it offers opportunity for dialogue and the modelling of what Sadler called "noticings", which in the current research pertains to the highlighting of relevant parts of the artefact that illustrate its overall quality level.

Another area of research in peer assessment, particularly for MOOCs, is that of "ordinal grading" or "comparative judgement". In this formulation, rather than judging artefacts on explicit and objective criteria, the object is to simply rate them as better or worse than others. Over time, with enough such ratings made, a global ranking can be created and from that, potentially the extrapolation of what is

good and great can be undertaken by the student. Shah et al., in "A Case for Ordinal Peer-evaluation in MOOCs" (2013), believe that, given a large number of participants in a MOOC will not be expert graders, the results given by ordinal grading are much more robust than those given by cardinal grading. Caragiannis et al. (2016)in "How effective can simple ordinal peer grading be?, propose sets of complex mathematical "aggregation rules" to get global rankings:

> Each student has to rank the exam papers in her bundle (in terms of quality) and an aggregation rule will then combine the (partial) rankings submitted by the students and come up with a final ranking of all exam papers; this will be the grading outcome. Information about the position of a student in the final ranking (e.g., top 10% out of 33 000 students) can be included in her verified certificate.

Pollitt (2012) proposed a method called Adaptive Comparative Judgement, which is ingenious in the way it adapts. Essentially, random pairs are produced for markers to judge the winner and loser. In the next round of judgements, markers are given two winners to judge, or two losers. The same sorting procedure is carried out for two more rounds, until finally, a provisional ranking of items is obtained, which is then submitted for a further round of comparative judgement. Pollitt sums it up as follows:

> When an assessor awards a mark to an essay using a rating scale they are, of necessity, comparing that essay to imaginary ones in their head, opening the way to the three sources of error: variations in their personal internal standard, in how finely they differentiate the qualities they are looking for, and in their individual interpretation of the rating scale. ACJ, in contrast, requires direct comparisons, relative judgements, of the kind the mind is very good at making. The first two kinds of error do not operate in comparative judgement, and the remaining one is essentially a matter of validity rather than reliability. It is therefore no surprise that ACJ achieves higher reliability than rating scales.

Unfortunately, most of these techniques are mainly concerned with improving the reliability of assessments and consequently, have important deficits on at least two levels:

- the kind of feedback they give to the originator of the work is merely ordinal. To be sure, there is some value to knowing you're an "average" student - however without knowing how to improve, and without some identification of where defects and excellences reside in your work, it may be difficult to make that improvement

- If you are judging things by ranking them or by choosing a winner or loser, how can you establish that skill in “noticing”, which is emblematic of any skilled practitioner in any field; the skill to be able to reflect on what you are doing and knowledgably recognise where things need to be improved?

Given the challenges of scale, live peer assessment would likely be difficult to introduce on MOOCs for grading or credentialing purposes, however, as a technique for helping students develop their own appreciative system, and as a way for them to get an accelerated sense of the habitus, the perspectives and values of a field, it might offer a very compelling addition. Moreover, it might lead to the generation of higher levels of reliability for subsequent, genuine peer assessment. While live peer assessment might not scale so well (for time reasons) if it is summative (at least not in the model “whole cohort marks each coursework submitted”, which has been the case in this study), if used as a formative feed-forward exercise (where a whole class rates prior students’ work), it could easily be used with thousands of people together.

10.7 The Value of Live Peer Assessment

What this study, hopefully has demonstrated, is that live peer evaluation is a powerful method for raising academic attainment in students. The availability of technology that can easily aggregate student opinions simultaneously in live sessions, means that a kind of dialogue between tutor and student can take place even in situations where students are great in number. It enables tacit knowledge to be manifested and reflected on, it models the noticing of salient features when evaluating the artefacts of any discipline, it incubates the development of independent judgement where the values of a field are assimilated into the student’s own world view and finally, it allows students to compare their marking/voting habits with those of their peers and that of the tutor. Moreover, it communicates that students are participating in a community of practice; learning the ways of a field. For these reason, in the field of computer science it may be more important than in many other disciplines to implement live peer evaluation – precisely because of the stereotypes associated with “hackers” which privilege an attitude of obsessional asociality. However, in professional programming, forms of sociality are absolutely embedded in the practices of industry, in the form of agile working methods, stand up meetings, buddy programming, code reviewing and the writing of stories. The use of live peer evaluation can offer an antidote to this kind of asocial practice by showing that judgement ultimately is a matter of the community and not the prodigy. The other area where it might have application is in

feed-forward events on MOOC like courses, where the use of live peer evaluation can also bring some ethos of community into what might be experienced as an atomised and disconnected set of individuals.

Chapter 11. Conclusions

At the beginning of this study the research questions put were as follows:

- RQ1 How might EVS and LPA be designed and implemented?
- RQ2 How might EVS and LPA be evaluated?
- RQ3 What are the benefits and limitations of using LPA and EVS?
- RQ4 How do LPA and EVS work?

Some of these questions overlap in terms of response, but I proceed in addressing them one by one and with reference to the sub questions to which they relate.

11.1 RQ1 How might EVS and LPA be designed and implemented?

In this study, LPA has been used in two modalities, namely, summative peer assessment, where the mark awarded by peers has contributed to a student's score for the assignment and formative peer assessment (or better, feed-forward), where the marking of a sample of a previous cohort's work has helped students better understand the marking criteria and the nature of quality in a discipline. There is a very big difference in format between the two modalities, since in summative peer assessment, the students present to the class, but in formative, the tutor presents other students' previous work to the class

The logistics and practicalities of summative peer assessment on an MSc course was covered in chapter 3 and the same for a BSc course was covered in chapter 5. For summative peer assessment, five minutes were given per presentation, plus five more minutes for supplementary questions and the marking of the work according to two criteria. Prior to this assessment, a training session was used to acquaint students with marking and the interpretation of the criteria. Given the assessees were being voted on by their peers and saw the results immediately, I believe it is important that only group work should be assessed this way and that students should always have the option of not presenting (effectively delegating to the tutor). Accountability of marking is extremely important in this modality, so it is essential that there is a recorded ID for every marker.

For the BSc course, the purpose was formative and so there is no worry about possible unfairness of marking. However, there was a danger of non-attendance and also, within the session, non-engagement, which I deal with later.

11.1.1 What kinds of rubrics or marking sheets are best used?

Peer assessment is ultimately about marking by students and therefore, the terms used need to be comprehensible to them. Sometimes, the goals of comprehensibility and reliability will be in conflict. A very tightly drafted rubric that uses a number of sophisticated educational terms might be valuable in getting academics to produce reliable and repeatable marks, but one which is more colloquial, brief and looser might be better for students. Training in advance will help concretise and make comprehensible the inevitable ambiguities or intangibilities that come from representing in abstract terms a range of quality levels.

There are also different types of rubric, those with few and broad criteria (holistic) and others with detailed and very specific ones (analytic), with the MSc involving the former and the BSc the latter. However, even in the case of the MSc, the original four criteria were reduced down to two (for the prototype assignment) and the original five (for the final assignment) were reduced to three. The style of the attainment descriptors was considerably more colloquial than that of its predecessors, which was described in Chapter 3. The rubric used on the MSc course did not change over all four iterations, whereas that for the BSc course had two canonical forms as well as occasional variations.

The rubric used on the BSc course, in line with the different learning outcomes of that course, had a highly specific and analytical rubric, which meant fewer exemplars could be shown, because it took much longer to mark each of them. More specific rubrics are more likely to generate higher levels of agreement (since the criteria incorporate a lot of “has this” “has that” tick boxes and hence, will be more geared to recognising the self-evident). This would account for the higher Kalpha scores (measure of agreement across raters) on the BSc course when marking compared to for the MSc (see the Kalpha scores in Chapter 3 for the MSc and Chapter 7 for the BSc). In terms of quality control, however, and also for students’ own development, more holistic rubrics (the type Falchikov and Goldfinch have termed G+) might be more suitable. This is because whilst they are marked by dimensioned attainment descriptors, they nonetheless have sufficient breadth to *make students think* when awarding marks, and also allow for a more meaningful correlation statistic between tutor and student marks to be published.

The development of the rubric used on the BSc course was most clearly marked by the use of words like “appropriate” in the later version (see Chapters 5 and 6), which require some prior familiarisation by students in order for them to get their meaning and which also make for more demanding outcomes. This shows one of the benefits of feed-forward exercises that fully bring out any tacit assumptions in the

rubric – it means that richer rubrics can be used in the knowledge that students will more readily understand their terms having seen them applied to concrete pieces of prior student work.

In short, it is wise not only to simplify the rubric, but also, to provide timely training events, such that terms that may indeed have tacit meanings (such as “appropriate”, “original” or “superior”) can be illustrated *in situ*. Because the concrete manifestations of these terms can be made visible to the students, one has less need to worry about using inferential and less legally-watertight terminology.

11.1.2 What practices can be used to enhance the reliability of marks in this modality?

From the focus group with the MSc students (Chapter 4), the most important prerequisite was accountability, i.e. making sure that it is clearly known to the tutor who is marking who. On the occasion when anonymity was practiced for Multimedia Specification a much larger effect size between student and tutor marks was observed than on other occasions, meaning the scores were being inflated (see Chapter 3). Another practice would be to allow a broad range of permissible marks and penalising any student who went outside it (for instance penalising students whose marks do not correlate positively, however slightly, with those of the tutors). However, attempting to seek any closer correlation might make students merely wish to imitate the tutor, rather than exercising their own judgement.

The literature also talks about training as being valuable, in particular, because it allows the tutor to frame the act of evaluation. That being said, on the BSc course the marking patterns of those students who had attended the training sessions compared to those who had not were very much the same. This may be due to a natural process of homogenisation that might take place during LPA, in that the scores of the rest of the class are known to the individual markers and therefore, they naturally self-correct within the session, not needing any prior sessions to so do. However, when thinking instead from a feed-forward point of view, i.e. the effect on the students’ own conceptions of quality, the training sessions appear to be very valuable, as witnessed in the opinions of the focus group for the BSc course (see Chapter 8). This is because the commentary itself adds to the authority of the rubric and also, *models the process of noticing* – of highlighting salient features of quality, whether these be high or low.

11.1.3 What considerations should be borne in mind when choosing exemplars for students to mark?

When looking from a more feed-forward point of view, the major decision is the choice of training set. To this end, it has been recommended to choose examples from previous years that involve the least divergence (in the case of multiple markers) or that are subjectively the easiest to mark (if this indeed can be documented in a way that does not complicate the process of marking).

The two occasions when the choice of exemplars came under scrutiny was during the 2011-12 and 2013-14 of the BSc course (described in Chapter 5). The exemplars in 2011-12 were chosen after the tutors attempted to find artefacts that had the least divergence among them, whilst those in 2013-14 were improved by being more realistic. In Chapter 7 it emerged how some criteria required more time for students to choose when making a judgement than others. In addition, it was seen that, as might be expected, the higher order criteria required more time. However, this effect was not completely straightforward, since the difficulty of marking might also be a question of the *relationship* between the wording of the criterion and the reality of the artefact. For this reason, some prior vetting of exemplars to exclude those that produce highly divergent marking is recommended.

11.2 How might EVS and LPA be evaluated?

The value of a study like this is that, taking place over four iterations, varying levels of success in terms of student engagement, student results and satisfaction or dissatisfaction can be examined. The principal things to evaluate are: does it improve student performance and if it does, does it do so uniformly across a cohort? Finally, how can the LPA practice itself be evaluated, i.e. what are the measures of well run and effective LPA as well as what can hinder this?

11.2.1 Can LPA improve marks in an assignment?

The improvement in performance on the BSc under conditions of LPA is the clearest finding of the whole study. The following table comparing the results for the first assignment with those for the final one (the multimedia CV) on the BSc course illustrates the improvement in student performance.

Table 11-1: Comparison of baseline multiple choice test average score with final assignment average score

	Without LPA		With Live Peer Assessment			
	2008	2009	2010	2011	2012	2013
First Test	55.67	58.89	56.41	54.16	**N/A**	52.28
Final Artefact	55.3	58.3	65.4	68.3	71.4	71.6

On a year by year basis, small fluctuations in the first multiple choice test can be seen, which appears to indicate that there was no real difference between the cohorts. However, from the moment Live Peer Assessment was first used for the four year period, the average score in the final assignment went up by 13%. Admittedly, the marking rubric was different in the final two years from the previous two, but if anything, the rubric was more demanding, requiring a much greater level of stylistic integrity than in previous iterations.

11.2.2 If marks improve in an assignment, how is that improvement distributed across the cohort?

It seems that even very basic feed-forward exercises, without the sophistication in the practice that I had arrived at by the end of the study, will improve the overall average. This is because of the improvement it causes among those in the lowest 30% of the marks range. In the lowest two deciles gains of around 15% in student scores were observed in the earliest use of the technique. From the opinions in the BSc focus group, the evidence suggests that this is due to the act of LPA making the rubric a living document and a definitive point of reference for students doing their work. (See focus group Chapter 8).

Table 11-2: Scores for Final Assignment on BSc course by percentile

	Academic Years				
Percentile	09-10	10-11	11-12	12-13	13-14
10	28	43.25	43	53.52	52
20	37.5	51.5	56.8	62	61
30	45	57.5	64	66	67
40	52.5	65	68	70	72
50	60	67.5	72	74	75
60	65	72.5	74	78	77
70	72.5	75	78	80	80
80	80	80	81	82	83
90	87.5	85	85	87	87
100	100	97.5	95	96	97

11.2.3 What measures exist to quantify the success of an LPA event?

These are all measures to be used with student voting (not the subsequent results in the assignments) and are intended to quantify the health of a specific LPA event for a course.

- Correlation and effect size between tutor and student marking. A high correlation (>0.7) and a low effect size (<0.4) is desirable here. Higher effect sizes would indicate grade inflation, whilst lower correlations would imply there is a lack of clear understanding of the rubric.
- Proportions of agreement, overmarking and undermarking. If there is a higher than expected effect size, find out the amount of overmarking as this might indicate how concerted the grade inflation is.
- Median of correlations between individual students marking patterns with those of the tutors. Correlation may still be high between tutor and student marks, but the range of marks may be compressed – meaning the high achieving artefacts are not necessarily being sufficiently rewarded. A low median correlation might indicate a “noisy” range of voting, where poor voting is cancelling each other out but reducing the overall range of marks.

- Skewness and Kurtosis of the distribution of individual student correlations with tutor marking (anything outside the range of -2 to 2) indicates something is wrong such that something other than genuinely academic marking is taking place.
- Krippendorf's Alpha (essentially a measure of the level of agreement between members of a group when rating the same things.) This as a measure is not absolute and will depend on the level of *objectivity* of the criteria being used). However, it is a very good measure of the relative success of different marking events with the same rubric.
- Magin's Reciprocity Matrix will indicate if there is systematic "I'll scratch your back" arrangements among students. The procedure has been explained in Chapter 8.

11.2.4 Are students generally competent enough to mark their peers under these conditions and are they more or less so when compared to more traditional forms of peer assessment?

If the baselines are taken as being those of Falchikov and Goldfinch (r=0.69 d-0.24) as well as Li et al.'s meta studies (r=0.63), then the results on the Multimedia Specification course, where *summative* peer assessment is used, are very positive. When considering the occasions where student and tutor averages exist just for summative events (over the four years for 61 groups with five judgements for each piece of work – two criteria for assignment 2 and three for assignment 4), for the 305 judgements overall, the correlation (r) between the student average and the tutor marks is 0.775 and the effect size (d) is 0.244. Meaning the effect size is almost identical to that Falchikov and Goldfinch, but the correlation is higher.

Among these results, however, is the iteration of the Assignment 4 during 2011, when it appeared there was unusually high correlation and unusually low effect size, most likely because students were copying the tutor's marking. If that iteration is excluded, very good figures are still delivered, with the correlation being (r) 0.749 and effect size (d) 0.303. Over the course of this study, whilst there is a marginally higher effect size compared to the Falchikov and Goldfinch baseline, there is a higher correlation. This higher effect size may be down to the anonymity of the final iteration. If this is also excluded, then from the five "unblemished" examples of voting (that is to say, excluding both assignments from the 2013-14 iteration and assignment 4 from the 2011-12 iteration) , correlation (r) 0.745 and effect size (d) 0.267 is found. This, again, is a higher correlation than the Falchikov and Goldfinch average and only marginally higher in terms of the effect size.

I believe these are impressive figures, since the totals in the meta-studies may have been influenced by some kind of publication bias, whereby studies with poor correlations may not have been submitted for publication. Moreover, it is (at least within the same course) a repeatable figure, although as has been seen, some iterations have been more successful than others. As mentioned above, the clearest example is the effect size increasing in the final iteration. Both happened under conditions of anonymity of marking.

Table 11-3:Correlation and Effect Size for Tutor vs Student Marks Assignment 2 MSc Course

Year	2010-11	2011-12	2012-13	2013-14
Event	Assessed	Assessed	Assessed	Assessed
Comment				
Number of Items Evaluated	22	19	9	11
Overall Number of Evaluations	44	38	18	22
Average Number of Markers per Judgement	33.68	30.03	22.89	28.64
Overall r (Tutor average vs Student Average)	0.87	0.71	0.83	0.76
Overall d (Tutor Average vs Student Average)	0.16	0.01	0.43	0.72

Table 11-4: Correlation and Effect Size for Tutor vs Student Marks Assignment 4 MSc Course

Year	2010-11	2011-12	2012-13	2013-14
Event	Assessed	Assessed	Assessed	Assessed
Number of Items Evaluated	22	19	9	11
Overall Number of Evaluations	66	57	27	33
Overall r (Tutor average vs Student Average)	**0.69**	**0.90**	**0.84**	**0.43**
Overall d (Tutor Average vs Student Average)	**0.34**	**0.02**	**0.53**	**0.71**

With the exception of 2013-14, the correlation is high and the effect size acceptable. In 2013-14 correlation remains high for assignment 2, but falls to the lowest of all in assignment 4.

The conclusion would therefore be: Live Peer Assessment generally delivers better results than ordinary (asynchronous) peer assessment, with the proviso that students do not assess under conditions of anonymity.

11.3 What are the benefits and limitations of using LPA and EVS?

To address this question, the two modes, namely summative peer assessment and feed-forward peer assessment, need to be considered separately. In terms of feed-forward peer assessment (when marking works by a previous cohort), the primary issue relates to levels of engagement. In chapter 7's analysis of voting patterns, it was shown that when used in a purely formative context, attendance would drop, and even among the attendees, there might be inattention evidenced by no actual clicking. A potential way of stopping this would be to force students to attend, and potentially penalise any marking patterns that were completely at variance with those of the tutors. However, introducing this level of compulsion might result in students becoming passive imitators of the tutors' ways of thinking, rather than truly engaging with the desired way of thinking. Notwithstanding the clear inattention of some attendees during feed-forward events, the communicative power of the event did not appear to be diminished – as demonstrated in the opinions of the students in the focus group (chapter 8) as well as in the continuing improvement in scores for the final assignment. In the final iteration of the course, such disengagement during the marking events was much more visible than in previous years (see chapter 7 for levels of engagement measures and chapter 6 for final assignment scores).

11.3.1 What pedagogical benefits does this kind of peer assessment bring with it?

A main pedagogical benefit appears to be the internalisation of the concepts of quality. By this I also mean the ability to notice the salient in any piece of work; to recognise what needs to be corrected and what is completed. The other key benefit is that it adds to the authority of the rubric, which ceases to be "reified" to use Wenger's term and instead, becomes a living document.

11.3.2 Do students become more competent in judging quality in the work of others and in their own work?

The answer to this is yes, and the reason I believe so, is that judging quality is in some ways a tacit experience, difficult to reduce into words. In the Multimedia Specification focus group (chapter 4) we saw how students "picked up" the things the tutors liked. That is, they were looking for "tells", while

the tutors were introducing the marking as to what they really thought about the work being presented. In the BSc focus group (chapter 7) it was elicited that students were capable of judging their own work, because "it helped us to think *what we were supposed to look out for* in the actual test so really helpful." (my italics). This is the strongest argument for the use of feed-forward exercises, i.e. it begins to establish a repertoire of comparable work in the minds of the students and makes it easier for them to understand when something is or is not up to standard. In addition, it models the process of dissecting the work and finding the salient points of quality, that is to say, it models the kinds of noticing that is essential to anyone producing quality work. As has also been seen from the focus group feedback, it establishes a shared language among students such that conversations which take place long after the feed-forward event still use the terminology of the event. As one student said:

> It is when you look up at the mark sheet and what you learn in the lectures they say the same thing so when you attend the lecture and you do that with the clickers then you know that immediately when you get to those sheets that that is what the requirements are.

11.3.3 What are the limitations of the technique?

The primary limitation is that this technique can really only be used for assignments where demonstration or perusal requires five minutes or less. Asking students to evaluate things like project reports, requires them to have, indeed, *read* the project report prior to the session. If they have not, but nonetheless continue to vote, the result will either be a slavish adherence to any opinions expressed by a tutor, or randomness. If using this technique for things requiring a long time to evaluate, then some qualificatory pre-requisite (e.g. a multiple choice test asking objective questions about the thing being evaluated) might be necessary in order to only show responses from students who have actually read the work. Another limitation (in summative peer assessment mode) is the tendency of large numbers of raters to compress the range of scores between lower maxima and higher minima. This means that potentially brilliant artefacts are marked as only slightly better than others, despite the fact that the amount of thought and work that went into both being massively different.

A danger, but not so much a limitation, is that live voting unmoderated by tutors might involve expressing non-academic opinions, for instance, the likeability or otherwise of the presenters. Consequently, in a live situation, the moderator may need to express surprise or disappointment at the presence of very low marks (or very high ones) regarding an artefact that doesn't deserve them. Another issue, in a live situation, might be work that is, or is suspected of being, plagiarised. Clearly, it would be wrong to challenge a presenter, if there is only mild doubt, but equally, if a student presents in

front of the class work that is plainly beyond their capability and the class has to vote, the tutor may feel strong discomfort at having to go along with the marking of a piece of work which appears beyond the capabilities of the presenters, and opinion which may be shared by the other students in the cohort.

Finally, there is the worry that presenting exemplars of prior student work in a feed-forward event might lead students to mimic the exemplar, rather than producing work of an independent nature. However, in the chapter on "inspiration" (Chapter 9), it was found that whilst a number of students had been influenced by exemplars, only one (out of a cohort of 180 students) had slavishly attempted to reproduce the exemplar. Of course, this will also be diminished if the assignment itself requires originality and is valued accordingly. An assignment where students have to produce outcomes that are likely to be very similar, for instance, an entity relationship diagram for a database for the same online application, might well not benefit from using LPA. Moreover, any challenge that comprises some notion of a "right answer" might induce mimicry from students rather than independent thought.

These are the limitations of this technique and do need to be borne in mind when making the decision to use LPA, however, there are steps that can be taken to reduce risk, as follows:

- Only ask students to evaluate things that can be demonstrated quickly;
- Make sure the proportion of the marks awarded from student grading is a small proportion of the overall grade for that assignment;
- Try to be neutral in LPA events, except on occasions when there are aberrant marks based on non- academic considerations, which might constitute an injustice;
- Be explicit about the reason for carrying out LPA in feed-forward events – the discovery of different kinds of quality, not of finding a "model" which everyone should emulate

11.4 How does LPA Work?

This section is the most speculative regarding the research questions. It is difficult to definitively know what gives LPA its demonstrable efficacy, however, the statistics and dialogues with students do suggest some plausible explanations and also some lines for further enquiry.

11.4.1 Does the act of assessing impact on the way students' set about their own academic work?

Yes, primarily at the level of focus, self-efficacy, judgement and inspiration. By focus, I mean it affirms the status of the rubric, and makes it a living document in the eyes of students. This is a clear finding of the focus group involving the BSc E-Media Design group. They talked of ticking things off in the rubric as they did them and they also emphasised words like "appropriate", which was used in the rubric. In

terms of self-efficacy, that same group demonstrated a level of comfort in judging whether some feature of the work was done, or whether it needed more work. Moreover, they demonstrated ability to focus on the salient as opposed to the ephemeral when assessing work. In short, the students appeared to have a solid *appreciative* system established, which allowed them to judge with confidence the quality of their work. Finally, as seen in the work produced by the students on the E-Media Design course, seeing a range of work produced by previous cohorts inspired them to use some of the things they had seen and rework them into their own designs.

11.4.2 What might be the explanation for the effects claimed for LPA in this study?

What makes live peer assessment different from traditional peer assessment, whether in its feed-forward or summative format, relates to two things: its immediacy and public nature. The results of a poll are available instantaneously and the public nature of voting means that students compare their own vote with those of the rest of the class. This was noted particularly by students in the MSc focus group, who talked of changing their subsequent votes based on a comparison between their own vote and the other students' previous votes (see Chapter 4).

11.4.3 What are the students' opinion and feelings about the process?

The MSc students found the process engaging and transparent, claiming they found inspiration in what they saw in their peers' work. Some had initial concerns about being marked by their peers, which is consistent with a lot of the literature, however, these were allayed by ensuring the accountability of the markers. The undergraduate students mainly reported about having clarity regarding what was expected of them. One caveat is that both of these findings came from focus groups, which may have recruited (not intentionally) the most positive students in this regard.

11.5 Conclusion

LPA could be a driving force towards the renewal of practice based courses in Computer Science. It can help with affirming and demystifying notions of quality within a discipline. It can involve establishing a repertoire of similar cases and commentaries about them, thus enabling students to be more knowing about the quality of their own work. It has been shown in this study to result in superior levels of correlation between tutor and student marking than is present in the literature. It is also exceptionally fast, hence enabling coursework marking to be completed within a day. However, it needs to be moderated with care, and also some level of accountability has to be present for individual markers, although how that is to be achieved will depend on the course and culture of an institution. Certainly,

absolutely anonymous marking (where the identity of the marker is not only not known to the assesses, but also to the tutors themselves) does not appear to deliver good outcomes.

A variety of measures are available to test the health of any peer assessment, including: tutor - student marking correlation and effect size, Magin's reciprocity matrix, Krippendorf's alpha, counts of over marking and under marking as well as the median correlation between individual students' marks and those of the tutor(s). The choice of exemplars in training and feedforward events is important, in that care should be taken to find ones that (a) do not exhibit overly divergent responses and (b) are credible comparisons with the work being undertaken by current students.

This format is extremely suitable for things that can be judged in presentation, such as speeches, posters and multimedia artefacts and generally, as aforementioned, anything requiring less than five minutes to judge. For more extended work, or short work that has obvious "right" answers, it is not really suitable. Given the stress of being judged by one's peers in an open public situation, notwithstanding the fact that this is common in design disciplines via the "Crit", it is still preferable that this technique be applied to work undertaken by groups rather than individuals. Also, it is important that the proportion of the marks awarded by peers is small in regard to those awarded by tutors. In my experience of using LPA on an impromptu basis as feed-forward segments on other courses (not part of this study) the result is always the same – students simply do better work.

In simple terms, LPA or feed forward is one of the easiest ways to implement principle 6 of Chickering and Gamson's *Seven Principles for Good Practice in Undergraduate Education* – "Communicates High Expectations". They write:

> Expect more and you will get it. High Expectations are important for everyone - for the poorly prepared, for those unwilling to exert themselves, and for the bright and well-motivated. Expecting students to perform well becomes a self-fulfilling prophecy when teachers and institutions hold high expectations of themselves and make extra efforts. (Chickering & Gamson, 1987)

With LPA, not only do we communicate high expectations, but we demonstrate concrete manifestations of those expectations. Moreover, we model the way quality can be judged and improvements made. We give students a framework in which they themselves can be an authoritative source of feedback for their own work. And also, we do it in a public and transparent way, such that quality can be discussed

and contested, and a dialogue about it can continue long after the evaluation events and feed forward exercises have concluded.

Appendices

References

Andrade, H. G. (2001). The effects of instructional rubrics on learning to write. Current issues in education, 4.

Barwell, G., & Walker, R. (2009). Peer assessment of oral presentations using clickers: the student experience.

Beck, K. (2000). Extreme programming explained: embrace change: addison-wesley professional.

Bennett, S., & Barker, T. (2011). The Use of Electronic Voting to Encourage the Development of Higher Order Thinking Skills in Learners. Paper presented at the Proceedings of the International Conference on Computers and Assessment CAA 2011.

Bennett, S., Barker, T., Thomas, P., & Lilley, M. (2015). Modelling and Motivating High Quality Academic Work With Live Peer Evaluation. Paper presented at the ECEL2015-14th European Conference on e-Learning: ECEl2015.

Biggs, J. (1999). What the student does: Teaching for quality learning at university. Buckingham. Open University Press.

Bloor, M. (2001). Focus groups in social research: Sage.

Bloxham, S., & West, A. (2004). Understanding the rules of the game: marking peer assessment as a medium for developing students' conceptions of assessment. Assessment & Evaluation in Higher Education, 29(6), 721-733.

Blythman, M., Orr, S., & Blair, B. (2007). Critiquing the crit. Brighton: Art, Design and Media Subject Centre.

Bostock, S. (2000). Student peer assessment. Learning Technology.

Boud, D. J., & Tyree, A. (1980). Self and peer assessment in professional education: a preliminary study in law. J. Soc't Pub. Tchrs. L. ns, 15, 65.

Brown, S., & Knight, P. (1994). Assessing learners in higher education: Psychology Press.

Bruffee, K. A. (1984). Collaborative learning and the" Conversation of Mankind". College English, 46(7), 635-652.

Carbone, A., Lynch, K., Barnden, A., & Gonsalvez, C. (2002). Students' reactions to a studio-based teaching and learning philosophy in a three year IT degree. Paper presented at the Proceedings of the 2002 Annual International Conference of the Higher Education Research and Development Society of Australasia.

Carter, A. S., & Hundhausen, C. D. (2011). A review of studio-based learning in computer science. Journal of Computing Sciences in Colleges, 27(1), 105-111.

Cheng, W., & Warren, M. (1997). Having second thoughts: student perceptions before and after a peer assessment exercise. Studies in Higher Education, 22(2), 233-239.

Cheng, W., & Warren, M. (2005). Peer assessment of language proficiency. Language Testing, 22(1), 93-121.

Cohen, J. (1992). Statistical power analysis. Current directions in psychological science, 1(3), 98-101.

De Vaus, D. (2002). Analyzing social science data: 50 key problems in data analysis: Sage.

D'Errico, J. (2014). Greenness of an RGB image. Retrieved from http://uk.mathworks.com/matlabcentral/answers/119804-greenness-of-an-rgb-image.

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. Studies in Higher Education, 24(3), 331-350.

Estey, A., Long, J., Gooch, B., & Gooch, A. A. (2010). Investigating studio-based learning in a course on game design. Paper presented at the Proceedings of the Fifth International Conference on the Foundations of Digital Games.

Fairbrother, R., & Black, P. et Gill, P.(eds.)(1995) Teachers Assessing Pupils: Lessons from Science Classrooms. Hatfield UK: Association for Science Education.

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. Review of educational research, 59(4), 395-430.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. Review of educational research, 70(3), 287-322.

Forde, K. (1996). The effects of gender and proficiency on oral self-and peer-assessments. English Language Studies Working Papers, City University of Hong Kong, 1(1), 34-47.

Gaillet, L. L. (1992). A Foreshadowing of Modern Theories and Practices of Collaborative Learning: The Work of Scottish Rhetorician George Jardine.

Gelotte, K. (2011). Image Color Extract. Retrieved from http://www.coolphptools.com/color_extract

Gibbs, G., & Simpson, C. (2004). Does your assessment support your students' learning. Journal of Teaching and Learning in Higher Education, 1(1), 1-30.

Gielen, S., Dochy, F., & Onghena, P. (2011). An inventory of peer assessment diversity. Assessment & Evaluation in Higher Education, 36(2), 137-155.

Grove, J. (2014). National Student Survey 2014 results show record levels of satisfaction. Retrieved from https://www.timeshighereducation.com/news/national-student-survey-2014-results-show-record-levels-of-satisfaction/2015108.article

Hamer, J., Kell, C., & Spence, F. (2007). Peer assessment using aropä. Paper presented at the Proceedings of the ninth Australasian conference on Computing education-Volume 66.

Hanrahan, S. J., & Isaacs, G. (2001). Assessing self-and peer-assessment: The students' views. Higher education research and development, 20(1), 53-70.

Hendrix, D., Myneni, L., Narayanan, H., & Ross, M. (2010). Implementing studio-based learning in CS2. Paper presented at the Proceedings of the 41st ACM technical symposium on Computer science education.

Hendry, G. D., Armstrong, S., & Bromberger, N. (2012). Implementing standards-based assessment effectively: Incorporating discussion of exemplars into classroom teaching. Assessment & Evaluation in Higher Education, 37(2), 149-161.

Hendry, G. D., White, P., & Herbert, C. (2016). Providing exemplar-based 'feedforward'before an assessment: The role of teacher explanation. Active Learning in Higher Education, 17(2), 99-109.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. Studies in Higher Education, 39(10), 1774-1787.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. Educational research review, 2(2), 130-144.

Korman, M., & Stubblefield, R. (1971). Medical shcool evaluation and internship performance. Academic Medicine, 46(8), 670-673.

Kuhn, S. (2001). Learning from the architecture studio: Implications for project-based pedagogy. International Journal of Engineering Education, 17(4/5), 349-352.

Lewis, R. K. (1998). Architect. A Candid Guide to the Profession (Revised edition) MIT Press, Cambridge, Mass.

Li, H., Xiong, Y., Zang, X., L. Kornhaber, M., Lyu, Y., Chung, K. S., & K. Suen, H. (2015). Peer assessment in the digital age: a meta-analysis comparing peer and teacher ratings. Assessment & Evaluation in Higher Education(ahead-of-print), 1-20.

Liu, C. C., & Tsai, C. M. (2005). Peer assessment through web - based knowledge acquisition: tools to support conceptual awareness. Innovations in Education and Teaching International, 42(1), 43-59.

Lu, R., & Bol, L. (2007). A comparison of anonymous versus identifiable e-peer review on college student writing performance and the extent of critical feedback. Journal of Interactive Online Learning, 6(2), 100-115.

Lynch, K., Carbone, A., Arnott, D., & Jamieson, P. (2002). A studio-based approach to teaching information technology. Paper presented at the Proceedings of the Seventh world conference on computers in education conference on Computers in education: Australian topics-Volume 8.

Magin, D. (1993). Should student peer ratings be used as part of summative assessment. Research and Development in Higher Education, 16, 537-542.

Magin, D. (2001). Reciprocity as a source of bias in multiple peer assessment of group work. Studies in Higher Education, 26(1), 53-63.

Miller, C. M., & Parlett, M. (1974). Up to the Mark: A Study of the Examination Game.

Morton, J., & Macbeth, W. (1977). Correlations between staff, peer and self assessments of fourth - year students in surgery. Medical Education, 11(3), 167-170.

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. Practical assessment, research & evaluation, 7(10), 71-81.

Narayanan, N. H., Hundhausen, C., Hendrix, D., & Crosby, M. (2012). Transforming the CS classroom with studio-based learning. Paper presented at the Proceedings of the 43rd ACM technical symposium on Computer Science Education.

Orpen, C. (1982). Student versus lecturer assessment of learning: a research note. Higher Education, 11(5), 567-572.

Orsmond, P., Merry, S., & Reiling, K. (1996). The importance of marking criteria in the use of peer assessment. Assessment & Evaluation in Higher Education, 21(3), 239-250.

Pachur, T., & Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. Cognitive Psychology, 65(2), 207-240.

Panadero, E., Romero, M., & Strijbos, J.-W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. Studies in educational evaluation, 39(4), 195-203.

Percy, C. (2004). Critical absence versus critical engagement. Problematics of the crit in design learning and teaching. Art, Design & Communication in Higher Education, 2(3), 143-154.

Petřík, J. (2011-2017). JPEXS Free Flash Decompiler. Retrieved from https://www.free-decompiler.com/flash/download/

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., & Koller, D. (2013). Tuned models of peer assessment in MOOCs. arXiv preprint arXiv:1307.2579.

Pollitt, A. (2012). The method of adaptive comparative judgement. Assessment in Education: Principles, Policy & Practice, 19(3), 281-300.

Pond, K., Ul‐Haq, R., & Wade, W. (1995). Peer review: a precursor to peer assessment. Programmed Learning, 32(4), 314-323.

Pryor, J., & Lubisi, C. (2002). Reconceptualising educational assessment in South Africa—testing times for teachers. International Journal of Educational Development, 22(6), 673-686.

Raes, A., Vanderhoven, E., & Schellens, T. (2015). Increasing anonymity in peer assessment by using classroom response technology within face-to-face higher education. Studies in Higher Education, 40(1).

Räihä, K.-J., Ovaska, S., & Ferro, D. (2008). Observations on Peer Evaluation using Clickers. IxD&A, 3, 127-134.

Reimer, Y. J., Cennamo, K., & Douglas, S. A. (2012). Emergent themes in a UI design hybrid-studio course. Paper presented at the Proceedings of the 43rd ACM technical symposium on Computer Science Education.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. Oxford Review of Education, 13(2), 191-209.

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. Assessment & Evaluation in Higher Education, 34(2), 159-179.

Sadler, D. R. (2009). Transforming holistic assessment and grading into a vehicle for complex learning Assessment, learning and judgement in higher education (pp. 1-19): Springer.

Sadler, D. R. (2013a). Making competent judgments Of competence. In O. Troitschanskaia, C. Kuhn, & J. Fege (Eds.), Modeling and measuring competencies in higher education: Tasks and challenges (pp. 13-27).

Sadler, D. R. (2013b). Opening up feedback. Reconceptualising feedback in higher education: Developing dialogue with students, 54(4), 12.

Sadler, P. M., & Good, E. (2006). The impact of self-and peer-grading on student learning. Educational assessment, 11(1), 1-31.

Sambell, K., McDowell, L., & Brown, S. (1997). "But is it fair?": an exploratory study of student perceptions of the consequential validity of assessment. Studies in educational evaluation, 23(4), 349-371.

Schön, D. A. (1983). The reflective practitioner: How professionals think in action (Vol. 5126): Basic books.

Schumacher, C. F. (1964). A Factor-Analytic Study of Various Criteria of Medical Student Accomplishment. Academic Medicine, 39(2), 192-196.

Seery, N., Canty, D., & Phelan, P. (2012). The validity and value of peer assessment using adaptive comparative judgement in design driven practical education. International Journal of Technology and Design Education, 22(2), 205-226.

Sitthiworachart, J., & Joy, M. (2004). Effective peer assessment for learning computer programming. Paper presented at the ACM SIGCSE Bulletin.

Sluijsmans, D. M., Moerkerke, G., Van Merrienboer, J. J., & Dochy, F. J. (2001). Peer assessment in problem based learning. Studies in educational evaluation, 27(2), 153-173.

Smith, C. (2011). Understanding Students' Views of the Crit Assessment. Journal for Education in the Built Environment, 6(1), 44-67.

Stefani, L. A. (1992). Comparison of collaborative self, peer and tutor assessment in a biochemistry practical. Biochemical education, 20(3), 148-151.

Stiggins, R.J. (1987). Design and development of performance assessments. Educational Measurement: Issues and Practice, 6, 33-41.

Swanson, D. B., Case, S. M., & van der Vleuten, C. P. (1991). Strategies for student assessment. The challenge of problem-based learning, 260-273.

Topping, K. (1998). Peer assessment between students in colleges and universities. Review of educational research, 68(3), 249-276.

Van Zundert, M., Sluijsmans, D., & Van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. Learning and Instruction, 20(4), 270-279.

Vanderhoven, E., Raes, A., Montrieux, H., Rotsaert, T., & Schellens, T. (2015). What if pupils can assess their peers anonymously? A quasi-experimental study. Computers & Education, 81, 123-132.

Weekley, J. A., & Gier, J. A. (1989). Ceilings in the reliability and validity of performance ratings: The case of expert raters. Academy of Management journal, 32(1), 213-222.

Wenger, E. (1999). Communities of practice: Learning, meaning, and identity: Cambridge university press.

Williams, E. (1992). Student attitudes towards approaches to learning and assessment. Assessment and evaluation in higher education, 17(1), 45-58.

Wimshurst, K., & Manning, M. (2013). Feed-forward assessment, exemplars and peer marking: evidence of efficacy. Assessment & Evaluation in Higher Education, 38(4), 451-465.

Xiao, Y., & Lucking, R. (2008). The impact of two types of peer assessment on students' performance and satisfaction within a Wiki environment. The Internet and Higher Education, 11(3), 186-193.

Zucconi, A. (2015). The incredibly challenging task of sorting colours. Retrieved from http://www.alanzucconi.com/2015/09/30/colour-sorting/

(2007). "Re-engineering Assessment Practices in Higher Education." Retrieved 20/2/2017, 2017, from http://www.reap.ac.uk.

(2016). "National Student Survey: student satisfaction climbs to 84% with record response rate." Retrieved 20/02/2017, 2017, from https://www.ucl.ac.uk/news/students/082016/082016-11082016-national-student-survey-2016-results.

(2017). "Poll Everywhere Home Page." Retrieved 20/2/2017, 2017, from https://www.polleverywhere.com/.

Ethics Approvals

Ethics Protocol: 1112/51

UNIVERSITY OF HERTFORDSHIRE

FACULTY OF SCIENCE, TECHNOLOGY AND CREATIVE ARTS

M E M O R A N D U M

TO **Trevor Barker**

C/C **N/A**

FROM **Dr Simon Trainis – Chair, Faculty Ethics Committee**

DATE 19 January 2012

Your Ethics application for your project entitled:

Using EVS and Peer Assessment

has been granted approval and assigned the following Protocol Number:

1112/51

This approval is valid:

From 19 January 2012

Until 1 June 2012

If it is possible that the project may continue after the end of this period, you will need to resubmit an application in time to allow the case to be considered

Ethics Protocol: COM/SF/UH/00014

UNIVERSITY OF HERTFORDSHIRE SCIENCE AND TECHNOLOGY

M E M O R A N D U M

TO **Steve Bennett**

CC **Trevor Barker**

FROM **Dr Simon Trainis, Science and Technology ECDA Chairman**

DATE **08/01/14**

Protocol number: COM/SF/UH/00014

Title of study: Feed Forward Evaluation

Your application for ethical approval has been accepted and approved by the ECDA for your school.

This approval is valid:

From: 08/01/14

To: 01/07/14

Please note:

Approval applies specifically to the research study/methodology and timings as detailed in your Form EC1. Should you amend any aspect of your research, or wish to apply for an extension to your study, you will need your supervisor's approval and must complete and submit form EC2. In cases where the amendments to the original study are deemed to be substantial, a new Form EC1 may need to be completed prior to the study being undertaken.

Rubrics

2010T Rubric

Using LIKERT scales provided, you are required to rate the websites according to the following criteria:

(A) Has correct number of pages with correct headings on each

Yes	Maybe	No
1	2	3

(B) Correct background colour

Yes	Maybe	No
1	2	3

(C) Correct width and height of the Flash file

Yes	Maybe	No
1	2	3

(D) Correct number of buttons with correct colours for them

Yes	Maybe	No
1	2	3

(E) Buttons navigate to correct frames

Yes	Maybe	No
1	2	3

(F) Contains at least two images

Yes	Maybe	No
1	2	3

(G) Small file-size

Yes	Maybe	No
1	2	3

(H) Motion Tweening Of Position of Things in the animation welcome screen OR Motion Tweening of Visibility of Things (for fade ins and fade outs) in the animation welcome screen

Yes	Maybe	No
1	2	3

(I) Correct and Nice Positioning of Buttons and Content

Yes	Maybe	No
1	2	3

(J) Good easy on the eye content (text and image) - not too little, not to much and all relevant

Yes	Maybe	No
1	2	3

(K) Button clicks have small sounds associated with them

Yes	Maybe	No
1	2	3

(L) Some very clever and visually elegant animation (using Shape Tweens, or Motion Guides or Masking) in the animation welcome screen

Yes	Maybe	No
1	2	3

(M) Extremely well positioned and pleasant looking buttons

Yes	Maybe	No
1	2	3

(N) Extremely well judged content

Yes	Maybe	No
1	2	3

(O) An immediately visible and functioning background music toggle

Yes	Maybe	No
1	2	3

(greyed out criteria indicate elements that were marked by the tutor but were not able to be used in the labs)

Marking scheme

For each website you will receive 1 mark for a 'correct' rating of the category.

Website 1 15 marks

Website 2 15 marks

Website 3 15 marks

Website 4 15 marks

Total 60 Marks

The 'correct' rating will be measured by how close your rating is to the tutors' ratings of the website.

FILES EVALUATED: Final

final	rehearsal
1. 10.swf 2. 11.swf 3. 12.swf 4. 13.swf	1. 1.swf 2. 2.swf 3. 3.swf 4. 4.swf

2010S Rubric

Criteria Used in Final (all yes/no/maybe):

1. Has correct number of pages with correct headings on each
2. Correct background colour
3. Correct number of buttons with correct colours for them
4. Make buttons navigate to correct frames using simple action script
5. Contains at least two images of you
6. Motion Tweening Of Position of Things in the animation welcome screen OR Motion Tweening of Visibility of Things (for fade ins and fade outs) in the animation welcome screen
7. Correct and Nice Positioning of Buttons and Content
8. Good easy on the eye content (text and image) not too little not to much and all relevant
9. Button clicks have small sounds associated with them
10. Some very clever and visually elegant animation (using Shape Tweens or Motion Guides or Masking) in the animation welcome screen
11. Extremely well positioned and pleasant looking buttons
12. Extremely well judged content
13. An immediately visible and functioning background music toggle

Tutor Rubric for that year: (bolded things not marked by students in test)

- **Q1 Publish an SWF file and upload it to Studynet**
- Q2 Has correct number of pages with correct headings on each
- Q3 Correct background colour
- **Q4 Correct width and height of the Flash file**
- Q5 Correct number of buttons with correct colours for them
- Q6 Make buttons navigate to correct frames using simple action script
- Q7 Contains at least two images of you
- **Q8 Small file-size**
- Q9 Motion Tweening Of Position of Things in the animation welcome screen OR Motion Tweening of Visibility of Things (for fade ins and fade outs) in the animation welcome screen
- Q10 Correct and Nice Positioning of Buttons and Content
- Q11 Good easy on the eye content (text and image) not too little not to much and all relevant
- Q12 Button clicks have small sounds associated with them
- Q13 Some very clever and visually elegant animation (using Shape Tweens or Motion Guides or Masking) in the animation welcome screen
- Q14 Extremely well positioned and pleasant looking buttons
- Q15 Extremely well judged content
- Q16 An immediately visible and functioning background music toggle

(YES/NO/MAYBE for all)

2011TE Rubric

Has correct number of screens with correct headings on each

- No headings/Way too few viewable screens
- Many Incorrect headings/Some screens missing
- Some wrong headings but all screens there
- A few problems [visual/spelling] but mostly OK
- All OK

Correct background colour

- Background colour completely wrong and bad to look at
- Background colour completely wrong but ok to look at
- Many screens wrong color but some have the right colour
- Good attempt at most screens but a few not correct
- All OK

Correct width and height of the flash file

- Totally wrong size
- Right size but screen elements don't fit
- Seems that screen elements weren't really designed for this size
- A few minor issues with resizing
- Everything OK

Correct number of buttons with correct colours for them

- No buttons
- Some buttons but no real attempt to follow the brief in their design
- Wrong number of buttons – or wrong colours – but tried to follow brief
- Almost correct – just a few problems
- All OK

Make buttons navigate to correct frames using simple action script

- No navigation
- Lots of problems
- Some buttons navigate well
- Most buttons navigate well/minor issues

Contains at least two images of you

- No images
- Poor image
- Just one good image

- Two images but some problems
- All OK

Small file-size

- Vast file size (>x10)
- File too large (x5)
- Not bad (X2)
- Just over (<x2)
- OK

Motion Tweening Of Position of Things in the animation welcome screen OR Motion Tweening of Visibility of Things (for fade ins and fade outs) in the animation welcome screen

- No animation
- Animation but very simple
- Fair animation but not well performed
- Quite good attempt showing good technique
- Good animation with good

Layout and Positioning of Buttons and Content

[Button Problems: inconsistent size of buttons, inconsistent fonts/styes, inconsistent margins/spacing/text color, positional juddering between screens] [Content Problems: too much/too little/inconsistency of fonts/styles/margins between screens, positional juddering between screens]

- No buttons / content
- Poorly laid out buttons or other major problem with content
- Buttons and/or content competently laid out but not visually attractive
- Buttons and content quite well laid out but maybe lacking coherence
- Very well laid out buttons and content

Well designed and appropriate content (text and image)

- Poor content with low quality images
- Images poor OR content poor
- Content and images fair
- Content OR images are GOOD
- Content AND images GOOD

Button clicks have small sounds associated with them

- No sound
- Sound almost imperceptible
- Inappropriate sound

- Most buttons have appropriate sounds
- All buttons have appropriate sounds

Some very clever and visually elegant animation (using Shape Tweens, or Motion Guides or Masking) in the animation welcome screen

- Poor or absent
- Fair
- Satisfactory
- Good
- Excellent

Buttons: Extremely well positioned, elegant, and suitable

- Poor or absent
- Fair
- Satisfactory
- Good
- Excellent

Content: text relevant, appropriate length and evenly distributed across cv, images appropriate to cv, images of very high quality

- Poor or absent
- Fair
- Satisfactory
- Good
- Excellent

An immediately visible and functioning background music toggle

- No toggle
- Toggle not immediately visible
- Not on all screens/Not functionally perfect
- On most screens but functionally perfect
- On all screens and functionally perfect

2011 Rehearsal (no width and height)

1. Has correct number of screens with correct headings on each
2. Has Correct Background Color
3. Correct number of buttons with correct colours for them
4. Make buttons navigate to correct frames using simple action script
5. Contains at least two images of you
6. Small file-size
7. Motion Tweening Of Position/Visibility in the welcome screen
8. Layout and Positioning of Buttons and Text
9. Choice of material
10. Button clicks have small sounds associated with them
11. Has either very clever or visually elegant animation
12. Contains some very well positioned
13. The text is relevant
14. An immediately visible and functioning background music toggle

2011 TS Rubric

1. Has correct number of screens with correct headings on each

- No headings/Way too few viewable screens
- Many Incorrect headings/Some screens missing
- Insufficiently Prominent or Bad Spelling
- A few problems [visual/spelling] but mostly OK
- All OK

2. Has Correct Background Color (Yellow)

- Background colour completely wrong and bad to look at
- Background colour completely wrong but ok to look at
- Background colour correct but occupies too small a percentage of the space
- Good attempt at most screens but a few not correct
- All OK

3. Correct Width And Height of the Flash Flash File (600x300)

- Totally Wrong Size
- Right size, but screen elements don't fit
- Seems that screen elements weren't designed for this size
- A few minor issues with resizing
- All OK

4. Correct number of buttons with correct colours for them (purple with white text)

- No buttons
- Some buttons but no real attempt to follow the brief in their design
- Wrong number of buttons – or wrong colours – but tried to follow brief
- Almost correct – just a few problems
- All OK

5. Buttons navigate to correct frames using simple action script

- No navigation
- Some wrong navigation
- Navigate correctly but often with problematic transitions
- Navigate correctly but sometimes with problematic transitions
- All OK

6. Contains at least two images of you

- No images
- Has a very poor image
- Has only one good image
- Two images but some problems
- All OK

7. Small file-size (Less than 200k)

- Vast file size (>x10)
- File too large (x5)
- Not bad (X2)
- Just over (<x2)
- OK

8. Motion Tweening Of Position/Visibility in the welcome screen

- No animation
- Animation but very simple
- Evidence of technique but does not look good
- Quite good attempt showing good technique
- Good animation with good technique

9. Layout and Positioning of Buttons and Text

- No buttons, or laid out so badly and inconsistently it is disorienting to navigate
- Poorly laid out buttons /not appear altogether on same screen, or other major problem
- Buttons and/or content competently laid out but not visually attractive
- Buttons and content quite well laid out but maybe lacking coherence
- Well laid out buttons and content

10. Choice of material, text and tone appropriate for a CV (text and image)

- Poor content with low quality images
- Images poor OR Content poor
- Content and images Average
- Content OR images are GOOD
- Content AND images GOOD

11. Button clicks have small sounds associated with them

- No sound
- Sound almost imperceptible
- Inappropriate sound
- Most buttons have appropriate sounds
- All buttons have appropriate sounds

12. Has either very clever or visually elegant animation

- Strongly Disagree
- Disagree
- Not sure
- Agree
- Strongly Agree

13. Extremely Well Positioned and Pleasant Looking Buttons

- Strongly Disagree
- Disagree
- Not sure
- Agree
- Strongly Agree

14. Extremely Well Judged Content The text is relevant, of appropriate length and evenly distributed across the cv, images are appropriate to cv and of a high quality

- Strongly Disagree
- Disagree
- Not sure
- Agree
- Strongly Agree

15. An immediately visible and functioning background music toggle

- No toggle
- Toggle not immediately visible
- Not on all screens/only plays per screen
- Abrupt transitions between loops
- On all screens and functionally perfect

2011TM RUBRIC

BASICS (40 – 49 marks - attain satisfactory achievement)

- Publish an SWF file and upload it to Studynet (5)
- Has correct number of pages with correct headings on each (5)
- Correct background colour (5)
- Correct width and height of the Flash file (5)
- Correct number of buttons with correct colours for them (5)
- Make buttons navigate to correct frames using simple action script (5)
- Contains at least two images of you (5)
- Small file-size (5)
- Motion Tweening Of Position of Things in the animation welcome screen OR Motion
- Tweening of Visibility of Things (for fade ins and fade outs) in the animation welcome
- screen (5)
- Correct and Nice Positioning of Buttons and Content (5)

INTERMEDIATE (50 - 69 marks - attain good to very good achievement)

- Good easy on the eye content (text and image) - not too little, not to much and all relevant
- (10)
- Button clicks have small sounds associated with them (10)

ADVANCED (70 – 100 to attain excellent to outstanding achievement)

- Some very clever and visually elegant animation (using Shape Tweens, or
- Motion Guides or Masking) in the animation welcome screen (10)
- Extremely well positioned and pleasant looking buttons (5)
- Extremely well judged content (5)
- An immediately visible and functioning background music toggle (10)

2012S Rubric

1. Has all required screens (welcome, details, hobbies, employment, education) and they are always accessible by button navigation

- Does not have all screens
- Has all screens but not all always accessible
- All screens but very large variations in location of heading
- Some variations in location of heading
- Each screen clearly headed by correct word at the same spot

2. Sufficient contrast in colours

- Some pages unreadable because of colours
- Some illegibility because of colour problems
- Poor contrast in colours or large variations in contrast
- Mostly clear contrast in colours
- Clear contrast in colours

3. Buttons are readable and have sounds

- Buttons are unreadable and have no sound
- Buttons have sound but lack readability
- Buttons are readable but none have sound
- Buttons are all readable but some may lack sound
- All buttons are readable and all have appropriate sounds

4. Good grammar and spelling and use of language

- Many glaring spelling or grammar errors
- A glaring spelling or grammar error
- Minor spelling errors
- Minor grammar errors
- No spelling/grammar errors

5. Aligned and Uniform Sized Buttons

- Extremely poor buttons
- Chaotic arrangements of buttons across and within screens
- Not all buttons visible on all screens or alignment issues
- Some small alignment, spacing or text size issues
- Perfect alignment, spacing and text size on all screens

6. Text of appropriate length and evenly distributed across cv (ignore questions of suitability of text here)

- Clearly insufficient text for the purpose
- Only the bare minimum of text
- Sufficient but unevenly distributed

- Mostly evenly distributed
- All text evenly distributed of appropriate length

7. Animation demonstrates meaning and visual appeal

- Has no animation
- Has only the most basic animation which lacks meaning
- Displays some creativity but lacks meaning
- Is meaningful but lacks visual appeal
- Is both meaningful and visually appealing

8. Layout is harmonious, regular and consistent

- All screens harmonious and balanced
- Some minor disharmonies on screens
- Small positional twitches between screens
- Some big disharmony or inconsistency between screens
- Very inharmonious or random screens

9. A functioning widget for turning music on and off independent of navigational functionality

- No music widget
- Music widget has obvious and immediately apparent flaws
- Has flaws but not immediately apparent
- On most screens but functionally perfect
- On all screens and functionally perfect

2012TM Rubric

Marking Scheme

Satisfactory work 40% - 49%

- Publish an SWF file of under 150k and upload it to Studynet (4)
- Has correct number of screens with correct headings on each (4)
- Appropriate choice of screen colour – providing good contrast (4)
- Correct width and height of the Flash file (4)
- Correct number of buttons with good colour selection (4)
- All buttons navigate to the correct frame script (4)
- Contains at least two images of you (4)
- Good spelling and use of language (4)
- An animation in the welcome screen (4)
- Aligned and Uniform sized Buttons (4)
- Text content is relevant and expressive and compact (5)
- Buttons have appropriate sounds on click event (5)

Good work 50% - 59%

- Images of self show high production values (5)
- Text and image presented well on the screen (5)

Very good work 60% - 69%

- Animation demonstrates originality and visual appeal (10)

Excellent work 70% - 100%

- Background music is appropriate and is controllable by user (10)
- The CV is suitable for being viewed by someone who might give you a job (10)
- The CV's design expresses a kind of brand identity of you as a person (10)

2013S Rubric

1. Publish an SWF file of under 250k and upload it to Studynet

- <250k
- 250k-300k
- 300k-350k
- >350k

2. Has correct number of screens with correct headings on each

- All screens correct headings
- Not exact but meaningful headings
- Odd Headings

3. Appropriate choice of screen colour – providing good contrast

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

4. Correct width and height of the Flash file

- Yes
- Proportions The Same, but Not Size
- No

5. Correct number of buttons with good colour selection

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

6. All buttons navigate correctly

- Yes
- Mostly
- No

7. Contains at least two images of you

- Yes
- Yes But Poor Quality
- Only 1 Image

- No Images

8. Good spelling and use of language

- Perfect
- Some imperfections
- Many imperfections

9. An animation in the welcome screen

- Yes
- No

10. Aligned and Uniform sized Buttons

- Yes
- Slightly imperfect
- Very imperfect

11. Text content is relevant and expressive and compact

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

12. Buttons have appropriate sounds on click event

- Yes
- Have Sounds but Not Appropriate
- Some Missing Sounds
- No Sounds

13. Images of self show high production values

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

14. Text and image presented well on the screen

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

15. Animation demonstrates originality and visual appeal

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

16. Background music is appropriate and is controllable by user

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

17. The CV is suitable for being viewed by someone who might give you a job

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

18. The CV's design expresses a kind of brand identity of you as a person

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

2013T Rubric

Satisfactory work 40% - 49%

- Publish an SWF file of under 150k and upload it to Studynet (4)
- Has correct number of screens with correct headings on each (4)
- Appropriate choice of screen colour – providing good contrast (4)
- Correct width and height of the Flash file (4)
- Correct number of buttons with good colour selection (4)
- All buttons navigate to the correct frame script (4)
- Contains at least two images of you (4)
- Good spelling and use of language (4)
- An animation in the welcome screen (4)
- Aligned and Uniform sized Buttons (4)
- Text content is relevant and expressive and compact (5)
- Buttons have appropriate sounds on click event (5)

Good work 50% - 59%

- Images of self show high production values (5)
- Text and image presented well on the screen (5)

Very good work 60% - 69%

- Animation demonstrates originality and visual appeal (10)

Excellent work 70% - 100%

- Background music is appropriate and is controllable by user (10)
- The CV is suitable for being viewed by someone who might give you a job (10)
- The CV's design expresses a kind of brand identity of you as a person (10)

List of Publications Related to Work Done in this Thesis

Barker, T., & Bennett, S. (2011). Marking complex assignments using peer assessment with an electronic voting system and an automated feedback tool. *International Journal of e-Assessment*.

Bennett, S., & Barker, T. (2011). *The Use of Electronic Voting to Encourage the Development of Higher Order Thinking Skills in Learners.* Paper presented at the Proceedings of the International Conference on Computers and Assessment CAA 2011.

Bennett, S., & Barker, T. (2012a). *Live peer marking for HCI design education.* Paper presented at the Teaching, Assessment and Learning for Engineering (TALE), 2012 IEEE International Conference on.

Bennett, S., & Barker, T. (2012b). Using Peer Assessment and Electronic Voting to Improve Practical Skills in Masters Students. *Proceedings of the 11th European Confernece on e-Learning: ECEL*, 53.

Bennett, S., Barker, T., & Lilley, M. (2014). An EVS Clicker Based Hybrid Assessment to Engage Students with Marking Criteria. *International Association for Development of the Information Society*.

Bennett, S., Barker, T., & Lilley, M. (2016). An EVS Clicker-Based Assessment for Radical Transparency in Marking Criteria *Competencies in Teaching, Learning and Educational Leadership in the Digital Age* (pp. 183-195): Springer.

Bennett, S., Barker, T., Thomas, P., & Lilley, M. (2015). *Modelling and Motivating High Quality Academic Work With Live Peer Evaluation.* Paper presented at the ECEL2015-14th European Conference on e-Learning: ECEl2015.