# UNIVERSITY OF LONDON THESIS

Degree PhD Year 2006 Name of Author WILSON Gillian May

# THE RELATIONSHIP BETWEEN MEDIA QUALITY AND USER COST IN NETWORKED MULTIMEDIA APPLICATIONS

Gillian May Wilson

A dissertation submitted in partial fulfilment
of the requirements for the degree of

**Doctor of Philosophy**
**of the**
**University of London**

Department of Computer Science
University College London

Submitted December 2005
Amended May 2006





ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

I declare that the work presented in this thesis is my own.

Date.....29th September 2006.....

# Acknowledgements

I would firstly like to thank my supervisor, Angela Sasse, for the unwavering support, encouragement and guidance she has given me. I have had many happy times and amazing experiences as part of this PhD and for the opportunities it has given me I am extremely grateful. Thanks also to BT for the financial support of the PhD and to my industrial supervisor, Ian Henning.

Many thanks to colleagues and friends at UCL: Anna Conniff, my mentor Anna Bouch, Daniel Bruneau, Maia Garau, Jens Riegelsberger, Piers O'Hanlon and especially John McCarthy for his help with statistics. Thanks also to UCL's Glasgow partners on the ETNA project: Anne Anderson, Jim Mullin, Matthew Jackson, Lucy Smallwood and Rachel McEwan. More recent thanks to Phil Bassett for his technical assistance. Personal thanks to Dr Casey for being there and supporting me during this journey.

Most importantly, I am indebted to my wonderful family: my Mum, Dad, and sisters, Caroline and Jennifer. Thank you all for the support you have given me and for believing in me. Without your encouragement, practical help and love, this thesis would never have been completed.

**For my parents**

# Abstract

The research reported in this thesis assesses the impact of media quality degradations in Internet multimedia conferencing on users. Low quality audio and video can be experienced, therefore it is important to determine the minimum levels of quality needed to perform specific tasks. This has most commonly been investigated using subjective measures, however the research reported in this thesis adopted a 3-factor evaluation framework of task performance, user satisfaction and user cost. User satisfaction was measured subjectively, whereas physiological indicators of perceptual strain were utilised to measure user cost. Physiological measures provide continuous data throughout a session, are not subject to cognitive mediation and taking such measurements does not interfere with the user's task.

Five experiments were performed investigating audio and video quality degradations. With the exception of one passive listening task, all tasks used were based on remote interviews, as they fully exploit the capabilities of the application. Results showed that physiological responses to media quality degradations can be detected in passive, perceptual tasks. However, active participation in a task made it more difficult to detect changes due to quality degradations. In all experiments physiological measures gave information on the nature of the tasks being performed and effects of variables such as order. The results of this research were then used in three further experiments in the areas of VR and web quality of service and design.

In conclusion, the physiological measures utilised in the research reported in this thesis can be employed to assess the impact of media quality degradations in passive perceptual tasks and to give general information about the nature of the task being performed.

# Table of contents

# List of tables

# List of figures

# List of abbreviations and acronyms

| Acronym | Meaning |
|---|---|
| ANS | Autonomic Nervous System |
| bpm | Beats per minute |
| BVBA | Bad video and good audio quality |
| BVGA | Bad video and good audio quality |
| BVP | Blood volume pulse |
| CIF | Common Image Format |
| CNS | Central Nervous System |
| DVI | Digital Video Interactive |
| ECG | Electrocardiogram |
| EEG | Electroencephogram |
| EMG | Electromyogram |
| ETNA | Evaluation Taxonomy for Networked Multimedia Applications |
| fps | Frames per second |
| GSM | Global System for Mobile Communications |
| GVBA | Good video and bad audio quality |
| GVGA | Good video and good audio quality |
| HCI | Human Computer Interaction |
| HR | Heart rate |
| HRV | Heart rate variability |
| IBI | Inter-beat interval |
| ITU | International Telecommunications Union |
| LPC | Linear Predictive Coding |
| MIT | Massachusetts Institute of Technology |

| | |
|---|---|
| MMC | Multimedia conferencing |
| MOS | Mean Opinion Score |
| ms | microsiemens |
| OR | Orienting response |
| PC | Personal Computer |
| PCM | Pulse Code Modulation |
| PNS | Parasympathetic nervous system |
| POMS | Profile of Mood States |
| QCIF | Quarter Common Image Format |
| QoS | Quality of Service |
| QUASS | Quality Assessment Slider |
| RAT | Robust audio tool |
| SBP | Systolic blood pressure |
| SC | Skin conductance |
| SCIF | Super Common Image Format |
| SCL | Skin conductance level |
| SCR | Skin conductance response |
| SNS | Sympathetic nervous system |
| SSCQE | Single Stimulus Continuous Quality Evaluation |
| ST | Skin Temperature |
| UCL | University College London |
| VE | Virtual Environment |
| vic | VIdeoConferencing tool |
| VMC | Video Mediated Communication |
| VR | Virtual Reality |

# Chapter 1 Introduction

## 1.1 Chapter Aims

This chapter begins with a background to and description of the research problem. Questions addressed in the thesis are specified and the approach taken to answer them is detailed. The final sections concern the scope of the research and the methodological and substantive contributions it makes. The chapter ends with an overview and description of each chapter in the thesis.

## 1.2 Multimedia conferencing over the Internet

The Internet has revolutionised communication. A prime example of this is desktop multimedia conferencing (MMC), which facilitates real-time communication between two or more users through the tools of audio, video and a shared workspace. Until early in the millennium, this application resided mainly in the higher education and research communities, to support tasks such as remote meetings and co-located tutorials. However, improvements in network infrastructures, the increasing availability of affordable equipment and the economic and political climate have encouraged the uptake of the application by many business and personal users, who enjoy the savings of time and travel expense that it allows.

However, with this increase in adoption comes a problem that could potentially be the downfall of such real-time applications. The Internet is a 'best-effort' service, which means that it does not use direct connections between computers and cannot guarantee any level of performance. The increase in real-time traffic for inelastic (requiring a rich set of performance guarantees) applications such as MMC can result in unacceptable audio and video quality being delivered to the end-user.

Subsequently, there have been many calls for Quality of Service (QoS) guarantees to be developed, such as bandwidth reservation. This would involve service resources being allocated according to the assumed objective QoS requirements of the application. In order to do this, the quality requirements of the application must firstly be determined. Central to this are the quality thresholds for ensuring that the needs of the end-user are met. However, broad-based assumptions are made about the adequacy of quality, when research (e.g. Anderson et al, 2000) shows that quality requirements vary greatly depending on the task being performed and the frequency of use of the application. Such broad assumptions can over-estimate the quality requirements needed, which in turn pushes up the financial cost unnecessarily and as a result may put the technology out of reach from many users who would benefit.

## 1.3 Description of the research problem

The research reported in this thesis focuses on determining the audio and video quality requirements of users in desktop MMC performing passive and interactive tasks, and the most appropriate methods of doing so.

Computer workstations and high bandwidth networks can deliver high quality audio and video, however this is expensive. Most users do not want to pay more than necessary for their communication technology. In addition, there will always be a market for lower quality at a lower financial cost (Podolsky 1998). Thus, the minimum levels of quality that support users undertaking specific tasks need to be determined. It would also be useful to discover the levels at which increasing the quality further is of no potential benefit to the user, as this would allow valuable bandwidth to be conserved, which could reduce the financial cost to the user. It is always important to take the task being performed into consideration, as different tasks will require different levels of quality. For example, an important consultation between two medical experts over a multimedia

conference link will require higher levels of audio and video quality than two friends using the application for an informal chat.

The establishment of such quality thresholds is essential for network providers and application designers. To date, the networking community has largely adopted the International Telecommunications Union (ITU) recommended rating scales, which have many shortcomings (see section 2.4.1.2), such as not taking account of the user's task. Moreover, there are drawbacks of utilising subjective assessment in isolation as it is cognitively mediated. This means that external variables can influence the ratings given. Thus, results obtained may give a misleading impression about the impact of the quality on the user.

Previous research conducted at University College London (UCL) (Watson 2001) has developed a rating scale, which is an improvement on the ITU scales. The research reported in this thesis adopted this scale and used it in the context of a 3-factor HCI (Human Computer Interaction) evaluation framework of task performance, user satisfaction and a focus on the previously neglected element of user cost, through physiological indicators of perceptual strain. The signals measured were Skin Conductance (SC), Heart Rate (HR) and Blood Volume Pulse (BVP). This technique was chosen because it is an unobtrusive, continuous and more importantly, objective method.

## 1.4 Research questions

The three main research questions addressed in this thesis are:

1. Can physiological responses to media quality degradations be detected?

2. Which media quality degradations have a negative impact on users, physiologically and subjectively?

3. How do physiological responses relate to subjective data?

## 1.5 The research approach

As a starting point, questions that exist in the networking field regarding media quality requirements were identified. This assisted in the formulation of experiments and shaped the research. It was also appropriate to look at previous HCI research investigating MMC.

A critical review of the ways quality has traditionally been assessed and the status quo was necessary to highlight the problems that exist and to give support to the central argument of this thesis, which is that an additional method was needed. To introduce the rationale behind the use of physiological measures a background was given of how, why and when such responses will occur. A review of work being carried out in similar fields using such measures is of prime importance to add support to the fundamental drive of this research.

Five empirical investigations were performed to determine which media quality degradations negatively impact upon the user, and which assessment methods are most suitable in different contexts. The use of the method was also extended through two HCI experiments investigating different applications and one experiment in the area of Virtual Reality (VR). From the results of this research, both review-based and empirical, conclusions and recommendations were formulated on the quality thresholds required for specific tasks and the use of physiological measures in media quality assessment.

## 1.6 Scope of the thesis

The focus of the research presented in this thesis is methodological: the technique of utilising physiological signals is explored and developed as part of the 3-factor framework with the aim that it will become a standard evaluation approach in HCI research.

Due to the large number of parameters that contribute to the perception of quality in multimedia conferences, it was impossible to investigate them all. Therefore, some of the most important were selected from past research and from building on the results from experiments in this thesis. Both the audio and video channels were investigated. Audio degradations caused by the network, hardware set-up and end-user behaviour were investigated, as were high and low levels of video frame rate.

Solely laboratory-based experiments have been carried out in the research reported in this thesis. Measuring physiological signals in the context of MMC is novel, thus the method had to be rigorously tested in a controlled setting before it could be used in a real-world scenario in the field due to the inertia of variables that exist in such settings and with such tasks. However, the recommendations that will be made from this research will allow the method to be used in a wider variety of contexts.

## 1.7 Contributions of the thesis

The investigation of the impact of MMC quality on the end-user is of the highest importance as he/she ultimately determines the success, or otherwise, of an application. In the area of HCI, there are no specified criteria or established methods to assist in determining quality thresholds, and the methods that are traditionally used have drawbacks. In tackling this issue, the research reported in this thesis makes a number of methodological and substantive contributions.

### *1.7.1 Methodological Contributions*

1. This thesis has tailored the 3-factor evaluation framework using physiological measures of user cost for to evaluate media quality. It has shown that physiological measures can be effectively used to measure the impact of media quality degradations as part of the 3-factor framework in passive perceptual tasks. Physiological data can also offer lower level

information about the nature of the task (whether it is perceptual or cognitive) in all tasks. In interactive tasks, responses to the task can drown out responses to the quality, due to the large number of variables in operation. This thesis also shows the potential of the methodology in a different application (the web) and in a different area (VR).

2. The use of physiological signals in the experiments reported in this thesis has led to the development of guidelines for their use in HCI. Physiological signals should not be used in isolation and more than one signal should be measured, as fractionation (e.g. where SC increases and HR decreases) between the signals gives information about whether the task is primarily perceptual or cognitive. Experiments should be carefully designed with a gender balance, the order of presentation of conditions randomised and consideration given to the repetition of conditions, as all can impact upon results. Experiments with a within-subjects design are best suited to physiological measurements, however this and the baseline session take additional time, which must be accounted for in the design of the experiment. The task that participants perform and the environment in which measures are taken must be controlled or variables other than those being investigated should be accounted for to attribute responses. Limitations of measuring physiological responses are also given, such as the length of time necessary to analyse the data, the difficulties of interpreting the results, and the fact that some participants are not comfortable wearing the sensors.

### *1.7.2 Substantive Contributions*

1. The research reported in this thesis is the first investigation of physiological responses to media quality and shows that, under certain conditions, responses to media quality can be

detected. When watching a recorded interview for the first time in which the frame rate changed from 5-25-5 frames per second (fps), SC increased significantly from the first presentation of 5fps to 25fps and to the second presentation of 5fps. SC responses to the second interview (with the same frame rate sequence) were slightly different: it increased significantly between the two presentations of 5fps, however this time SC increased significantly between 25fps and the second presentation of 5fps. Most previous research into audio degradations has focussed on audio packet loss, however an experiment conducted as part of the research reported in this thesis examined a variety of degradations caused by the network, hardware and end-user behaviour in a passive listening task. Results showed that echo and loud volume differences between speakers caused significant increases in SC from a condition with normal quality and 5% packet loss caused a significant decrease in HR from a condition with normal quality. Significant differences were also found in the physiological signals due to variables such as the repetition of conditions, yet this did not impact upon subjective responses. Therefore, physiological signals can offer additional data in an experiment that subjective data may not pick up on.

2. The relationship between subjective assessment and physiological signals has been investigated in the experiments reported in this thesis. Results showed that only 12% (in the 5-25-5fps group) and 17% (in the 25-5-25fps group) of participants in a passive, engaging task (experiment 1) noticed that the frame rate had changed, whereas significant differences to the frame rates in the physiological responses of participants in the 5-25-5fps group were found. SC responses in the passive listening task were similar to subjective responses as echo and loud caused significant increases in SC and were among the three worst rated

conditions. In HR, 5% packet loss caused a significant decrease compared to a normal condition and this was rated as being of the second best quality. In the subsequent three experiments, there were no significant main effects of media quality degradations in the physiological signals (with the exception of BVP in experiment 5 where post-hoc tests showed no significant differences between conditions), therefore in such experiments subjective responses may better suited.

## 1.8 Publications

A short background to the research reported in this thesis was published in 2 papers: (Wilson 1999b) and (Wilson 1999a). The results of experiment 1 were published in 3 papers and in a book chapter: (Wilson & Sasse 2000d), (Wilson & Sasse 1999), (Wilson & Sasse 2000b) and (Wilson & Sasse 2000a). The results of experiments 1 and 2 have been reported in 3 papers: (Wilson 2000), (Wilson & Sasse 2000e) and (Wilson 2001). The full results of experiment 2 were reported in 1 paper (Wilson & Sasse 2000c). The results from experiments 3 and 5 were reported in 1 paper (Wilson 2001). An outline to the research reported in this thesis as an introduction to a panel session on the role of electrophysiology in HCI was presented in 1 paper (Allanson & Wilson 2001), and as an introduction to a workshop on physiological computing (Allanson & Wilson 2002a). The proceedings from a workshop on Physiological Computing co-organised by the author of this thesis were published (Allanson & Wilson 2002b). The contribution of the research reported in this thesis to the ETNA (Evaluation Taxonomy for Networked Multimedia Applications) project was stated in 1 paper (Mullin et al 2001). Finally, a paper describing the research reported in this thesis and commenting on detecting emotion from physiological signals was published in a journal article (Wilson & Sasse 2004).

## 1.9 Structure of the thesis

Chapter 2 presents the background to the research reported in this thesis and is comprised of four sections. Section 1 gives a background to the area of HCI and describes traditional evaluation methods. Section 2 gives an introduction to MMC and the degradations that can occur, the findings of which informed the degradations investigated in the experimental studies that form the core of this thesis. The audio and video tools used in the experiments carried out in the research reported in this thesis are described. The third section investigates how communication over a MMC link is affected by degradations that commonly occur in MMC. The final section gives a background to the area of media quality assessment and a critical review of the existing subjective methods widely used to assess the impact of quality on the user.

Chapter 3 begins with an introduction to psychophysiology, where the signals used in this thesis are introduced. A critical literature review of the area of Physiological Computing is then presented and details the state of the art in this area. It describes how physiological measures are used in a number of areas relevant to the research reported in this thesis, from HCI evaluation to Affective Computing and VR. It ends with a comparison between the research reported in this thesis and the studies detailed in this chapter with regard to factors such as experimental design, signals measured, features used and baselines.

Chapter 4 sets out the methodology utilised in the empirical investigations in this thesis, which was a 3-factor framework incorporating measures of task performance, user satisfaction and user cost. Measures used in each of these factors are specified and their use justified. The methodology for the measurement of the physiological signals in the experiments in this thesis is presented.

Chapters 5 and 6 describe empirical research that has been conducted as part of this thesis. Chapter 5 describes three experiments investigating the impact of video frame rate and audio degradations in passive tasks. Experiment 1 investigated the impact of high and low video frame rates in recorded interviews. The variable of interest in experiment 2 was audio degradations (without the video channel), of which six plus a reference condition were investigated in a passive listening task. The results from this experiment fed into experiment 3, which included the video channel and examined four audio degradations from the previous experiment in recorded interviews.

In experiment 1 there were significant differences in SC of the group who saw 5-25-5fps (in the first interview, 25fps and the last presentation of 5fps were significantly higher than the first presentation of 5fps and in the second interview, the second presentation of 5fps was significantly higher than the first and 25fps). There was also a significant interaction between group and interview in SC and significant differences between the interviews in HR. Experiment 2 showed a significant main effect of degradation in all signals (BVP showed no significant differences in post-hoc tests): echo and loud caused significant increases in SC from a normal quality condition and 5% packet loss caused a significant decrease in HR from a normal quality condition. There were also significant main effects of presentation in both SC and BVP and gender in BVP. Finally, there were significant interactions between presentation and gender (SC) and degradation and presentation (BVP). Results from these experiments show that physiological responses to media quality degradations can be detected. The physiological results from experiment 3 were not significant. The direction of the means showed that the tasks in experiments 1 and 2 were perceptual and the task in experiment 3 was cognitive.

Chapter 6 describes two experiments that investigated the impact of audio and video degradations in interactive tasks. Experiment 4

investigated the impact of high and low video frame rates in an interactive interviewing task with inexperienced interviewers. Experiment 5 investigated high and low audio packet loss combined with high and low video frame rates in the same interviews in an interactive interviewing task with experienced interviewers. Significant main effects of media quality degradations in these active tasks were not obtained (with the exception of BVP in experiment 5, however post-hoc tests showed no significant differences). This is most likely due to the large number of variables that were operating, which made it difficult to tease apart effects of the quality from the other variables, such as task stress. In experiment 4 there was a main effect of gender, a significant interaction between the frame rate and the order in which the frame rates were seen and another interaction between the frame rate, the order in which the frame rates were seen and the order of the interviewees seen. The direction of the means showed that the task in experiment 4 was cognitive and the task in experiment 5 was perceptual. This may be because the participants in experiment 5 were experienced interviewers whereas those in experiment 4 were not.

In chapter 7, the methodological and substantive contributions of the thesis are presented, followed by the limitations of the research reported in this thesis. Recommendations for HCI researchers on the use of the 3-factor framework and for network providers on the minimum levels of audio and video quality required for various tasks are given. Finally, a research agenda for this area is presented.

## Chapter Summary

This chapter has presented a background to and description of the research problem and the three main research questions being tackled in the thesis. The scope and contributions of the research was specified, and an outline to the thesis has been given.

# Chapter 2 Background

## Chapter Aims

This chapter is divided into four sections. The first section introduces the area of HCI. It begins with a description of its history, background and current status. The role of evaluation and the evaluation methods that exist are outlined and the evaluation approach used in this thesis is described. The second section describes MMC over the Internet. It explains the tools used and factors that can affect its perceived quality. The third section of this chapter investigates how communication in MMC is affected by degradations that commonly occur in MMC. The chapter finishes with a critical review of the way in which MMC quality is most commonly assessed.

## 2.1 Human Computer Interaction

At the beginning of the twentieth century, studies into the performance of people working manually with machines in factories began. This area of research became more important during the Second World War, driven by pressure to produce more effective weapons systems. In 1949, the Ergonomics Research Society was formed and was mainly concerned with how the physical characteristics of machines and systems impacted upon the performance of the people operating them. As the use of computers became more widespread, the field of man-machine interaction emerged. This centred on the interaction between people and computers with an emphasis on the physical and psychological aspects of the interaction. In the 1980s, with the emergence of the personal computer (PC), this area became more popular, and the term HCI was created.

The main aim of HCI is to make systems more usable. Usability is comprised of three elements: effectiveness, efficiency and satisfaction (e.g. Frokjaer et al, 2000). Effectiveness is the accuracy

with which users can complete their tasks. It can be measured by, for example error rates. Efficiency is the relationship between effectiveness and the resources used to achieve completion of the task. This can be measured by, for example the time taken to complete the task. Satisfaction is the user's comfort in using the system.

In order to make systems usable an understanding of five areas is required.

1. The users, including their psychological and physiological abilities and limitations. This encompasses areas such as communication, information processing, language, interaction and ergonomics.
2. The computing technology, for example input and output techniques and computer graphics.
3. Tasks the user performs.
4. The usability of the system (Dix et al 2004), which is *"...a measure of the ease with which a system can be learned or used, its safety, effectiveness and efficiency, and the attitude of its users towards it"* (Preece et al 1994).
5. The environment in which the user is operating. This encompasses physical (e.g. lighting) and psychological and social aspects (e.g. job structure).

Due to the diversity inherent in these five areas, HCI is an interdisciplinary field. At its centre are computer scientists (who create the technology), psychologists (to contribute knowledge of the perceptual, cognitive and problem solving skills of humans) and experts in ergonomics (who design tools for different environments with the capabilities and capacities of users at the centre). HCI also attracts researchers from areas as diverse as art and design, sociology and linguistics. The multidisciplinary makeup of this area contributes to the fact that there is no *"...general and unified theory of HCI"* (Dix et al., 2004).

The first stage in conducting research in HCI is to identify knowledge from related areas that can be adopted. The second stage regards the type of knowledge to be furthered. Both substantive (research findings) and methodological knowledge (system development methods and evaluation techniques) require contributions. A combination of quantitative (which involve measurements, such as rating scales) and qualitative (which involve descriptions and anecdotes, for example from interviews) research methods can be used to ensure that the overall findings are comprehensive. The final stage regards the application of the knowledge gained. If it is not relevant or comprehensible to system designers, then it will not be adopted and users will be disadvantaged by having to interact with a poor system. Thus, to ensure its applicability, the information should be in the form of guidelines, methods, or incorporated into tools.

The study of HCI has tended to come late in a computer scientist's training, if at all. In addition, with increasing concerns about the health and safety of employees, there is pressure on employers to provide systems that are not just safe, but also usable. At a fundamental level, if a system is not usable and there are a variety of systems available that support the performance of the same task in a more effective and efficient way, then the system will not be adopted and utilised by users.

### *2.1.1 HCI Evaluation Methods*

There are methodologies and models that assist a designer in creating a usable interactive system. Yet, even where these are employed, it remains important to test the system to show that it meets the requirements of the user. This is the role of evaluation, which is defined as *"...an assessment of the conformity between a system's actual performance and its desired performance"* (Whitefield 1991).

When planning an evaluation, there are many factors that need to be taken into account with regard to the most appropriate assessment methods to use (Dix et al., 2004).

1. The information elicited.

Evaluation can be either summative or formative and the distinction depends on the information elicited. Summative evaluation is carried out on the final product and aims to determine if the product is usable and if the goals of the design have been achieved. On the other hand, formative evaluation intends to improve the usability of a product by detecting usability problems and understanding why they occur so that the problems can be fixed in the next version of the product.

2. The style of evaluation.

Evaluation can take place in a laboratory or in the field and can be either experimental or non-experimental (referring to the control and formality of the investigation). Most commonly, experiments are performed in the lab and non-experimental studies are undertaken in the field. Experimental studies have a great degree of control but lack context, whereas non-experimental studies are subject to unrelated events, yet are more ecologically valid. Ideally, both styles should be performed in both settings to ensure that problems with the system are identified and dealt with before it is implemented.

3. The type of measures provided.

Quantitative measures, for example ratings scales, are usually numerical and can be analysed using statistics. On the other hand, qualitative measures are usually non-numeric, for example verbal statements from interviews. The researcher has to consider the type of data that will be most useful in answering his/her questions and the time he/she has to analyse responses. For example, interviews collecting qualitative data can take large amounts of time to perform and analyse, yet offer in-depth data and the opportunity to tailor the

interview to the participant by, for example asking him/her to expand on certain points.

The method used depends largely on the experimental questions. If the experimenter is unsure of the problems with their application, then qualitative data can help to uncover them. However, if the issues are known, then a study using quantitative data (such as closed questions) can determine the relative frequency in a larger sample. Thus, interviews using qualitative data can be suited to smaller groups of participants, whereas analysing quantitative data from rating scales or questionnaires can be quicker and, therefore better suited to larger groups of participants. In addition the responses can be analysed more rigorously than, for example qualitative interview responses. However, if quantitative questionnaires are being used the experimenter has to be certain of the questions he/she wants to ask at the outset and accept that the questions will be less probing than he/she would ask in an interview.

4. The level of subjectivity or objectivity of the evaluation technique.
Evaluation methods differ in their degree of subjectivity or objectivity. For example, some techniques are reliant on the skills of the evaluator, such as the cognitive walkthrough and think aloud techniques (Lewis 1982). These can provide information that may not be available from objective methods, however evaluator bias is an inherent factor. Objective evaluation methods, such as controlled experiments, should allow many evaluators to achieve the same results. The disadvantage is that they may not *"...reveal the unexpected problem or give detailed feedback on user experience"* (Dix et al., 2004).

Another level to be considered is whether the technique used depends on the subjective response of the participant. For example, interviews, questionnaires and methods like think aloud (Lewis 1982) are subjective, in that they are based on users' opinions, therefore

they may *be "...a 'rationalised' account of events, rather than a wholly accurate one"* (Dix et al., 2004). Objective methods, such as physiological measures, are not subject to biases because they are outside the control of the participant, thus *"...ideally, both objective and subjective approaches should be used"* (Dix et al., 2004).

6. Information provided

The information required from an evaluation can be different at various stages of the design process. Generally, experiments are better suited to providing low-level information (for example, is a specific size of video window found to be more satisfying?), which can allow a particular decision about the design to be made, whereas methods such as interviews elicit more complex, high-level responses about the overall system, for example *"...is the system usable?"* (Dix et al., 2004).

7. Immediacy of response

Techniques such as think aloud (Lewis 1982) give participants' responses at the time of the interaction, yet they can be intrusive and influence the way the participant performs the task. They may also force the participant to focus on the evaluation as opposed to the application they are interacting with. On the other hand, questionnaires or post-task walkthroughs elicit information after the event. This means that they can be subject to inaccuracies in the recall of information and could be affected by primacy and recency effects. These are characterised by participants remembering more about what they experienced at the beginning or end of a time period, as opposed to what occurred during the time period. However, they do allow the participant to concentrate more on their task.

8. Resources

Elements such as time, money, equipment, participants and the expertise of the evaluator are all considerations when planning an

evaluation and will influence the methods adopted. A comprehensive classification of evaluation techniques can be found in Dix et al. (2004).

### *2.1.2 The 3-factor approach*

The framework utilised in the research report in this thesis is based on that developed by Shackel (Shackel 1981). He stated that in the measurement and evaluation of systems, there are three criteria that must be measured.

1. Dimensional criteria, which rely on physical measurement, mainly the *"...size, shape and other characteristics of the tool in relation to human size".* They allow only a pass or fail judgement.
2. Performance, which is often measured in HCI as time or the number of errors.
3. Attitude, which assesses *"...the user's view of the cost and relative difficulty in achieving the performance"* through subjective measures such as scaling techniques. It is, in effect, a subjective judgement of the ease of use of an application. However, when this paper was published in 1981 there were very few studies that were based on the attitude criterion.

Shackel argued that the three factors should be viewed as being complementary to each other in usability evaluation, as opposed to focussing on only one. For example, the performance criteria cannot be the only criterion because a person *"...may readily achieve a given performance, but still not prefer to do the task or use the tool because it is very inconvenient and awkward, so that he may well prefer (i.e. find more useable) another similar tool which gives less speed or more errors but is easier or more convenient".* Despite Shackel putting forward the importance of measuring user cost it has not been incorporated into evaluation practice, especially not objectively. Therefore, the evaluation framework used in the research

reported in this thesis is comprised of three factors: task performance, user satisfaction and user cost. With reference to Shackel's criteria, performance and attitude (user satisfaction and user cost) are being measured. Dimensional criteria are not relevant to the research reported in this thesis because it investigates quality levels in an application, and the physical set-up across all experiments was similar.

This framework is similar to the effectiveness, efficiency and satisfaction factors that comprise usability in that it measures task performance (which is comparable to the effectiveness and efficiency factors) and also measures user satisfaction. However, it extends the efficiency criterion from being the relationship between effectiveness and the resources used to achieve the goal, to being a factor on its own with an equal status to task performance and user satisfaction. In an application like MMC where quality can be poor, examining the cost to the user of interacting with the application is increasingly important.

The relative weighting of importance of each of the three factors is largely dependent on the task being performed. For example, in a safety critical application, task performance will be the most important dimension. It is also based on the goals of the evaluation, which are in turn related to the time available, the skills of the researchers and resources available.

The methods used in the evaluation framework are questionnaires (subjective and quantitative), rating scales (subjective and quantitative), and physiological measures (objective and quantitative). Outcomes from the tasks used in the experiments reported in this thesis could not be directly measured because the process of communication was more important than the end product. However, measures of time were taken in the interactive interviews

and participants were given a meaningful task in four of the five experiments to provide a focus.

## 2.2 Multimedia Conferencing

The preceding section introduced the area of HCI and evaluation. This section describes the specific application, MMC over the Internet, on which the research reported in this thesis was focussed in an attempt to determine quality requirements and improve evaluation methods.

The use of MMC is increasing in areas such as distance education, remote health care, personal communication and business meetings with participants in different locations. It offers its users a large number of benefits, such as savings in time and money by reducing the need for travel and also offers the potential to communicate with large numbers of people at the same time. However, the quality in multimedia conferences over the Internet is subject to unique degradations, which differentiate it from conferences over dedicated links. For example, some packet loss is possible when involved in MMC. Generally it is in the region of 2-5% (Handley 1997), yet this figure can be significantly higher, depending where on the network the computer is. Packet loss is complex and fluctuates according to network conditions. Even with future improvements in the network and QoS guarantees, there will be users who will accept lower quality levels as a trade-off for lower financial cost.

### *2.2.1 Multicast Multimedia Conferencing*

MMC involves real-time communication between two or more users through the tools of audio, video and a shared workspace, where necessary. Figure 1 shows the typical set-up for this application. The user sits at a desktop machine and a camera placed on top of the screen captures their image. He/she wears a headset with a microphone attached in order to hear and speak. Three types of digital content are sent and received: audio, video and a shared workspace. The latter tool was not investigated by the research

reported in this thesis, as audio and video are the more important channels in the tasks used in the experiments.

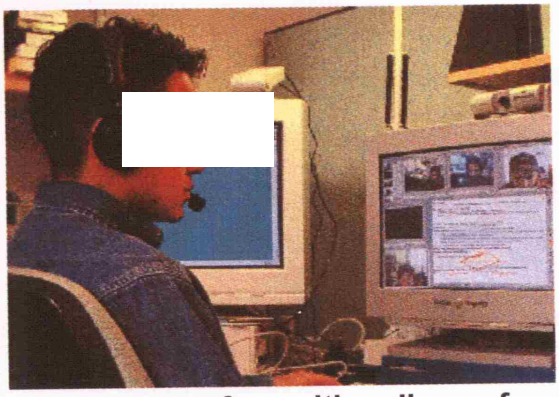


**Figure 1: Typical set-up of a multimedia conference**

#### 2.2.1.1 Audio

In order to send audio over the Internet, firstly it needs to be digitised using a speech-coding algorithm. Such algorithms can be waveform based, which attempt to preserve the speech signal, or knowledge based, which model the speech production process. After digitisation, compression takes place in order to minimise the bit rate required. This is made possible by the fact that a large amount of the speech stream is redundant (Fluckiger 1995).

The tool utilised throughout this thesis to send and receive audio is called the Robust Audio Tool (RAT[1]), and was devised at UCL for use in unicast and multicast conferences. It is a previous generation of tools to those such as Skype[2] that exist today. However, it was used because technical support was available for any problems that arose and it allows the user to control a large number of parameters that contribute to quality, thus making it well suited for the experiments carried out as part of the research reported in this thesis. It was designed to cope with a variety of different Internet conditions and was developed for use with other audio tools and

---

[1] http://www-mice.cs.ucl.ac.uk/multimedia/software/rat/
[2] http://www.skype.com/

across a variety of platforms, thus there are a number of options given to the user. The default settings are shown in Table 1.

| Setting | Definition | Options | Default |
|---|---|---|---|
| Encoding scheme | Speech-coding algorithm | - 16-bit linear<br>- Pulse Code Modulation (PCM)<br>- Digital Video Interactive (DVI)<br>- Global System for Mobile Communications (GSM)<br>- Linear Predictive Coding (LPC) | DVI |
| Packet size | Long, medium or short in duration. All have benefits and drawbacks | 160ms, 80ms, 40ms, 20ms | 40ms |
| Silence suppression | Only audio above a certain level is transmitted (conserves bandwidth) | On/off | On |
| Full/half duplex | Full is where users can speak and listen at same time. Half is where the speaker's microphone is muted as soon as someone else speaks | Full/half | Full |
| Repair methods | Methods employed to repair audio affected by packet loss | Receiver based (silence substitution, packet repetition) or sender-based (interleaving, redundant transmission) | Packet repetition |
| Reception statistics | Statistics box giving information on audio being received | | Click on participant's name to view |

**Table 1: Default settings of RAT**

#### 2.2.1.2 Video

Compression is vital with the video stream. There is a trade-off between frame rate and resolution, as increasing one implies decreasing the other with a limited bit rate. A widely used standard

for video compression in videoconferencing is H.261. It supports three image sizes.

1. QCIF (Quarter Common Image Format), which is 176 x 144 pixels.
2. CIF (Common Image Format), which is 352 x 288 pixels.
3. SCIF (Super Common Image Format), which is 704 x 576 pixels.

With this standard, the whole frame is updated on a block-by-block basis, depending on whether there is new information in that block. Thus, at low bit rates there can be a partial update of the faces of participants in the conference. The video tool used in the experiments conducted as part of the research reported in this thesis is vic[3] (VideoConferencing tool) and was devised at Lawrence Berkeley Labs (McCanne & Jacobson 1995) (Figure 2). It was adopted because it was being developed by researchers at UCL, thus technical support was available.

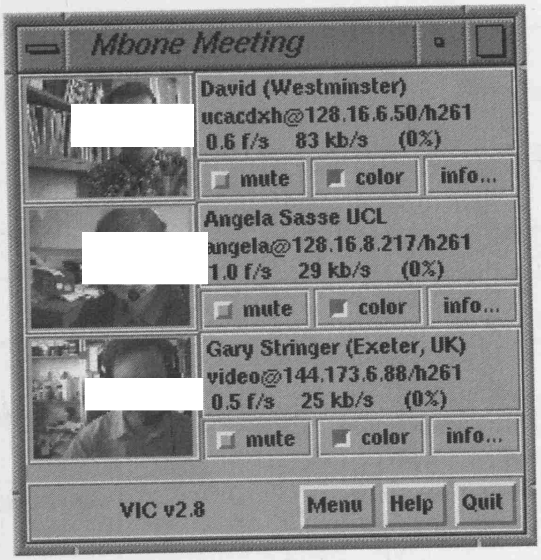



**Figure 2: Interface of vic with QCIF images**

Like RAT, vic was also designed to operate in a number of different conferencing environments and platforms, thus it has a number of

---

[3] http://www-mice.cs.ucl.ac.uk/multimedia/software/vic/

settings that can be selected by the user, which can be seen in Table 2.

| Option | Default |
|---|---|
| Colour/greyscale | Colour |
| Bit rate | 126kbit/s (ranges from 10kbit/s to maximum for that session, determined by distance packets have to travel) |
| Frame rate | 8 fps (maximum is 30) |
| Reception statistics | Shown next to QCIF image, specifying the frames, kbit/s and packet loss being received from participants |

**Table 2: Default settings of vic**

### *2.2.2 Factors affecting perceived quality in MMC*

There are many research questions in this area, which need to be addressed in order to move the technology forward. Many of these can only be answered by determining the impact of the many factors that can affect the perceived quality of a multimedia conference (see figure 3). The network, end-user behaviour or the hardware set-up can cause these. Clearly there are many and, due to time constraints, a selection of the most important were investigated in the research reported in this thesis.

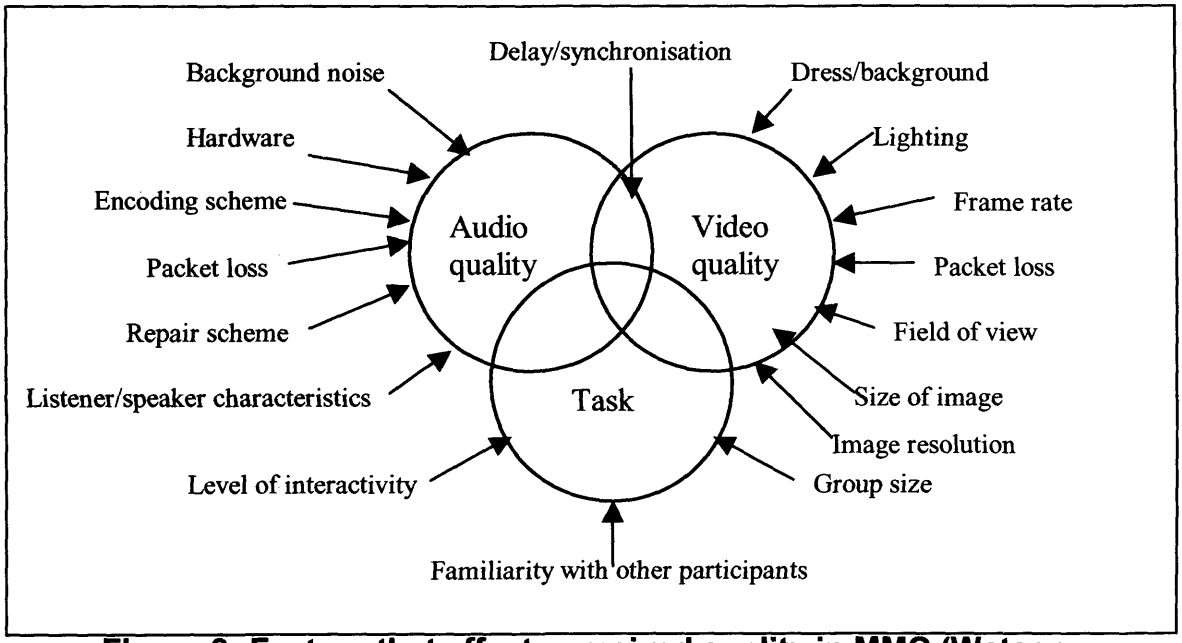



**Figure 3: Factors that affect perceived quality in MMC (Watson 2001)**

## 2.3 Communication in multimedia conferencing

This section summarises communication in MMC, specifically the impact that the commonly experienced degradations have upon it. Audio in isolation is considered, then the impact of video and audio is described. An extensive review of this area can be found in Watson (Watson 2001).

### *2.3.1 Audio degradations*

Communication over MMC can suffer as a result of the degradations it is subject to. Investigating this area is of the utmost importance because if users cannot effectively communicate using this technology, then they will not utilise it. Audio is the most important channel in MMC (e.g. Sasse et al., 1994). As detailed in figure 3, there are a large number of degradations that can affect it, however the most important is packet loss. Packet loss can result in a loss of phonemes or syllables. Packets typically contain 40 or 80 milliseconds of information, which is roughly the size of a phoneme. A loss of a phoneme or a syllable can lead to impaired speech intelligibility, however the human brain can be extremely resilient to such disruption. Thus, there is no direct relationship between intelligibility and degradation and this is largely due to the effects of the surrounding speech content (e.g. Warren, 1970).

Another degradation that can occur in MMC is delay. This can disrupt the flow of a conversation because, for example it can be unclear whose turn it is to speak, and this can lead to interruptions and ill-timed backchannels, which are utterances like 'uhuh' that serve as positive evidence that the listener has understood the speaker.

There are also factors due to the hardware set-up of MMC that can degrade the audio stream, which are largely due to having to wear a headset and speak into a microphone. As a result of no sound localisation it can be difficult to determine who is speaking, especially in a conference with many participants. In addition, speaking into a

microphone can lead to people talking loudly or quietly, as they cannot hear their own voices. Echo and feedback can occur as a result of having a 'leaky' headset and these can be distracting to both the speaker and listener (Watson & Sasse 2000). Finally, problems due to end-user behaviour, such as volume differences between speakers, can affect the audio quality of a multimedia conference. These can be distracting and uncomfortable to listen to (Watson & Sasse 2000).

The degradations described above impact upon the audio channel when considered in isolation. However, when the video channel is added, determining the impact upon communication becomes more complicated.

### *2.3.2 Adding video to audio*

With the addition of video, more information becomes available to the viewer, such as lip movements, facial expressions, gaze, gestures and body language. This section examines the additional data that becomes available with video and ways in which communication can be affected by degradations in it. It must be noted that there has not been a lot of research investigating how communication is affected over MMC. The majority of the research that will be presented here investigates what happens to communication when it is mediated by video (Video Mediated Communication, VMC) compared to face-to-face communication. This research is applicable to MMC communication, however it must be remembered that: MMC quality fluctuates and it is not guaranteed; eye contact is not possible in MMC; more than two users can participate in the same conference and image sizes will be relatively small. Therefore, results of experiments performed with a different set-up must be applied to MMC with caution. There are also difficulties in comparing data from experiments in VMC. This stems from the fact that the studies have been carried out in a large number of environments with different criteria for measuring the effects. In addition, a vast number of tasks

have been used and different methods of analysis employed. Despite this, valuable information can be obtained from such experiments and applied to MMC.

#### 2.3.2.1 Lip-reading

A major benefit of the presence of the video channel is that it allows participants to lip-read, which includes the tongue, teeth and movements of other parts of the face. This is particularly valuable in applications like MMC, where it is likely that there will be a degree of background noise (Summerfield 1992) or packet loss on the audio stream. In addition, it is helpful for people who have hearing impairments and situations where the language spoken in the multimedia conference is not the person's first language (Reisberg et al., 1987).

The frequent mismatch between the audio and video streams in MMC could lead to difficulties in perception arising, such as the McGurk effect (McGurk & MacDonald 1976). This is where the sound for one syllable is dubbed onto the lip movements of another sound and participants generally report hearing an unrelated syllable or the visual 'sound'. This effect is taken as evidence that speech perception is multimodal and that the visual channel dominates perception. Thus, an asynchrony between audio and video could result in the McGurk effect occurring frequently. The literature suggests this is unlikely due to the small size of the video, its poor quality, low frame rate, and the fact that it tends to lag behind audio. There has not been a vast amount of research in the networking community regarding the audio-visual delays tolerable for perception, yet it has been suggested that the end-to-end delay should not be more than 300ms (Roy 1994).

#### 2.3.2.2 Frame rate

In the United Kingdom, full motion video is delivered at 25fps, whereas in the United States it is 30fps. The influence of vision on

speech is maximised above 15fps (e.g. Barber & Laws, 1994). Frame rate does seem to be the most important determinant of video quality in verbal tasks, with users willing to sacrifice other aspects of image quality to keep the frame rate above 5fps (Pappas and Hinds, cited by Kies et al., 1997).

Due to bandwidth restrictions and the processing power of machines, full motion video is not always achievable and frame rates can drop to as low as 5fps. At the time this thesis research was conducted, rates as low as 5fps were common. Nakazono (1998) found that frame rates below 5fps can impair speech perception, yet it has been found that any increase in the visual representation of a speaker increases the listener's tolerance for background noise (Summerfield 1992). Additionally, there is a difference between interacting in real-time, such as in experiments 4 and 5 in this thesis (see chapter 6) and streaming, which is widely used and can improve quality by having a start-up delay and large buffer.

A study by Anderson et al. (2000) found that participants did not subjectively notice the difference between 12 and 25fps when they were involved in an engaging task and there were no significant differences in task performance. However, when the data were short video clips in isolation the difference was noticed. This finding lends support to the argument that different tasks have different quality requirements.

#### 2.3.2.3 Gaze

Gaze in MMC is a complex issue. In face-to-face communication, it occurs around 50% of the time (Argyle 1990). There are two types of gaze: regular gaze, where one person looks at the other (usually at the mouth) and mutual gaze, which is where eye contact is made. Mutual gaze occurs much less often than regular gaze does. Gaze serves three main purposes: to indicate that the person being

communicated with is paying attention; to emphasise a particular word through brief eye contact; and to facilitate turn-taking.

Gaze in MMC, especially mutual, is extremely difficult to achieve. This is because the user will typically be watching the other person's image on the screen as opposed to looking into the camera, which is usually placed on top of the screen. In communication, people usually end their utterance by a sustained gaze, thus when considering the difficulties to obtain gaze in MMC, it would be predicted that turn-taking would become more difficult. However, Anderson et al. (1997) found that turn-taking becomes more structured in VMC as more turns and words are used but interruptions are fewer. It is possible that this increase in formality would reduce with experience.

Monk & Watts (1995) performed a study investigating the impact of the size of a video image (small or large) on the focus of gaze of participants (whether they looked at the video image or somewhere else) and their speech (whether they were speaking or silent). Each session involved two participants. The tasks were a screen-based questionnaire on their joint interests and a card game where one participant tried to deceive the other. Results showed that the small video resulted in less fluent verbal interaction, however there were no significant differences in gaze between the large and small video windows.

#### 2.3.2.4 Image size

The image size utilised is generally a determinant of the bandwidth available, the screen space available and the processing power of the specific machine being used. Most of the research into the impact of image size has occurred in the remote education field and the general consensus has been that despite people subjectively preferring larger images, it does not have an impact upon their learning (e.g. Hearnshaw, 1999).

With the addition of the video channel comes the benefit of being able to observe the gesture and posture of the person being communicated with, both of which are valuable sources of communication. The most common view in desktop MMC is of the head and shoulders, thus it is not possible to observe gestures, however this allows the viewer to take in more information from the face.

### *2.3.3 The benefits of video*

VMC seems, as yet, unable to emulate face-to-face communication, adding video does not add much to task performance, it can make the process of the interaction lengthier and it is hypothesised to add a cognitive load that can impair conversational behaviour (e.g. O'Malley et al., 1996). However, the important point to make is that subjectively, participants prefer the video channel to be there (e.g. Tang & Issacs, 1993), and the opinions and preferences of the user are of the utmost importance.

Generally, the research cited throughout this section has looked at the way that communication is affected when mediated by video. This type of research is extremely useful, yet has minimal input into the development of the technology. Thus, the research reported in this thesis takes a traditional usability approach by investigating how media quality degradations impact upon the user.

### *2.3.4 Quality requirements for different tasks*

The ETNA project[4] (Mullin et al., 2002) attempted to classify multimedia applications in terms of the task or activity, user and other aspects of the situation in which they are used, in a way that is useful for determining what levels of audio and video quality they require. An additional aim of the project was *"...to introduce new methods for establishing audio-visual quality requirements for a range of real-time*

---

[4] See http://www-mice.cs.ucl.ac.uk/multimedia/projects/etna/

*multimedia applications"*, because *"...until recently, there were few HCI-specific methods for assessing audio and video quality and the usability of videoconferencing systems"*.

The types of MMC tasks were divided into telepresence and teledata. Teledata, which is applications where the audio and video channels are used to carry other sorts of information such as images of shared work objects (e.g. Nardi et al., 1993), was further divided into foreground and background tasks. Telepresence, which is applications that are used to support communication or awareness between users, was divided into foreground and background tasks, the former of which was divided into interactive, non-interactive, social and cognitive tasks. Cognitive tasks were further subdivided into difficult, urgent, emotive and standard tasks. Knowledge of the task the user is to perform along with information about the user and situation characteristics were argued as essential information to be considered when determining the quality requirements of a multimedia conference.

A brief summary of the quality requirements for different tasks is given.

1. In foreground telepresence tasks it has been found that users adjust the clarity of their speech to compensate for a lack of visual cues due to poor video (Blokland & Anderson 1998). The communication of content and process is affected by audio delay (O'Conaill et al., 1993), yet visual feedback has little effect on task performance (e.g. Anderson et al., 1996), although it may become more important if the task or communication process is more difficult (Rudman et al., 1997). Counter-intuitively, it has also been found that video at a high frame rate can actually impair conversational behaviour with regard to turn-taking and the amount of speech generated, both of which increase compared to face-to-face and audio only

(Anderson et al., 1997). This could be because the participants found the video distracting or used it more than they normally would due to its novelty. The ETNA project concluded that video quality may be relatively unimportant if lip reading or access to social cues is not needed, however there may be a minimum acceptable threshold.

2. In interactive tasks, certain aspects of video quality, such as image size, can affect the perceived or actual interactivity of a conversation (e.g. Monk & Watts, 1995). Audio quality may be more important in a non-interactive setting because, for example there is no opportunity to ask the person to repeat what they said.

3. Cognitive tasks (that involve co-operative problem solving or the transfer of information) do not seem to benefit from the addition of video (e.g. Short et al., 1976). However, audio quality would need to be at least adequate. A study by Veinott et al. (1997) found that non-native speakers who performed the Map Task[5] in English benefited from the addition of the video channel, thus it may be helpful to non-native speakers.

4. More complex tasks (e.g. Olson et al., 1994) or tasks where communication is more difficult (e.g. Veinott et al., 1997) may place a greater importance on the video channel.

5. Social tasks, such as conflict resolution rely more on the presence of visual information and non-verbal communication and as a result may be more affected by

---

[5] The Map Task involves participants working out a route from two maps that differ slightly. It was devised to compare face-to-face communication with audio only (Brown et al., 1984)

the quality of the video. Thus, frame rate may need to be higher for social tasks. Horn (2001) found that at very low frame rates (3fps) it becomes significantly more difficult to discriminate when someone is lying. Thus, in a sensitive or high-stake interaction where detecting lies or similar cues is vital, a higher frame rate will be required. In addition, as affective information is also carried by the voice it may be that high quality audio is important.

6. Rudman et al. (1997) found that participants said the video was useful in monitoring the understanding of the person they were communicating with in a complex task. In addition, Boyle et al. (1994) found that participants performing the Map Task gazed at each other more when communication was more difficult. Veinott et al. (1997) found that the video channel was advantageous when communication was harder. Olson et al. (1994) came to the same conclusion when the task was more complex, whereas many studies that found no advantage of the video channel used simple tasks (e.g. Chapanis et al., 1972). Therefore, users performing difficult tasks may benefit from higher quality video and also audio, as such tasks may benefit from an interactive conversation style.

7. When a task is urgent, it may be especially important that communication is clear through the provision of high quality audio and video.

8. Olson (1994) proposed that tasks with a strong emotional content will require a greater degree of visual information because they may require affect cues, for example from facial expression, and also higher audio quality, as emotion is also carried by the tone of voice.

The findings of the ETNA project illustrate the complexity of determining quality requirements in MMC and also the difficulty of measuring task performance in MMC in ecologically valid scenarios. However, the taxonomy goes a long way to combining the myriad of findings and strands of research that exist in this area.

## 2.4 Media Quality Assessment

This section reviews the way in which media quality is assessed. Most research concerned with investigating quality in MMC has involved participants subjectively rating the quality. Quality has traditionally been considered in this area of research as unidimensional. This approach is questionable because there are many individual variables that contribute to the overall perception of quality. Variables such as loudness, intelligibility, naturalness, listening effort and pleasantness of tone have all been identified as contributing to audio quality (Kitawaki & Nagabuchi 1998). With regard to video, variables like colour, brightness, background stability, speed in image reassembling, outline definition and the mosaic/blocking effect all contribute to its perceived quality (Manzanaro et al., 1991).

### *2.4.1 Audio and video assessment tools*

#### 2.4.1.1 The ITU Scales

In order to standardise the wealth of different testing methods and conditions employed in the investigation of speech and video, the ITU was formed. It is the recommendations of this organisation that are most widely employed in assessing MMC quality. The ITU is an international organisation within which governments and the private sector co-ordinate global telecommunication networks and services. It is the leading publisher of telecommunications regulatory and standards information. The ITU-T (standing for telecommunications) set of recommendations were originally designed to evaluate toll (i.e. telephone) quality audio and the ITU-R (standing for

radiocommunications) were originally devised for testing audio and video over entertainment and broadcasting systems.

The ITU-T scales for assessing speech transmission are gathered together in the P series. The ITU-T P.800 (ITU-T) is comprised of conversation opinion and listening opinion tests. The scale used for both types of test is the 5-point category scale, the results from which are averaged across participants to give a Mean Opinion Score (MOS). The material in the listening tests are groups of short sentences around 5-15 seconds in length. With the conversation test, the conversation has to have a natural beginning and end and has to be long enough to ensure that the participant can gain a full impression of the quality of the connection.

There are many problem when using these scales to assess the quality of speech transmitted over a MMC link. For example, the test material used is not long enough to allow the participant to experience the full range of degradations that typify MMC audio. In addition, the conversation difficulty scale with its 'yes' or 'no' choice of answer means that it is difficult to determine which part of the conversation is being rated, given that quality in MMC is variable.

Scales for assessing image quality come under the ITU-R set of recommendations. Stimuli can be rated on their own using the image quality or impairment scale or compared to a reference stimulus: DSIS (Double Stimulus Impairment Scale), DSCQS (Double Stimulus Continuous Quality Scale) or stimulus comparison scale. The test material with these methods is generally 10 seconds in length. There are also numerical and non-categorical methods that can be used here. This test can be beneficial when there is no reference condition available. The latter involves the participant assigning a rating either on a point on a line drawn between semantic labels or to attribute a number that reflects a value on a specific dimension. ITU P.910 (ITU-T) regards the assessment of non-interactive low and medium quality

digital images and finally the P.920 (ITU-T) brief covers interactive test methods, which are based on conversation opinion tests.

Similar problems to audio exist when the ITU scales are used to assess video quality in MMC. For example, the test material is frequently short, thus the full range of degradations that can affect MMC quality will not be experienced. Finally, given that MMC video quality is frequently poor, it is highly unlikely that the video channel would ever be used in isolation, therefore it is unwise to assess video without audio. This point becomes even more important when the evidence presented in section 2.3.2 is reconsidered, regarding the interactive influence of both channels upon each other.

The ITU released a recommendation for the assessment of low and medium quality digital images (ITU-T P.910) (ITU-T). The scales involved are again the 5-point quality and impairment scale and the pair comparison method. However, there are problems with the use of these scales, as has been discussed above and the pair comparison method is unsuitable for MMC video because it demands that the images are of almost equal quality. Conversation opinion tests (P.920) (ITU-T) are recommended for interactive test methods in multimedia services. These are better suited to audio-visual quality than conventional tasks because they focus attention on the video channel.

#### 2.4.1.2 Problems with the ITU scales

In addition to the problems mentioned above, there are two more fundamental issues with the ITU recommended scales, which are investigated in Watson & Sasse (1998). Firstly, the intervals represented by the category labels are claimed to be equal, however this has been called into question. For example, it was found that the terms 'bad' and 'poor' were relatively similar in meaning, whereas the distance to 'fair' was much greater (Jones & McManus 1986). Having unequal intervals means that the use of parametric statistics on the

data, as are mainly used, is invalid because this requires a normal distribution.

Secondly, there is a concern that the scale labels have not been adequately translated into different languages because the positional rankings of the scale labels in different languages are not equal (e.g. Jones & McManus, 1986). Add to this the fact that participants have different concepts for semantics and will use rating scales in different ways, it is clear that there are many problems with the ITU recommended scales.

#### 2.4.1.3 Alternatives to the ITU scales

In order to counteract the problems with the ITU scales, specifically with the vocabulary of the scale labels, a new rating scale was developed (Watson & Sasse 1997). It is a 200mm unlabelled scale with a plus and minus sign at either end to indicate polarity.

It was found that listening quality ratings gathered from 24 participants were consistent where two different packet loss repair schemes (LPC and Packet repetition) were employed. In addition, it was discovered that the unlabelled scale reduced the tendency to avoid the end points of the scale (Watson & Sasse 1997). This scale can be adapted for different purposes, for example to measure overall adequacy. The scale was later developed to be a 100-point scale with markers at every 10 points and the labels "Very poor quality" at point 1, and "Very good quality" at point 100 (Watson 2001) (see appendix C for an example of this scale). This modification occurred due to the requirements of an experiment in which ratings were gathered at the end of a number of multimedia conferences on a paper-based scale and also during multimedia conferences on a web based scale. The polar scale was impractical to implement in HTML, therefore participants had to provide a number between 1 and 100 relating to the quality or adequacy of the audio.

### 2.4.2 *Continuous assessment*

To this point, all the rating scales discussed have been post-hoc in nature. This means that they are administered after the stimulus material has been experienced. There are fundamental problems with such an approach. For example, it is not clear which part of the quality is being rated, as the judgment is made over the entire piece of material. Considering the variability of MMC quality, this problem becomes more marked.

The standard length of test material with the ITU scales is around 10 seconds, which is not long enough to allow the full range of degradations typical in MMC to be experienced. However, increasing the length of the test material to, for example 30 seconds can cause the recency effect to occur (Aldridge et al., 1995). This is where participants remember what they experienced most recently as opposed to the preceding material. Thus, it is clear that post-hoc assessment can be unsuitable in many situations. In order to attempt to address these problems, the SSCQE (Single Stimulus Continuous Quality Evaluation) was developed (ITU-R). The SSCQE was developed to assess the quality of digital television pictures in long test sequences. The tool is a slider of 10cm, which is labelled with the terms *"Excellent, Good, Fair, Poor and Bad"*. The position of the slider is registered at 500ms intervals and the means and standard deviations are calculated across all participants.

Studies using the tool have shown that it can be successfully utilised to assess perceived picture quality and also other variables, such as bit rate and delay. It has also been successfully used to assess bandwidth-impaired audio. It does recommend that the audio channel be included in the test material, which is sensible due to the interactive influence of the channels upon each other and makes the test material more ecologically valid. It has been found that there is a discrepancy between the post-hoc rating and the continuous ratings. This is to be expected because post-hoc ratings are unrepresentative

of quality over an extended period, as they are influenced by many factors. Thus, it can be assumed that the rating given by continuous methods for long segments of material may be less subject to bias.

In order to attempt to improve on the SSCQE and extend on the polar continuous scale (section 2.4.1.3), the QUASS (QUality Assessment Slider) tool was developed (Bouch et al., 1998) (Figure 4). This is a dynamic software slider that ranges from 0 to 100. A reading is taken of the position of the slider every second.

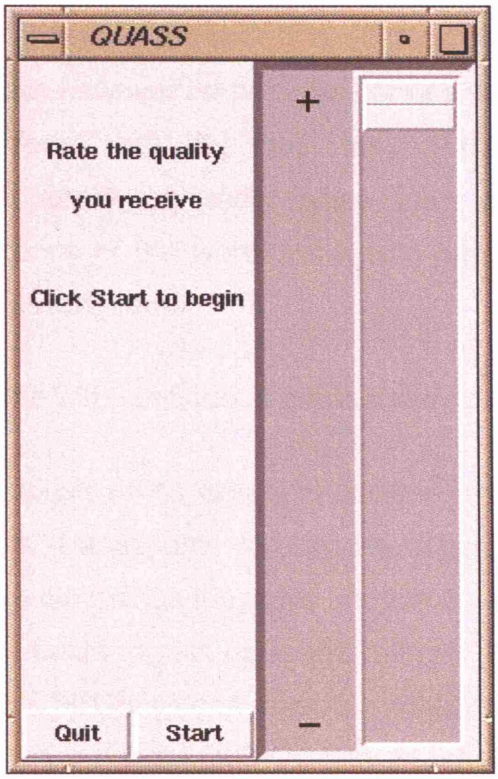



**Figure 4: QUASS interface**

A study using a passive listening task (Watson 2001) found that users could monitor time-varying changes in quality, which illustrates that users do not need quality labels on the rating scale in order to do this. An investigation was then performed to determine if the tool could be used effectively in a real-world multimedia conference (Watson 2001). Results showed that participants found it difficult to rate the quality whilst performing their main task, as the cognitive load was too great. Thus, it was concluded that QUASS can be used

effectively when the task is passive, however rating interferes with the user's task in a real-world context.

Passive rating does not give an indication of the quality required in a real-world context, thus it was decided to allow users to *control* the quality they received (Watson 2001). It was predicted that participants would only move the slider when the quality began to interfere with the task being performed. This would then allow the determination of what quality levels were good enough to allow the particpant to perform his/her task. Results showed that the main task (playing a word game) and the task of controlling the quality could be performed together, however participants commented that controlling the quality interfered with the main task. Participants found it *"impossible"* to operate the slider when they were engaged in providing descriptions of the word, yet it was attainable when they were listening to descriptions.

#### 2.4.2.1 Problems with continuous assessment methods

The act of continuously rating quality in addition to performing a task can be problematic. Participants can become fatigued, they can stop paying attention to the ratings they give and the severity and duration of the impairment could impact upon the ratings. Investigations into fatigue using the SSCQE (Aldridge et al., 1998) showed that in a comparison of two session lengths (10 and 24 minutes), participants in the longer session group were better at reacting to and identifying coding errors in video. So rather than becoming worse at the rating task over a long period of time, participants actually improved. Subjective data also revealed that fatigue was not playing a role. In the same study it was found that drift did not occur, at least in sessions up to 24 minutes in length. Finally, it was found that participants did not register the difference between 5 and 10 seconds of poor quality, thus it is implied that the duration of an impairment does not have an impact on overall judgments, yet the severity of it does. However, this was solely a rating task and it is likely that

results would be different when the user is rating the quality *and* performing a task.

## Chapter Summary

This chapter presented an overview of HCI and a background to evaluation. MMC and the degradations that can occur were then described. The impact upon communication of a mediated link and of media quality degradations were detailed. Finally, the current methods of assessing quality, which centre on the ITU recommended rating scales, were described and criticised for their inability to give an accurate indication of the effects that degraded media quality have on users. In addition, it was highlighted that subjective assessment does not give the full picture of the impact of degradations on users when used in isolation.

Taking the evaluation criteria from Shackel (1981) into account, it is clear that the attitude criterion is the only one being met by the current approach in MMC quality evaluation. Therefore, the research reported in this thesis adopted a 3-factor framework (see chapter 2), which incorporated measures of task performance, user satisfaction and the neglected element of user cost. There are subjective methods of measuring user cost, however due to the issue of cognitive mediation it was decided to investigate the use of an objective method: physiological signals.

# Chapter 3 Physiological Computing

## Chapter Aims

This chapter begins with an introduction to psychophysiology, starting with a description of the human nervous system and a focus on the autonomic nervous system (ANS). The sympathetic nervous system (SNS) response of SC, HR and BVP are presented. An examination of relevant psychophysiological concepts to the research reported in this thesis is then carried out. Despite the term *Physiological Computing* being coined relatively recently (Allanson & Wilson 2002a), there has been research using physiological signals in various areas of computer science since the 1980s, which has grown in popularity with the advent of portable physiological measuring equipment. This chapter gives a critical review of studies of relevance to the research reported in this thesis, which are separated into the areas to which they make a contribution.

## 3.1 Introduction to Psychophysiology

Psychophysiology is defined as *"... the study of relations between psychological manipulations and resulting physiological responses, measured in the living organism, to promote understanding of the relation between mental and bodily processes"* (Andreassi 2000). The psychological processes studied in this area are extremely diverse and range from sleeping and perception, to the body's response to stress. Physiological responses can be measured from the heart, brain, muscles, skin and eyes. Most measures are taken from the surface of the body, thus are viewed as being physically non-invasive. All of these techniques are aimed at learning more about the physiological substrates of behaviour.

### *3.1.1 Human Nervous System*

The nervous system of humans is divided into the central nervous system (CNS) and the peripheral nervous system. The CNS includes the brain and the spinal cord, whereas the peripheral nervous system

is comprised of nervous tissue outside the brain and spinal cord. This system allows communication between the brain and the spinal cord. The peripheral nervous system is further divided into the somatic nervous system, which consists of motor nerves that control voluntary muscle and the ANS.

### *3.1.2 The Autonomic Nervous System*

The ANS is the regulator and controller of many important bodily activities, such as digestion and aspects of emotional behaviour. Its activities are generally thought to take place without conscious control, hence its name 'autonomic'. The main function of the ANS is to maintain a constant internal body environment, despite internal or external changes that could alter the balance.

The ANS is divided into the parasympathetic nervous system (PNS) and the SNS. The SNS controls the activities that are mobilised during the 'fight or flight' response (Cannon 1915). Thus, the sympathetic reaction includes the expenditure of energy, the acceleration of HR, increased blood pressure and blood sugar, pupil dilation, an increase in blood flow to the voluntary muscles, increased sweating, and a decrease in the amount of blood sent to the internal organs and extremities. There are many evolutionary functions of the sympathetic reaction that increase the chance of survival. For example, the increase in HR and glucose helps to bring oxygen and extra nutrition to the muscles, and the increase in the tension of the muscles prepares the body to run or physically defend itself. Sweating cools the body in preparation for activity and may also protect the skin and dilated pupils allow more efficient vision, as they allow more light to enter the eye.

The functions of the PNS are rest, repair, and relaxation of the body and restoration of energy stores. This system controls responses such as decreases in HR and blood pressure, stimulation of the digestive system, constriction of the pupils, rest, sleep and sexual

arousal. The PNS and SNS act in a complementary way that allows for a smooth flow of bodily activities and behaviour.

The signals that were measured in the research reported in this thesis were SC, HR and BVP. These are all autonomic responses. They were selected because they are measured from the surface of the skin, thus are physically non-invasive. They provide a continuous measure over time as opposed to, for example measuring stress hormones[6] through a saliva sample, which can only indicate a difference between two points in time. Finally, they are widely used in the area of Physiological Computing, as will be shown in section 3.2.

### *3.1.3 Skin conductance*

The skin has two types of sweat glands: the appocrine and the eccrine. The eccrine glands are found in a large number of places over the body's surface, especially in the palms of the hands and the soles of the feet. It is these sweat glands that are important in the area of psychophysiology. These glands respond weakly to heat, yet strongly to psychological and sensory stimuli.

The term 'Galvanic Skin Response' is terminology that has in recent times been replaced with SC. The SC signal has both tonic (Skin Conductance Level (SCL)) and phasic (Skin Conductance Response (SCR), which are often referred to as a Startle or Orienting Response (OR)) elements[7]. SCL increases with a SNS response. The unit of measurement of SC is the microsiemen, and is also known as micromho. Throughout this thesis, the term microsiemen (ms) will be used. The mean level of SC (SCL), which was measured in the

---

[6] Stress hormones, such as cortisol, are secreted by the adrenal glands. Under stress, the body secretes higher levels of such hormones. Higher and prolonged levels of cortisol in the bloodstream can have many negative effects, such as a decrease in muscle tissue.

[7] Skin resistance, as opposed to conductance, can also be measured, yet SC is generally preferred as it is more suitable for averaging and performing statistical analyses.

research reported in this thesis, is usually between 2 and 20 ms (Cacioppo et al., 2000).

### *3.1.4 Heart Rate*

The heartbeat is representative of the contraction of the heart when pumping blood to other body areas. At rest, the average HR of a human is 72 beats per minute (bpm). The PNS slows the heartbeat, whereas the SNS produces an increase in HR. HR is determined by many factors, such as sympathetic activation and cortex factors, of which the psychological impact may be small, therefore HR should not be the only physiological signal measured.

In studies of human performance, it is the bpm or the Inter-Beat Interval (IBI) that are most commonly used. Generally, researchers interested in changes in heart activity that occur within a single cardiac cycle make use of IBI, whereas those interested in longer term changes that occur over a period of 30 seconds or more measure bpm.

### *3.1.5 Blood Volume Pulse*

Photoplethysmography refers to various techniques of measuring blood volume changes in a limb or segment of a tissue (Brown 1967). Shifts in the blood volume in various parts of the body occur in different mental and physical tasks. These depend on the arterial blood flow into an area and the venous outflow from an area.

Cook (1974) stated that there are two elements of plethysmographic change that can be measured: the slow engorgement of an area (blood volume) and a rapid component, which is referred to as pulse volume or amplitude. The latter represents the pumping action of the heart as represented in the local blood vessels. A decrease in BVP amplitude is caused by constriction of the blood vessels and this is caused by increased activity in the SNS.

### *3.1.6 Other signals*

There are many physiological signals available that could have been measured in the research reported in this thesis and that have been measured in similar areas by other researchers (see section 3.2). These are described in table 3.

| Measure | Description |
|---|---|
| Electromyograph (EMG) | Measures the electrical activity of muscles using electrodes attached to the skin e.g. from the jaw |
| Electrocardiograph (ECG) | Measures the electrical activity of the heart using electrodes attached to the skin. From this Heart Rate Variability (HRV) can be calculated, which is a measure of the oscillation of the interval between consecutive heartbeats. |
| Electroencephograph (EEG) | Measures the electrical activity of the brain using electrodes attached to the scalp. |
| Respiration | Measures the rate of breathing via a belt placed round the sternum or diaphragm. |
| Blood Pressure | Measures the force of the blood pushing against the walls of the arteries. Systolic blood pressure (SBP) is the maximum pressure, whereas diastolic is the lowest pressure. Measured from the upper arm or fingertip. |
| Skin Temperature (ST) | Measures the temperature of the skin. |

**Table 3: Physiological measures used by other researchers in similar area**

Attaching electrodes to the skin (in the measurement of EMG, ECG, EEG) is more invasive that sensors that go on the fingertips, therefore these were less suitable for the experiments carried out as part of the research reported in this thesis. The respiration sensor is a band that goes round the chest. People may be aware of it as their chest contracts and expands whilst breathing and this may remove focus form the task, thus this was not appropriate in this context.

ST is a slow responding signal (takes more than 2 minutes to respond), however this is the length of the shortest conditions in experiment 2, therefore it may not have registered a difference.

Blood pressure was an option to measure (if it was from the fingertip in order to be minimally invasive and to provide a continuous measure), however BVP is more commonly measured in the area of Physiological Computing (see table 4) and can be measured from the same sensor as used to measure HR, therefore was viewed as being more suitable.

## 3.2 Physiological Computing

### *3.2.1 Psychophysiology and the Electronic Workplace*

Gale & Christie (1987) published the first book that conjoined psychophysiology with information technology, the aim of which was to illustrate the relevance of psychophysiological research to the 'electronic workplace', which is *"...the application of information technology to the office".* The main argument behind the use of physiological measures was that subjective techniques and methods that interrupt the user's task cannot give an accurate, continuous indication of the user's experience in real-time. The use of measures of behaviour, task performance and subjective measures of mood were recommended to aid interpretation of the physiological signals.

Gale and Christie stated that psychophysiology is not *"...a coherent theory or integrated body of knowledge. Its identity comes largely from the particular set of techniques that have been developed by researchers working on a range of different problems from a similar general perspective."* This issue remains today and consequently practitioners in this area strive for one theoretical framework to reconcile the existing data, however this may always be out of reach due to the many complexities that operate in this area. A number of concepts that apply to the research reported in this thesis will now be briefly described.

### *3.2.2 Relevant theories*

#### 3.2.2.1 Arousal

Arousal is *"...the theoretical concept which has been most used in psychophysiological theorizing, ever since Lindsley (Lindsley 1952) postulated a continuous dimension of arousal, from extreme excitation and emotional disturbance through to deep coma"* (Gale & Christie 1987). Arousal is frequently described with reference to the inverted U shaped curve, where high levels of performance are accompanied with moderate levels of arousal and high and low levels of arousal are associated with poor performance. However, the theory of arousal has been widely criticised for the notion that the signals must either all increase or decrease in the same direction as evidence has been found to the contrary (Cacioppo et al., 2000) (see section 3.2.2.2). Gale & Christie (1987) conducted a review of the status of arousal as a theory. They reported that the search for an independent psychophysiological measure of arousal, such as HR, had been unsuccessful due to the signals being determined by many factors. Additionally, the association between physiological measures is too loose for them to be considered as all demonstrating the same process. Thus, it was concluded that the concept of arousal is unhelpful in this area.

#### 3.2.2.2 Directional fractionation

The intake-rejection school of thought has been dominated by the Laceys: (Lacey et al., 1963) and (Lacey 1967). Lacey (1959) and Lacey et al. (1963) gave examples of directional fractionation, which is situations where HR decreased and SC increased and situations where they changed in the same direction. They also presented evidence that tasks involving cognitive functioning are accompanied by an increase in SC and HR whereas those involving perceptual activities lead to an increase in SC and decrease in HR. It was suggested that the latter leads to an increased uptake of environmental stimuli, whereas HR acceleration attempts to reject

stimuli that would be disruptive to the performance of a cognitive task.

#### 3.2.2.3 Orienting theory

Orienting theory was originally devised by Pavlov (1927) and stated that the Orienting Reflex was a combination of behavioural and physiological responses to events, the primary function of which was to increase the organism's sensitivity to environmental events. This reflex (characterised by, for example increases in SCL, dilation of the pupils and decreases in HR responses) is thought to facilitate a possible response to a new stimulus. Due to the fractionation between signals, parallels were drawn between orienting and intake (see section 3.2.2.2). The OR habituates with repetition. This can be contrasted with the defensive response, where HR increases and does not habituate, as its function is to reduce sensitivity to environmental events. Thus, the OR can be compared to the rejection component of the intake-rejection theory (Graham & Clifton (1966), Sokolov (1963)).

#### 3.2.2.4 Workload and task difficulty

Physiological responses to workload reflect those of arousal. There is no definition and quantification of cognitive (mental) workload that is universally agreed upon. In the absence of this, existing theories are based on the premise that attention can be analogous to the limited processing capacity of a computer (e.g. Kahneman, 1973). However, as HR is multiply determined any measure of mental workload that comes from HR will be incorporating emotional factors, as well as the cognitive element. There is a direct relationship between HR acceleration and task difficulty (e.g. Kahneman et al., 1969) and HR deceleration and task difficulty (e.g. Duncan-Johnson & Coles, 1974) depending on the information processing requirements of the task. SCRs are also related to task difficulty and complexity (e.g. Kaiser & Sandman, 1975).

### 3.2.2.5 Integrating the theories

As has been illustrated in section 3.2.2, research in this area is fragmented as a result of different measures used and different states being measured, thus no universal theory exists. The risk is if this integration problem is not addressed, the use of psychophysiology in HCI and in related areas will diminish. An attempt to integrate the field was made by a workshop at an international conference (Allanson & Wilson 2002b), which gathered together many of the researchers working in the area. One of the questions tackled was: *"What can we, as a community, do to ensure that Physiological Computing has staying power this time around?"*, from which several key themes emerged.

- The field needs proper leadership.
- The field will have more longevity if behaviour is incorporated, which was proposed by Gale & Christie (1987).
- Techniques that are used in the area need to be standardised.
- This area is frequently over-hyped, thus researchers in the community need to be explicit about what they can and cannot do, the latter of which is frequently overlooked.
- Sensors should be integrated into clothing and movement should be built into them to make them less restrictive and field trials more amenable.

In order for the field to progress, Gale & Christie (1987) recommended that more longitudinal studies and field trials take place to increase the bulk of research in these areas, thus bringing them into line with the more commonly used short-term lab-based experiments. In addition, they called for more real-life tasks to move away from artificial lab-based tasks with the knowledge that there will be confounding variables. A review of research in a similar area to the research reported in this thesis will now be presented. Table 4 gives a summary of the studies that will be discussed, in order to allow a comparison to be performed.

| Study | Signals | Partic | Bl length | Other measures | Design | Features used |
|---|---|---|---|---|---|---|
| Dillon et al. (2001) | HR, SC | 119 | 20 secs for SC and 100 secs for HR | Subjective | Mixed | Means |
| Dillon (2002) | HR, SC | 24 | 1 minute | Subjective | Within | Means, maximum positive and negative deviations |
| Picard & Healey (1997) | SC, HR | 5 | Not stated | None | N/A | HR - mean 10 secs before task minus mean 10 secs after task SC – diff btwn bl and 1st sig local max |
| Healey (2000) | SC, HRV, RESP and EMG | 6 | 15 minutes | Subjective, video taped participants | Within | Set of 22 features |
| Mandryk & Inkpen (2004) | SC, EMG, HR, Respir. | 10 | 5 minutes | Subjective, TP | Within | Means |
| Meehan (2001) | ST, HR, SC | 10-52 | 40-90 secs | Subjective, behavioural | Within | Means |
| Rowe et al. (1998) | HRV | 13 | 5 minutes | Subjective | Within | Not stated |
| Scheirer et al. (2002) | SC, BVP | 24 | Not stated | Behavioural measure, video taped participants | Within | Set of 5 features |
| Simons et al. (1999) | SC, EMG, HR | 34 | Not stated | Subjective | Within | ORs, ½ sec averages for HR |
| Ward et al. (2001) | SR | Not stated | 10 minutes | None | N/a | Variance and number of ORs |
| Ward & Marsden (2003) | SC, blood volume, HR | 20 | 5 minutes. 1st minute of task used as baseline | TP | Btwn | Normalised SC, mean HR and BVP % change from baseline. |
| Wastell et al. (1982) | EEG, HR | 15 | Not stated | Subjective, TP | Within | Bpm, HRV |
| Wastell & Cooper (1996) | HR, SBP | 18 | Baseline was paper based system | External work demands, subjective | Within | Levels at a specific point |

**Table 4: Summary of methods used in studies reported in chapter 3**

Key to table 4:

Partic = participants

TP = task performance

Btwn = between-subjects design

Bl = baseline

NB. The Bersak et al. (2001) publication is not included in this table as it was not an experiment.

### *3.2.3 Usability*

#### 3.2.3.1 Identifying significant HCI events

The preliminary study reported in Ward et al. (2001) involved participants in five conditions, three of which were intended to represent the following events in computer interaction: a web search for photographs of Huddersfield; extracting information from an information resource on the web, during which an alert box appeared on the screen that was accompanied with an auditory tone and extracting information from a web-site, which had usability problems. The other two conditions were a baseline session of ten minutes and taking three deep breaths, which were performed to illustrate the effect that necessary bodily functions can have on Skin Resistance[8], which was the only measure taken. Variability was defined as a decrease of 3% or more over a 10 second period. An OR (see section 3.1.3) was defined as a sudden decrease of 7% or more over a 10 second period.

During the baseline session, a 'settling down' period of 2-3 minutes was observed, after which an increase in resistance was found (from 400-800 kOhms). This is indicative of participants relaxing. When the first 3 minutes were excluded, no sudden changes in SC occurred. No data points were 3% or more lower than the data point 10

[8] This is the inverse of SC and is measured in kOhms, thus it decreases with a SNS response.

seconds before it. Thus, this session was classed as low stress, with 0% variation, 0 events and 0 ORs.

During the 10 minute photograph searching task, responses showed a similar pattern to the first 3 minutes of the baseline session. SR then stayed in the 400-500 kOhms range. 3.1% of data points decreased 3% or more over 10 seconds, but none were more than 5%. This was attributed to maintenance of arousal levels by orienting to elements that needed attention. However, there were no drops greater than 7%, so these cannot be classed as ORs by the author's own criteria. In addition, without subjective responses it is difficult to attribute these results to a maintenance of arousal levels. This session was classed as medium stress, had a variation of 3.1%, 0 events and 0 Ors.

The breath sequence lasted for 10 seconds. There was a 2.5 second delay after the impact of the first deep breath registered in the signal. SR then decreased by 20% over 9.5 seconds. It recovered 2 seconds after normal breathing resumed. 7.2% of data points showed a decrease of 3% or more over 10 seconds. This session was classed as low stress, with a variation of 7.2%, 1 event and 1 OR. However, it is not sensible to class this event as being of low stress, as the signal was responding purely to a physical event, which would outweigh any responses to psychological events.

In the information extraction task (which lasted for 14 minutes) there was a one second latency following the appearance of the alert box (which appeared after 5 minutes), then a decrease in SR from 440-161 kOhms over the next 9 seconds. This is defined by Ward et al. as being an OR. There was also considerable variability, with 15.1% of data points decreasing by 3% or more over 10 second periods. This session is classed as high stress, a variation of 15.1%, more than 1 event and 4 ORs. However, responses were due to the alert box not the rest of the task, thus this session in its entirety should not

be classed as high stress. In addition, the classification of the appearance of the box as a highly stressful event is questionable because it startled the participant, which is different to causing stress.

In the final session, there was minimal variability. There was only one drop of 9.7% when the participant was looking for the price of the book. Excluding the first 3 minutes, 7% of the data points decreased by 3% or more over 10 seconds. It was classed as medium stress, had 7% variation, 1 event and 1 OR.

It is noted that these results were replicated using different participants, and that further investigations were being carried out to determine their generalisability. However, it is not stated how many participants were used and whether they all responded in the same way.

From these results the authors concluded that SR can be used as a measure of stress and that HCI events can be categorised according to the stress they produce. In addition, the number of ORs and variability are suggested as being two features of SR that can be used in this area. However, as mentioned earlier it is not clear how many participants were involved. Furthermore, only three HCI type events were investigated and are narrow in their generalisability. Moreover, looking at short-term ORs, for example to the alert box, has minimal use in HCI when longer-term (i.e. more than 10 minutes) responses to a technology are measured. In addition, subjective assessment methods were not used in this study, therefore making it more difficult to interpret physiological responses. Finally, the conclusions concerning the two useful characteristics of SR to measure do not mention overall levels and interpreting variance and the number of ORs without looking at overall levels, would be flawed.

#### 3.2.3.2 Evaluating well and poorly designed web sites

A follow-up study (Ward & Marsden 2003) involved a historical directory of a Yorkshire town. There were two directories, one of which was well designed through the adoption of good web and information design and the other was poorly designed with excessive use of pull-down lists, poor navigation cues and functions and functionless animation and adverts that either caused screen content to change position or they appeared in pop-up boxes, which had to be closed before the user could continue.

There were twenty participants in this between-subjects experiment who had to answer questions in 10 minutes, using either the well or poorly designed website. Within-subjects designs are more common in the area of psychophysiology, as they allow each participant's response to each condition to be compared, rather than comparing responses from different participants who will have different magnitudes and response ranges.

SC, blood volume and HR were taken as measures of arousal. However, as mentioned in section 3.2.2.1, arousal is a concept in psychophysiology that has not fared well, which is due to the finding that signals can directionally fractionate. Thus, this is not an appropriate term to use. A five-minute settling in period was given before the experiment started. The first minute of physiological signals in the task was used as a baseline for each participant. SC was normalised (by subtracting the baseline and dividing by the range) and expressed as a % change. Mean HR and BVP were measured.

It was found that SC decreased by a mean of 0.0308ms in the 10 seconds before the pop-up adverts appeared, compared with an increase of 0.026ms in the 10 seconds after the advert appeared. This measure was the only significant result to have come out of this experiment. The lack of significant results was attributed to

differences between participants outweighing responses to the directories. This is most likely due to the design being between-subjects and offers support for the utilisation of within-subjects designs in such experiments. It was found that users of the well-designed directory answered an average of 22 questions, whereas users of the poorly-designed directory answered an average of 12 questions. However, again no subjective data was gathered.

The authors concluded that, despite the lack of significant results, mean responses, as opposed to variability and the number of ORs as recommended in their previous study *"...are able to distinguish differences in arousal levels in different computer based situations and therefore can provide an indication of software usability"*. Despite the mean levels responding in the direction predicted, it would have been more convincing to obtain significant results, which may be achievable if the experiment was replicated with a within-subjects design.

### *3.2.4 The impact of moving images*

An example of directional fractionation was found in a study by Simons et al. (1999), which investigated the effects of moving and still images. 34 participants saw 27 images that were extracted from films and television programmes. Each participant saw a still and moving version of the same image for six seconds, thus the experiment employed a within-subjects design. However, responses may have been influenced by boredom or frustration as a result of seeing exactly the same images, albeit one in video, twice. In addition, extracting a still image from a television show is not an ecologically valid piece of experimental material, as it has been taken out of context. Moreover, there is not a task to become engaged in, other than subjective rating, as applies with the ITU scales (see section 2.4.1.1). Therefore, these results should be applied to interactive MMC with caution.

The images were associated with a wide range of ratings on the emotion dimensions of arousal and valence[9]. SC, facial EMG and HR were measured. In addition, participants gave valence, arousal and dominance subjective ratings to each of the images. A complicated method of analysis was used for the physiological signals. SCR magnitude was defined as the difference between the largest peak (that occurred between 0.5 and 4 seconds after the condition began) during the condition and the onset of the response. However, these were identified visually, thus some degree of error will be expected. In addition, it is not stated how many people did this, therefore inter-rater reliability may have been a factor. The HR data was analysed by dividing the data stream into 14 half-second averages. Such a complicated method may have been necessary to detect differences in the short conditions (6 seconds).

The results showed that SC increased linearly according to the arousal properties of the stimuli and also increased for moving images. EMG was significantly related to the valence properties of the images. HR decreased shortly after the presentation of both still and moving images and remained below baseline for the duration of the condition. HR change was significantly related to valence (with positive valence being higher than neutral and negative) and arousal (with high arousal being lower than medium and low). Motion also had a significant impact on HR, with moving images causing HR to slow more than still images. Subjective responses showed that moving images were rated more positively and as more arousing than still images.

Image motion was concluded to influence the arousal of an image more than a participant's opinion of the content. The difference in deceleration between the moving and still images grew more marked towards the end of the viewing period. This was taken as evidence

[9] At a basic level, emotions can be described by their strength (arousal) and by their meaning e.g. a positive or negative feeling.

that there are several processes involved in the HR response. Firstly, a short deceleration occurs, which indicates orienting to the image. Secondly, the affective properties of the image dominate and finally motion becomes the dominant factor. HR returns to baseline if the image is still but if it is moving HR remains slowed. Simons et al. posit that this reflects sustained attention and that motion continues to present new information to viewers and holds their attention once it has been captured by content. This study, despite its limitations, shows that a decrease in HR may be expected with MMC due to the video and that any differences in the arousal properties of the multimedia conference may be reflected in SC. It also illustrates the differences in features that can be extracted from physiological data.

### *3.2.5 Affective computing*

Roz Picard of the Massachusetts Institute of Technology (MIT) Media Laboratory is widely recognised as being the founder of Affective Computing: *"...if we want computers to be genuinely intelligent and to interact naturally with us, we must give computers the ability to recognise, understand, even to have and express emotions"* (Picard 1997). The rationale behind this is based on the essential role that emotions play in decision making, perception and learning, thus in order for computers to be intelligent and to interact naturally with their users it is claimed that they should be emotionally intelligent. One of the main ways in which emotion is detected at the Media Laboratory is through their physiological responses. The Affective Computing group research served as a starting point to the research reported in this thesis, as it showed that physiological signals can be successfully measured in a number of areas in computer science and to measure different states, from stress and frustration through to emotions. In addition, the physiological measuring equipment was the same as was used in the research reported in this thesis (see section 4.3.4).

#### 3.2.5.1 Ambulatory sensing

One of the aims of the Affective Computing research group is to develop wearable sensors, for example the Galvactivator (Picard & Scheirer 2001), and wearable computer systems that allow physiological signals to be measured over long periods of time and also potentially allow the application to learn and respond to a user's state (e.g. Healey & Picard (1998) & Healey et al. (1998)). However, in order for this to occur it is necessary that the data is accurately labelled in order to interpret results. This is one of the main challenges of ambulatory affect sensing because physiological responses due to emotion can often be less strong than responses due to, for example walking or coughing. In the experiments conducted in the research reported in this thesis, participants were stationary at a desk, thus movement was minimal. This is one way of addressing the issue. However, in tasks where movement is necessary, constraining it detracts from the ecological validity of the study. Therefore, incorporating it is difficult but necessary.

To estimate the extent to which movement can affect physiological responses, Picard & Healey (1997) performed an experiment involving five participants performing the following activities: standing, walking, jogging and coughing. There were two walking tasks but it is not made clear whether the data were averaged over the two sessions. In addition, there was little consistency on the rests given between the tasks. For example, after the second task, which was to walk round the room for one minute, there was a rest of two minutes where participants sat in a chair. However, after the jogging task, which lasted for one minute, there was only a rest of one minute. Additionally, after the standing task there was no rest. These rests presumably serve the function of returning the levels to their baseline, yet this may not have been achieved with such short breaks. In addition, the fact that they differed in their length after tasks makes results difficult to compare.

SC (from both the hand and foot[10]) and HR were measured. For SC, the difference between the baseline and the first significant local maximum was calculated. For HR, the average HR ten seconds before the task began was subtracted from the average HR ten seconds after the task. The results showed that there can be large changes in the values of the signals due to physical events. For example, there was an increase in HR of 53.7bpm whilst coughing for one participant and an increase in SC measured from the hand of 16.1ms when jogging for another participant. When it is considered that in psychological studies (e.g. Lang et al. 1993) and Winton et al. 1984), the greatest increases in SC and HR can be small (0.6ms and 8bpm) this study highlights that simple physical activities can potentially outweigh responses to other variables. However, there were a number of discrepancies in the data that are not explained. For example, for one participant the SC from the hand is greater during walking than jogging, the opposite of which is true for the other participants. In addition, the SC from the foot of another participant is higher for standing than jogging, a result which was not found in the other participants.

Despite this being a small, anecdotal experiment the results illustrate the sensitivity of physiological measures and provide support for the experiments used in this thesis being desk-based, with the exception of experiment 8 where the data were annotated.

#### 3.2.5.2 Detecting driver stress

The main experiment that constitutes the PhD thesis of Healey (2000) detected stress in car drivers. The ultimate aim of the research was to automatically manage applications in a car, such as a mobile phone, based on the driver's stress levels to ensure that the driver was not distracted when he/she was under stress. The car was

[10] The SC was taken from the foot as it was viewed as being more practical for ambulatory experiments. A further experiment detailed in this paper showed that SC from the hand and foot are highly correlated.

also seen as an ideal situation in which to measure long-term changes in a person's stress level because a drive performed frequently will encounter a relatively constant sequence of events that can be compared from day to day. In addition, a car does not facilitate free movement, which makes it suitable for monitoring physiological responses.

A system was developed using an onboard computer with video cameras (to account for movement), a microphone, and four physiological sensors (SC, ECG, Respiration and EMG). Data from 6 participants were used, which is a very small number. The task was a ninety-minute drive, which followed a set route through fifteen different events. From the results of questionnaires completed by the participants, four categories of stress level were created: low, neutral, high or very high stress, and 545 one-minute segments of the physiological signals were classified as belonging to one of the four stress categories.

A number of features were extracted from each signal and a linear discriminant function was employed in order to rank each of these features individually based on their recognition performance. An algorithm was then used to find an optimal set of features for recognition patterns of driver stress. The results of this showed that by detecting patterns across multiple features, performance for recognising stress is 88.6%. Motion artefacts did not affect this result. However, of the six participants there were problems with the data set for three of them (caused by problems with the sensors for two of them). It would be beneficial to have had more participants to examine whether the algorithm works as well with a larger data set. In addition, it needs to be tested on a route that is not pre-set by the experimenter, in order to gain ecological validity.

The results from this experiment are encouraging for the automatic detection of stress and the potential for its application in cars is

substantial. The method of analysis is more sophisticated than used in the research reported in this thesis, which detects simple changes from baseline, because Healey was attempting to automatically detect stress, which is more complex. In addition, the measurements were taken in a real-world scenario with a number of variables operating as opposed to strictly controlled lab-based experiments, thus such sophisticated analysis was necessary. However, the number of data sources used (physiology, video, self-report) and the use of multiple signals gives support for the 3-factor evaluation framework as used in the research reported in this thesis.

#### 3.2.5.3 Inducing and detecting frustration

An experiment performed by Scheirer et al. (2002) is a starting point to address the issue of contextualising the physiological responses of a user within their surroundings. It involved 36 participants, although accurate data was only obtained for 24 (problems with the sensors meant that data for 3 participants had to be dropped). The task was to play a computer game with the goal of completing a series of visual puzzles as quickly and accurately as possible to gain a reward. However, at repeated intervals the mouse was set up to behave as if it were faulty. This was done by the experimenters with the aim of eliciting frustration in the participants, which is known to cause an increase in physiological arousal. Several streams of data were gathered for analysis: mouse-clicking behaviour; SC; BVP; a video of participants and of the elapsed time of the experiment and events in the game. However, subjective measures of frustration were not taken, thus some responses that are classed as being frustration may be of another state, such as stress.

Features were extracted from the physiological data and an automatic technique for classifying the features using Hidden Markov Models was developed. The technique was significantly better than random for 21 out of 24 participants at classifying events where frustration was likely from those where it was not, which suggests

that there is some discriminating information in SC and BVP. Mouse-clicking behaviour was also synchronised to events where frustration was likely and this revealed distinct behavioural responses to stimuli.

This experiment shows that physiological signals can be used to detect states, other than perceptual strain, when interacting with computers. However, it induced frustration in order to generate an algorithm that would allow the computer to automatically detect when the user was frustrated and potentially act upon this information. The research reported in this thesis uses physiological signals to detect responses to media quality degradations, the information from which is aimed to improve MMC. So rather than an adaptive tool as in the Affective Computing research, the research reported in this thesis is using physiological signals as an evaluation tool.

#### 3.2.5.4 Affective feedback

In 2000-2005, the Media Lab Europe in Dublin operated as a joint venture between MIT and the Irish Government. The MindGames research group used a gaming framework to investigate 'Affective Feedback'. This is an extension of biofeedback, which involves teaching people to understand and control their physiological responses and is used in areas such as the management of anxiety and phobia. The aim of the research was to furnish the user with skills that can be utilised on an everyday basis.

An example of this is given in Bersak et al. (2001). The 'Relax to Win' video game was developed as an aid to therapists in the treatment of children with anxiety disorders. It involves two people competing against each other. Each player's SR controls a 3 dimensional dragon in a race. The dragon can walk, run or fly and this is controlled by the player's SR, with flying occurring when the player is relaxed.

In this game, SR is used as a measure of relaxation. The general trend of SR increasing was used in an algorithm to indicate relaxation, as opposed to a change from baseline. Trials with users showed that they have more control over their stress levels and perform better the more times they play the game. It was also observed that the game, which could be regarded as being stressful, seemed to enhance the feedback, as the users were aware of their performance and were keen to win. The researchers predicted that playing this game would facilitate relaxation in real-life.

This investigation was noted to have produced a proof of concept, yet the fact that the system was in its infancy was acknowledged. Future work aimed to include many physiological signals, such as HR, EMG and EEG, and input from the user's behaviour using facial recognition and gestures in order to *"completely characterise the player's physiological state"* in a fully immersive environment. The researchers were also interested in creating wireless sensors due users feeling constrained by the traditional sensors. This project gives an exciting example of the potential of physiological signals to modify applications and in turn to modify the state of users. However, the measurement of additional signals is necessary to ensure that the application is inferring the correct state of the user.

In conclusion, whilst the Affective Computing and MindGames groups operate from the premise that recognising and responding to user emotions (through physiological signals) is advantageous, the research reported in this thesis puts forward the idea that detecting basic changes in physiological signals is of more immediate benefit to evaluation in HCI for two reasons. Firstly, despite Picard et al. (2001) developing an algorithm which had a more than 80% accuracy for long-term emotion detection, Cockton (2002) points out that there is a longstanding debate about whether emotions can be inferred from physiological signals and this is likely to continue for the foreseeable future. Secondly, it is questionable whether knowledge

of the emotions the user is experiencing when interacting with MMC is necessary. It is more important that they can a) perform their task, b) are not being put under adverse amounts of pressure due to the quality and c) they are satisfied with the quality. Thus, the research reported in this thesis argues that physiological measures such as the ones employed by the Affective Computing group can be employed in a more traditional context of usability assessment.

### *3.2.6 Workload*

#### 3.2.6.1 Evaluating the impact of a modernised telephone system

The research of Wastell has used a similar evaluation approach to the research reported in this thesis to measure workload in the field and illustrates the many layers of information that taking such an approach offers. In a study by Wastell et al. (1982) physiological measures were employed to determine why a system that required less work on the part of the operator (a cordless telephone switchboard) induced a decrease in task performance and job satisfaction.

The study took the following measurements from 15 female operators who were experienced with both types of switchboard: subjective experience (mood questionnaires), task performance (call traffic levels and task related events) and physiology (EEG, HR and HRV). The data were averaged over 15 minute periods. Measurements were taken for one day with both types of switchboard when the operators performed their daily jobs, thus a within-subjects design was employed. Results showed that the cord-based design induced feelings of greater well-being and had higher levels of performance for effective call connections. Physiological data showed that HR was faster on the cord design (89 vs. 85.5bpm), but this difference was only reliable at the 10% level. This correlated with the greater workload and with the higher levels of physical activity required with the system. There were no significant

differences between the two systems in EEG or HRV. In a detailed analysis of call-traffic variation, operators working with the cord system showed lower levels of mental effort, which was shown by an increase in HRV (whereas a decrease in HRV reflects an increase in mental effort).

Therefore, the cord-based system, which required more workload than its replacement, evoked superior performance and job satisfaction without showing higher levels of physiological stress. This finding was attributed to the cord system being flexible and allowing operators to offer a high quality service. However, it is extremely difficult to determine for each individual the level that, if exceeded, would result in stress. In addition, it is possible that some of the operators were put under stress by the old system but did not notice because they were so used to it, therefore taking long-term measures of, for example stress hormones could help to investigate this. In addition, subjective data showed that the operators had a *"...general disposition against cordless switchboards",* which also may have influenced results. Finally, it is unlikely that participants were as experienced in both systems, thus the comparison may not have been equal because factors such as task learning and novelty may have been in operation with the new system, which may have increased their levels of physiological stress. There were a lot of variables operating in this study. The research reported in this thesis avoided some of these problems by employing more abstract studies with fewer variables.

#### 3.2.6.2 Measuring the impact of computerised ambulance control operations

In a second study (Wastell & Cooper 1996), a similar approach was used to evaluate the computerisation of ambulance control operations. The impact of the new system was measured for a period of 6 weeks before and 4 months after its implementation. The measures taken were: external work demands (the number of

simultaneous jobs being handled by the despatcher at the time of sampling); HR; SBP from a fingertip sensor; two aspects of subjective state (anxiety and fatigue) using visual analog scales and a post implementation questionnaire. 18 despatchers gave data for the real-time measures (physiological measures, external work demands, anxiety and fatigue).

The assessment of external work demands showed clear evidence of an improvement in performance with the new system. Responses from the post-implementation questionnaires showed that the despatchers said they felt the new system had improved their job satisfaction. The differences in stress and ability to cope were not significant between the two systems. Both subjective anxiety and fatigue increased significantly with workload in the paper based system. With the computer based system, the increase in anxiety was significant but not for tiredness.

For each despatcher, two data points were extracted relating to when they were least heavily loaded and under the most pressure. This was determined from the number of simultaneously active jobs. The differences in HR and SBP for the paper based and computerised systems were not significant, however HR and SBP increased (significantly for SBP) as a function of workload for both systems. When the analysis was made more individual by calculating the change in SBP divided by the difference in workload for high and low workload conditions, the differential increase in SBP was found to be statistically significant ($p<0.01$) for the paper based system.

The length of time that despatchers had been working with the old systems may have contributed to the lack of significant differences in HR and SBP when the data were averaged over all participants. It may be that despatchers who were less used to the old system would have adapted quicker, thus their responses may have been drowned out by those who had used the old system for a long time

and took longer to adapt to the new system. On the other hand, it is not surprising that HR did not show any significant differences. This was a field trial, so there would have been many variables at work in the environment and also in the job itself, which can be extremely stressful due to factors including its cognitive complexity and the uncertain dynamic environment. HR may have been too fine-grained a measure to detect overall differences due to the system.

Wastell was one of the first researchers to use physiological measures in HCI and to realise that *"psychophysiology has an immensely valuable role to play in deepening our understanding of HCI, especially in real-world settings and helping us to design better systems"* (Wastell 1990). The evaluation approach used by Wastell paved the way for the 3-factor approach adopted in the research reported in this thesis, as it showed how physiological measures may give a misleading picture if used in isolation. For example, in the telephone switchboard study (see section 3.2.6.1) the observed increase in HR could have been due to stress, yet the corresponding behavioural results showed better task performance and subjective responses showed greater feelings of well-being. The same is also true for using subjective assessment or task performance in isolation. Wastell's research is also important for illustrating that physiological measures can be effectively used in the field with a number of variables operating. That is the next step for the research reported in this thesis, however it had to begin in a strictly controlled experimental setting in order to determine if responses to media quality degradations could be detected. As has been shown, the research reported in this thesis has its grounding in the measures used by the Affective Computing group (as reported in section 3.2.5) and the evaluation methodology utilised by Wastell.

#### 3.2.6.3 Measuring mental effort

A study related to the research of Wastell was performed by Rowe et al. (1998) who performed an experiment to explore the use of HRV as an indication of the state of users. The scenario they used was an

air traffic control computer game, as this allowed the degree of mental effort to be manipulated. Participants had to monitor planes on a screen, where aircraft targets moved across the screen following fixed routes and others followed off route ('free flight') paths. The aim of the games was to keep planes at the same altitude from colliding. Mental effort was predicted to rise with the number of free flyers. This task is well suited as a starting point to determine how HRV responds to mental effort, however it is not ecologically valid (except in the context of air traffic control).

A total of 13 participants (5 of whom had experience with air traffic control) took part and had to play 5 games, which were a control game with no free flyers and games with 4, 8, 12 and 16 free flyers. Baseline ECG was measured for 5 minutes before the experiment began and a training and observation session was given. After each game participants completed rating in six workload areas: mental demand; physical demand; time pressure; performance; effort and frustration. The order of the presentation of conditions was not randomised, which may have affected results. In addition, this is not a large sample of participants, however the authors acknowledge this. An additional flaw is that the number of people with and without experience of air traffic control is not balanced.

Analysis of the mental demand scales showed that free flyers had a significant effect on the subjective measure of mental effort. Contrasts between the control and conditions were significant for the 12 free flyer condition. The number of free flyers also had a significant impact on overall workload (a composite of the six sub-scale scores. Across all participants there were no significant differences in HRV to the conditions. However, the group with air traffic control experience showed significant sensitivity to the number of free flyers: there was a consistent decrease in HRV as free flyers increased from 0 to 12, which indicates an increase in mental effort. However, the increase from 12 to 16 free flyers resulted in an

increase in HRV, indicating a decrease in mental effort. The authors interpreted the latter result as being evidence that as the number of free flyers rose from 12 to 16, participants could not handle the data they were presented with and subsequently disengaged from the task. Performance data was gathered, however it does not appear to have been analysed, which is a limitation of this study because it is vital to see how an increase in mental effort impacts upon task performance.

The authors suggested that HRV could be used to predict and uncover problems with interfaces and that interfaces could adapt to the state of the user. Given the differences that can occur between physiological signals, measuring more than one signal would give additional data and would allow further conclusions to be made. Despite its limitations, this study illustrates a clear effect of mental effort on HRV and opened many doors for the use of physiological signals in HCI studies.

### *3.2.7 Presence*

#### 3.2.7.1 Evaluating immersive television

Another area in which the use of physiological signals has been applied is that of presence, which is defined as *"...the observers subjective experience of 'being there' in a remote environment"* (Freeman et al., 1999). Cath Dillon has performed two studies investigating, most fundamentally, whether physiology can be used as a measure of presence in the context of immersive television.

Dillon et al. (2001) investigated the impact of three variables, each with two conditions: content of a clip (either a 100 second boat or rally driving sequence); Vistral (either on or off); and view (either monoscopic or stereoscopic). A Vistral is described by Freeman et al. (2001): *"The negative cue associated with the edge of the screen is reduced by using a Vistral screen surround manufactured by CRL.*

*This picture frame device generates a Moire effect from two layered patterns of dots either side of a glass plate, and it is extremely difficult to focus the eyes on. The results is that the stereo image really does appear to float, unattached from the back wall of the PIT, and gives an alarming sense of reality".*

Content and Vistral were between-groups factors, whereas view was a within-groups factor. This is a complex experimental design and as a starting point it may have been simpler to look at the differences due to one variable, such as content. If a significant difference was found, then further variables could be investigated. In addition, it would have been more methodologically sound to have had content and Vistral as within-groups factors to allow a direct comparison to be made for each participant on each of the factors.

A total of 119 participants were involved and they viewed both the monoscopic and the stereoscopic versions of either the rally or boat sequence with the Vistral on or off. This is a large sample of participants, which was most likely due to the complexity of the design and the presence of between-subjects factors. The assessment measures taken were: HR and SC recordings 100 seconds before, during and after each of the two presentations that participants viewed; a subjective measure of presence (ITC-Sense of Presence Inventory) following each presentation; and a subjective measure of mood (POMS: Profile Of Mood States) before and after each presentation.

This is a similar evaluation approach to the research reported in this thesis with the exception of the task performance element. Passive viewing without the context of a task is similar to the approach taken by the ITU (see section 2.4.1.1), yet the problems with this are twofold. Firstly, the task of watching a rally car or boat for one and a half minutes is not representative of television watching in the real

world. Secondly, without the context of a task it is difficult to maintain the attention and engagement of participants in the task.

SC was analysed by dividing the 100 seconds into 5 segments, from which the baseline (20 seconds before each video began) was subtracted, therefore different baselines were used for each clip. The mean baseline HR (the 100 seconds before the video began) was subtracted from the mean 100 seconds viewing for each participant.

The results showed that stereoscopic video received higher presence ratings than monoscopic video, which was expected. The boat clip was rated as being higher in presence than the rally clip, however this result may have been influenced by familiarity, as many of the participants in this experiment had visited the area in which the clip was filmed. Participants reported more negative effects for the rally clip than the boat clip. This was most pronounced when the Vistral was on, which the authors attributed to the Vistral being disorienting. The POMS results showed that the experience of viewing stereoscopic video was more positive than monoscopic video, regardless of the content of the clip.

SC significantly reduced over time. A main effect of time was found in the Vistral on and off conditions and there was a significant interaction between time and content in the Vistral on condition. In this condition, SC reduced more during the boat sequence than the rally sequence and this was most obvious at the end of the clips. There was a main effect of content in the HR data.

The physiological signals did not respond in the same way as the presence or mood measures, which the authors interpreted as evidence against physiological signals being used to measure presence. Yet, as Wastell (1990) points out, it is more interesting when behaviour and physiology do not concur because this indicates that a more accurate picture of the user's experience is being

obtained. Dillon et al. suggested that, due to SC discriminating between contents when the Vistral was used, SC may be related to some effects of the Vistral that are unrelated to presence and that SC may be useful in examining negative physical consequences of viewing, whilst HR may be useful for examining different contents. The authors came to the conclusion that physiological measures can give useful additional information, yet should not be used in isolation.

A second study (Dillon 2002) explored the impact of visual angle and the content of clips (emotionally neutral, amusing and sad) on presence ratings, HR and SC. Both variables were within-subjects, which is better suited to the use of physiological measures. 24 participants viewed the twelve clips (four in each category) at either 21 or 42 degrees one week and the week after saw the clips again at the other angle. No detail is given about the length of the clips, therefore it is difficult to judge whether they will be representative of television viewing in the real world. Despite this, boredom due to watching the same clips twice may have influenced responses. Alternatively, some adaptation to the content of the clips or the visual angle may have occurred. In addition, taking physiological measurements a week apart may have affected results, as participants may have been in a completely different physiological state compared to the other week and this may have had nothing to do with the experiment. Therefore, comparing visual angle within the same experimental session may have been more effective

HR and SC readings were taken during each clip and for 100 seconds before and after each clip. The mean of the last 60 seconds prior to each clip being viewed was used as a baseline measure for that clip. These means were then subtracted from: the mean when the clip was being watched; the 1$^{st}$, 2$^{nd}$ and last 60 seconds of each clip; and the maximum positive and negative deviations during the first two minutes of each film period. The means from the last 60 seconds of each clip were used as a second baseline period and

were subtracted from the means of the 60 seconds recovery period. This method of analysis may be overly complicated and its use is not justified. It may have been more effective to determine if simple differences from the baseline could be detected and if not, then to explore the use of alternative analysis techniques.

The subjective results showed that the amusement and sadness categories were representative of those emotions and were rated as being more arousing, interesting, engaging, of better image quality and adequacy and lower in negative effects than the neutral clips. In addition, the sadness category was rated higher than the neutral category on a question regarding whether the television environment at times became more real or present than the real world. There was no effect of angle on ratings.

SC was generally higher during the amusing clips (except on the maximum negative deviation from baseline where the amusement and sadness categories were higher than neutral). This result shows that an increase in SC does not always indicate a negative state. There was a lower change from baseline for the 21 degree angle, which indicates that there was less presence with the smaller angle.

HR reduced during the amusing and sad clips. Dillon states that the maximum and minimum deviations from baseline were greatest for the neutral clips, yet from examining the means this only holds for the maximum deviation from baseline. This fits in with the directional fractionation hypothesis (see section 3.2.2.2), as HR reduced more during the emotive clips. The significance values are not given in this paper, therefore it is difficult to determine how reliable the results are.

A surprising result that is not addressed by Dillon is that the mean SC is below the baseline for all conditions in the experiment. Watching emotional video clips would be expected to produce an increase in arousal, however given that SC did discriminate between

the content of the clips this observation could indicate that the baseline period used may have been unsuitable because it may have been stressful. HR decreased for all conditions, which fits in with the results of Simons et al. (1999) regarding HR deceleration during video. The lack of a significant effect of angle on subjective ratings surprised Dillon and was said to limit the interpretation of the results. This was hypothesised as being due to: measuring the effects of the angle one week apart; task demands; confounding variables or not enough participants. In conclusion, the research performed by Dillon et al., although applied in a different area to the research reported in this thesis and measuring a different state (arousal as opposed to perceptual strain), gives support to the use of physiological signals in that they can pick up effects that are not registered in subjective assessment (such as the effect of angle on SC).

#### 3.2.7.2 Investigating the use of physiological signals as a measure of presence

Meehan (2001) has also investigated the use of physiological signals as a measure of presence, but in a different application, which was a Virtual Environment (VE). The environment utilised was a Pit Room, with an unguarded hole in the floor leading to a room 20ft below. This extreme environment was chosen deliberately: if physiological signals could not pick up differences in such an environment then their use in less stressful environments may not be appropriate. Meehan posited that if physiological signals could measure presence, then they could be utilised in place of subjective assessment. However, as has been illustrated throughout this chapter, physiological signals should be measured in conjunction with subjective assessment because they do not always concur.

Meehan measured changes from baseline in three physiological signals: SC, HR and ST, the latter of which is known to decrease in response to heights and others stressors. The baseline used was in a virtual training room. In addition to the physiological signals

Meehan used subjective measures of presence and also behavioural measures (such as taking small steps). The task for participants was to move a book from the training room to the pit room.

Three experiments were performed which investigated: whether being exposed to the VE many times reduced its impact; if adding a wooden ledge increased presence; and the effect of four levels of video frame rate. The first experiment employed a within-subjects design: 10 participants were involved 3 times a day on 4 separate days (HR was not successfully measured). The results of this study showed order effects for SC and ST, both of which decreased after the first exposure only. Reported and observed behavioural presence also reduced: the former after the first session and the latter over the entire session. Order effects were also found in the frame rate study for HR and for SC and ST after the first exposure only, which highlights the importance of counterbalancing order in experiments utilising physiological measures. Order effects were not found for the passive haptics study, which could be because the addition of the ledge invoked strong responses and participants only did that trial on two occasions.

The second experiment involved more participants (52) and took measurements over 2 days, with and without a wooden ledge. The aim of the experiment was to discover whether adding a 1.5-inch wooden ledge increased the presence evoking power of the VE. Results showed that HR, SC and reported behavioural presence were significantly higher with the wooden ledge. Subjective presence was higher with the ledge but this difference was not significant.

The final study investigated the impact of four levels of video frame rate on presence: 10, 15, 20 and 30fps. 33 participants entered the VE four times on one day and were presented with the same VE at a different frame rate each time. Keeping the task consistent is positive, however participants may have become bored or frustrated

by having to do the same thing four times. It was hypothesised that the higher the frame rate, the greater the evoked presence would be. Results showed that the hypothesis was confirmed for 15, 20 and 30fps for HR and reported behavioural presence.

Meehan concluded that HR performed the best out of the physiological measures and that it correlates the most with subjective measures. However, the lack of significant results for ST may be because, as Meehan acknowledges, ST takes longer than 2 minutes to respond, whereas the conditions he used were around 1.5 minutes in length.

Meehan also attempted to determine if physiological measures could be used as a between-subjects measure. To do this, the first task for each experiment was analysed and showed that: physiological responses were significantly higher in the Pit Room compared to the training room; physiological signals did not correlate well with subjective measures, although this is not necessarily bad because it indicates that a more complete picture of the user experience is emerging; and physiological responses did differentiate among conditions in the passive haptics and frame rate (for HR only) studies. Consequently, Meehan concluded that HR showed the most promise as a between-subjects physiological measure of presence.

### *3.2.8 Evaluating collaborative entertainment technologies*

A study that utilised physiological signals (SC, HR, EMG, and Respiration) to measure enjoyment was performed by Mandryk & Inkpen (2004). It involved 10 participants who were experienced computer game players. They played the same game twice for five minutes, once against a friend and once against a computer. Rest periods of five minutes were incorporated before each condition to allow responses to return to the baseline. From examination of the means it appears that the baseline was not subtracted from the responses during the conditions, which is not common practice in

this area (e.g. section 3.2.7.1). Subjective measures were taken after each condition, which included 5-point Likert scales to measure challenge, ease, engagement, excitement, frustration and fun. It was hypothesised that subjective and physiological data would correlate with each other and that physiological signals would correlate with events during a game. Measures of task performance were also taken (whether the participants won, lost or drew the game).

Subjective responses showed that it was significantly more boring to play against the computer and significantly more engaging, exciting and fun to play against a friend. Half of the participants found it more challenging to play against the computer, which was attributed to the computer being a better player. A significantly higher level of SC was predicted when playing against a friend and this hypothesis was confirmed. SC also matched up with events during the game (e.g. when a goal was scored and during a fight). This matching up between events and physiology is stated by the authors as being one of the main advantages of measuring physiological signals. Higher EMG was predicted when playing against a friend due to trying harder as a result of greater competition: this was also confirmed. There were no significant differences in HR or respiration between the two conditions.

Correlating the physiological and subjective responses showed that normalised SC was correlated with fun and inversely correlated with frustration. However, the research of Scheirer et al. (2002) showed that SC increases during frustration, thus it may be that the simplistic way of measuring frustration in this experiment (a post-hoc Likert scale) may not have been in-depth enough to investigate this dimension and not warrant such a strong conclusion.

## 3.3 Conclusions

### *3.3.1 Substantive conclusions*

The studies reported in this chapter have taken place in diverse areas into which they have provided a number of substantive contributions. In the area of HCI, Ward et al. showed that HCI events can be categorised with regard to the amount of stress they produce and that the appearance of a pop-up box caused an increase in SC. Specifically related to the application of interest in this thesis was the study by Simons et al. which showed that SC significantly increased for moving and arousing images and HR significantly decreased for moving, arousing and negative valence clips. EMG was related to the valence of the images and subjective responses showed that moving images were rated as being more positive and arousing than still clips.

Wastell et al. investigated two computerised systems and found different results. The first showed that a computerised system for telephone operators led to a decrease in feelings of well-being, performance, activation and mental effort. This was attributed to factors, such as a reduction in skill variety. The second study showed a reduction in anxiety and increase in job satisfaction over 4 months due to the computerisation of ambulance control operations. Finally, Rowe et al. showed a clear effect of mental effort on HRV.

Dillon et al. and Meehan investigated the use of physiological signals as a measure of presence. Dillon et al. showed that: SC reduced more when a Vistral was on and when watching clips of a boat than when watching a rally car; HR discriminated between the rally and boat clips; stereo video and the boat clips were rated as being higher in presence; the rally car clip had more subjective negative effects, which were increased by the Vistral; and stereoscopic video was rated as being more positive than monoscopic video. In addition, Dillon showed that: SC increased significantly for amusing clips; a

smaller visual angle induced less presence; and HR significantly decreased for amusing and sad clips. Subjectively, the amusing and sad clips were rated as being more arousing, interesting, engaging, being of better image quality and adequacy and lower in negative effects than the neutral clips. The sadness category was also rated higher than the neutral category on a question regarding whether the television environment at times became more real or present than the real world.

Meehan showed that: 15, 20 and 30fps video increased presence as shown in HR and subjective measures; the addition of a wooden ledge significantly increased HR and SC and behavioural presence; and that subjective, physiological and objective responses to a VE reduced with multiple exposures.

The research of the Affective Computing group at MIT showed the successful performance of algorithms to detect stress and frustration. In addition, the MindGames group illustrated the potential of adaptive computer games to teach children with anxiety disorders how to reduce their stress levels. Finally, Mandryck and Inkpen found an increase in SC and EMG when playing computer games against a friend than a computer. Participants said it was more fun, engaging and exciting to play video games against a friend than a computer and SC was correlated with fun and negatively correlated with frustration.

The research reported in this chapter has used physiological signals to measure presence, workload and stress in HCI, to develop algorithms to detect stress, anxiety and frustration and to evaluate the impact of moving and still images and collaborative play. Generally, the research reported in this chapter has supported the use of physiological signals. It is only the research of Dillon et al. that has questioned the use of physiological signals as a measure of presence. This question was posed as a result of the first study

where the physiological signals did not respond in the same direction as the subjective results. Yet, dissociation between physiological and subjective results indicates that the physiological signals are uncovering information that the subjective responses are not, thus the lack of concurrence is a result in itself.

### *3.3.2 Methodological conclusions*

The studies reported in this chapter have utilised a variety of methods (see table 4). With the exception of three studies (Ward et al. (2001), Rowe et al. (1998) and Bersak et al. (2001)), all have measured more than one physiological signal and these have varied in number between 2 and 4. The most commonly measured signals were SC and HR and the number of participants used varied from 1 to 119.

There are three elements of the baseline recording session that differ in the studies reported in this chapter, the first of which is the length. The research reported in this thesis used 15 minutes, as was recommended by one of the leaders in the field at the time the research reported in this thesis commenced (Healey 2000). This time period was viewed as being long enough to allow the participants and consequently their physiological signals time to settle down and relax and, thus generate a true resting baseline. The second element is whether the baseline is an experiment baseline or a resting baseline. For example, Ward & Marsden (2003) used the first minute of the experiment as a baseline and Meehan (2001) gathered baselines in the training room. The research reported in this thesis used a resting baseline, where participants read a newspaper for 15 minutes, in order that changes from rest would be generated, as opposed to changes from a heightened state in an experiment due to the experiment beginning. The final element is the number of baselines used. For example, Dillon et al. (2001) used a different baseline for each condition, whereas the research reported in this thesis used one resting baseline in order to ensure that each comparison was the same.

In most of the studies in chapter 3, subjective measures were used, thus support is given to this element of the 3-factor approach. Task performance was measured in fewer studies. Within-subjects designs were employed in most studies, as is more common in psychophysiological studies and were also used in the research reported in this thesis. Finally, the most common feature of the signals used was the means, which offers support for the analysis conducted in the research reported in this thesis. Therefore, the evaluation approach utilised in the research reported in this thesis gains support from the studies reported in this chapter, whilst remaining novel in the area to which it was applied.

## Chapter Summary

In summary, this chapter began with an introduction to psychophysiology, the human nervous system and the signals measured in the research reported in this thesis. A background to the area of Physiological Computing through the critical examination of previous research in the area was then given. The research reported in this thesis is placed between Affective Computing and Wastell's workload research, as it uses information from Affective computing regarding gathering and analysing the physiological signals and uses the signals in a similar evaluation framework as Wastell's. The next chapter describes the methodology used in the experiments in the research reported in this thesis.

# Chapter 4 Methodology

## Chapter Aims

In this chapter, the rationale for the methodology used in the research reported in this thesis will be introduced. The chapter begins with a description of the 3-factor framework for assessing the impact of media quality degradations on users and previous work on which it is based. Each factor is then discussed in turn, ending with a description of the novel approach to measuring user cost in this thesis: the measurement of physiological signals.

## 4.1 Introduction

As was illustrated in Chapter 2, current methods of assessing MMC quality, mainly those recommended by the ITU, are unsuitable in this context and do not give an accurate indication of the impact that the quality has upon the user. This is especially pertinent when any of these assessment methods are used in isolation, since the results may mislead application designers and network providers about appropriate levels of media quality for long-term safe and enjoyable use. This means that users may not adopt the technology, or use it for a prolonged period of time. Thus, the developers of new applications or services may lose out financially. Alternatively, a user may continue to utilise an application, but it could have an adverse effect on them physiologically or could impair their task performance. So, how can an accurate indication of the impact that MMC quality has upon the user be gained?

## 4.2 Evaluation Methods used in the Thesis

There are currently no established methods in HCI for evaluating media quality in networking applications or services. This problem is amplified in the application domain studied in this thesis, Internet

MMC. Media quality in networked multimedia applications and services is difficult to evaluate for a number of reasons.

1. MMC typically involves two or more users and it requires communication between users to be considered, not just their interaction with the system.
2. Media quality often fluctuates within a session, thus it can be inappropriate to administer solely a post-hoc questionnaire.
3. The frequently poor and variable quality makes the methods used to evaluate high quality videoconferencing inappropriate.
4. The tasks performed utilise generally two or three different types of media (audio, video and a shared workspace), which all have an impact on each other.
5. Finally, because many participants have not experienced such quality levels they describe it in different ways, which makes subjective responses difficult to analyse.

To date, it is mainly subjective methods, such as asking users to rate the quality, that are used to assess the impact of media quality degradations in this area. However, these have drawbacks when used in isolation. Therefore, a 3-factor HCI evaluation framework of task performance, user satisfaction and user cost (as described in section 2.1.2) has been revisited. This thesis focuses on developing a measure of user cost, as this element has largely been neglected.

The assessment approach taken in this thesis is specified below.

- It was based on experimental trials carried out in a laboratory. This was done to minimise the number of variables that may have influenced physiological responses. If the impact of the degradations could be reliably measured in a controlled way, then the research could be moved into the field. Attempts were made to give the tasks ecological validity (see section 4.3.2.1).

- Objective and subjective quantitative data were collected through physiological signals and post-hoc rating scales and questionnaires.
- Data at a low level of granularity were given by the post-hoc questionnaires and rating scales, and data at a high level of granularity were given by the physiological signals, which were measured continuously during the experimental conditions.
- Immediate responses (not intrusive to the thought process or task) were garnered through physiological measures, and post-hoc methods were used for recollection of events.

The resources necessary to conduct the experiments were:

- physiological monitoring equipment
- participants, who were mostly students and staff at UCL and Glasgow University
- an experimenter skilled in measuring physiological responses
- PCs with MMC tools. These were connected to each other either across the network or on a standalone, isolated network when a more controlled environment was required.

## 4.3 The 3-factor framework

The research reported in this thesis used a 3-factor framework of task performance, user satisfaction and user cost, which was based on the criteria developed by Shackel (1981) (see section 2.1.2). Support for the framework comes from Wastell (1990): *"The study of HCI requires a greater depth than can be provided by any single domain...be it behavioural, subjective or physiological"* and combining these measures *"...provides a richer and firmer base for the study of HCI".*

As illustrated in chapter 2, measures of user satisfaction and task performance have been used when there has been a need to assess media quality. However, the cost to the user has largely been neglected. In a society where work stress is prevalent, having

applications that could contribute to this is unwise. Also if a user feels under strain from using a conferencing application where the quality is poor and annoying, it is unlikely that he/she will continue using it. Thus, the user should be satisfied, comfortable and able to complete their task with the quality delivered.

The weighting given to each of the three factors is not uniform and is largely dependent on the context in which the application is being used. For example, if the system is to be deployed for a safety-critical application, then task performance is likely to be the most important dimension because the task must be completed successfully. If the context is one of entertainment, then user satisfaction should be of prime importance, as it is vital that the user is happy with what they are interacting with. Finally, if the user has to perform an important job-related task daily, such as regular business meetings via MMC then the most important dimension will be the physical cost to the user.

### *4.3.1 User Satisfaction*

Current methods of assessing MMC quality are focused on user satisfaction, where users are asked to give their opinions on the quality they have received and the rating scales used most frequently are those recommended by the ITU. There are many drawbacks to these scales (section 2.4.1.2). In addition to problems with the scales, there is a more fundamental drawback to using subjective assessment: it is cognitively mediated. This means that in a situation where a lot of variables are operating, the rating a user gives to the quality can be biased. For example, it has been found that variables such as budget (Bouch & Sasse 1999) and task difficulty (Wilson & Descamps 1996) can influence a user's ratings of quality.

Subjective rating scales are most commonly administered at the end of a session. This means that the rating given will be influenced by what was experienced towards the end of the session (the recency effect). The alternative is continuous assessment throughout a

session using a tool such as QUASS (section 2.4.2.1). However, the activity of continuous rating can interfere with a user's main task. Moreover, it is argued by Knoche et al. (1999) that subjective assessment is fundamentally flawed, as it is not possible for users to register what they do not consciously perceive. Clearly there are drawbacks to using subjective assessment, particularly in isolation. However, this does not imply that their use should be completely abandoned, as they are extremely useful in garnering users' overall opinions. After all, if a user is not subjectively satisfied with the quality levels they receive, they will not adopt and continue to use the application: *"attitude criteria are no less valid than any other; indeed in many respects they are more valid with regard to usability, because ultimately it is the human user who must express the judgment of this characteristic"* (Shackel 1981). Thus, user satisfaction is measured as part of the 3-factor approach in the research reported in this thesis.

#### 4.3.1.1 User Satisfaction Measures used in this Thesis

In experiment 1, user satisfaction was measured by a post-hoc questionnaire (appendix A). In experiment 2, the 100-point rating scale (see appendix C for an example) was utilised. This is a 100-point scale of 100mm, which has numbered markers every 10mm, and is labeled with "very poor quality" at one end and "very good quality" at the other. In experiments 3, 4 and 5 the same rating scale as used in experiment 2 was used to rate the quality and adequacy of the audio and video channels separately. Despite them having an interactive influence upon each other (section 2.3.2) it is still important to attempt to assess their impact separately.

### *4.3.2 Task performance*

If a user cannot successfully complete their task with an application, then he/she will not utilise the application in the future to perform the same task. For example, if the task is one of distance learning, and the pupil cannot understand the tutor due to inadequate audio

quality, they may fail their assessment and opt for face-to-face lectures in the future.

Some classic measures of task performance (which all concern effectiveness and efficiency) are described below.

1. Task completion: can the user successfully complete their task?

2. Task length: how long did the user take to complete the task as opposed to the same task performed using another medium for example, face-to-face?

3. Number of errors: how many errors did the user make when performing the task, as opposed to performing the same task over a different medium?

The problem with these measures in the context of the tasks utilised in this thesis is that they are not well suited when a subjective opinion is required or a judgment needs to be made, such as in the context of an interview. They are more suited to simple tasks and tasks where a clear answer has to be reached. However, tasks such as used in this thesis are more ecologically valid as they are representative of typical tasks performed using a MMC link.

A study that found a discrepancy between subjective evaluation and measures of task performance was performed by Kies et al. (1996) who found no differences in task performance at video frame rates of 1, 6 and 30fps. However, low frame rates were subjectively rated as dissatisfying and caused distraction. As a result of this, some participants disregarded the video. The authors viewed this as having the potential to reduce learning over the long-term. This study illustrates that task performance measures do not give the full indication of the impact of media quality degradations on users, thus should not be used in isolation.

#### 4.3.2.1 Tasks used in this thesis

For the purpose of this thesis, a task is defined as that which forms the main activity of a session. Due to the multitude of variables that can impact upon physiological measures stemming from the environment and cognitive events, it was decided to use tasks that would attempt to minimise the number of variables operating. A passive task involves the participant watching/listening to some material, as opposed to actively doing something, such as talking to someone. Therefore, this was considered to be an appropriate task to begin with. In order to ensure that the participants did not become bored, they were given an overall task to perform. In experiment 1, the material to watch was a recorded interview between a university admissions tutor and a candidate for the Computing undergraduate degree course at UCL. The task was for the participant to decide if the candidate should be successful in their application. This task and the same context were used in experiment 3. In experiments 4 and 5 participants had to actively perform interviews. The remote interviewing scenario was deemed as sensible as it is one that fully exploits the capabilities of the application.

In order to ensure that the material of the interviews in experiments 1, 3 and 5 was ecologically valid, the author of this thesis conducted interviews with experienced admissions tutors. These interviews identified the goals of such interviews, their style, sub-tasks, the importance of body language and perceived expectations of candidates. From these interviews, a document was produced which detailed a typical interview scenario, overall high-level goals, sub-goals and tasks. To build on this body of knowledge, four anonymous admission forms were obtained with permission and were studied to form an opinion of what a typical application form looks like and to get an impression of the capabilities of a typical candidate.

After this, two potential computer science degree applicants were interviewed by the author of this thesis in order to identify how they

would respond to the questions previously identified. From all this information, interviews were scripted and acted in both experiments 1 and 3. This was done to ensure standard questions and responses over all the interviews, thus every participant would watch interviews with similar contents to allow responses to be compared across interviews and to ensure that the material was ecologically valid. In addition, both candidates had to be of similar calibre in terms of academic qualification and ability. If one were obviously worse than the other responses to the candidates, as opposed to the quality, may have been observed. For this reason, the task performance measure is largely subjective and difficult to measure: there is no right or wrong answer. Thus, this is not an element that will be examined extensively. Yet, it was still important to give the participants a task in order to focus their attention and increase the ecological validity of the experiment.

In experiment 2, the most basic task was employed, which was a passive listening task that required participants to rate the quality. This task is not claimed to be ecologically valid. It was utilised as the variable being tested, audio quality, is complex and subject to many degradations. Thus, the aim was to isolate its effects as far as possible from the video channel and from task effects, prior to including video and making the task more difficult. This task is very similar to that recommended by the ITU.

In experiment 4, the task changed slightly. The interview scenario held, as it was viewed to be an appropriate task, yet this time the scenario was not university admissions, but loan assessment[11] and the task was to actively perform an interview, as opposed to passively watching interviews. The interviewers were given a checklist of details that they had to discover from the applicants (who were actors) in the duration of the interview. This included aspects like the number of credit cards the applicant had and their income per month. After each interview, the participants had to fill out an applicant assessment form and they were asked if they thought that the applicant should be offered the loan.

In experiment 5, the interactive remote interview task was again used, yet for this experiment admissions tutors were the participants and they had to interview candidates for the Computing degree course at UCL (who were actors). The admissions tutors were not given scripted questions, however the candidates were given a list of typical questions that could be asked and responses they should give. Therefore, there was some control over the interview content. The admissions tutors were told that the interviews were mock but that the candidates were genuine applicants for the Computing degree course. The interviewers were told that the purpose of the interviews was to determine if MMC would be a suitable medium over which to perform such interviews in the future. In order to enhance the realism of this scenario, interviewers were given application forms that they were told the candidates had completed, however one of the co-experimenters had completed these, as the candidates were actors.

In experiments 1, 3 and 5 participants were given genuine candidate assessment forms used in the Computing department of UCL to complete (see appendix B). These require the interviewers to mark

---

[11] The loan scenario was chosen because it would focus participants' attention on the video screen in order to determine the suitability of the applicant for a loan. This task chosen by the ETNA project team.

on a scale of very high, high, departmental norm, low or very low, the candidate's:

1. general intelligence
2. motivation
3. creativity
4. social skills
5. enthusiasm
6. academic intelligence
7. self esteem
8. 'other' strengths.

The candidate assessment forms were administered during the baseline session and served the purpose of structuring the interview, increasing its ecological validity (as these forms are used in the Computing department at UCL) and giving the participant a task to focus on.

Experiments 6, 7 and 8 extended the scope of the method to assess its potential in areas other than MMC (see appendix D). These were joint experiments and the author of this thesis had minimal input in the design of the experiments. Experiment 6 investigated delay and pricing in a web-based library application, where the task for participants was to complete a series of web searches and download a number of files. Experiment 7 involved participants performing four tasks on four websites: to search for something, to read the latest news, to shop for a digital camera and to send an email. Finally, experiment 7 took place in a VE, where the task for participants was to explore the space for four minutes so that they could report on what they had experienced.

### *4.3.3 User Cost*

#### 4.3.3.1 Subjective measures

The physical cost to the user, for example strain and muscle tension, of interacting with a system particularly measured objectively, has been largely neglected in HCI evaluation and practice. This may be because application designers were more concerned about users' opinions of the application and whether they could perform their task with it. However, it is vitally important to consider the impact that applications have upon the user. With regard to MMC, if the quality is irritating to users, then they will not be comfortable using the application and may cease using it. Alternatively, if they continue to use it, this could induce negative health effects in the long term.

There are subjective measures of user cost. However, these are subject to cognitive mediation and recency effects may influence post-hoc ratings. In addition, they cannot discriminate between different segments of quality within the same session. It is due to the many problems with subjective assessment that the research reported in this thesis investigated the use of an objective method of measuring the impact of media quality degradations on users. One way of doing this is to measure physiological signals of perceptual strain. Such signals are not subject to cognitive mediation, are continuous throughout a session, therefore can offer the potential of being able to discriminate between different segments of quality within the same session and do not interfere with the user's main task.

#### 4.3.3.2 Perceptual Strain

The working hypothesis of this research is that when users are presented with insufficient audio and video quality, they have to expend extra effort at the perceptual level. If they struggle to decode the information, this should induce a response of perceptual strain, even if the user remains capable of performing his/her main task. Thus, the research reported in this thesis posits that under

perceptual strain, SC and HR will increase and BVP will decrease, as is typical of a SNS response (see section 3.1). These three signals were measured for the following reasons.

- They can be measured from three fingertips of one hand, thus are physically non-invasive, as opposed to measuring stress hormones through blood samples.
- They are widely used in the area of Physiological Computing (see chapter 3).
- SC has been found to be one of the most robust and non-invasive physiological measures of ANS system activity (Cacioppo & Tassinary 1990).

#### 4.3.3.3 Features of the signals measured

Before any experimentation began, it was imperative to measure the baselines of all 3 physiological signals. This was done for a period of 15 minutes in which participants were asked to relax and to read a newspaper. The baseline figure was always subtracted from the experimental response.

Changes in the mean levels of SC (ms), HR (bpm) and BVP (%) were investigated. This was done for the following reasons.

1. If significant changes in the means could be detected this method was more likely to be adopted in the HCI community, as opposed to more complicated method of analysis, such as extracting the number of SCRs.
2. The technique of using mean levels has been used in related research in the area of Physiological Computing (see section 3.2).
3. If overall SC and HR increased, whereas BVP decreased (all of which are indicative of a SNS response – see section 3.1) this would be an indication that the degradation had impacted upon the participant generally, as opposed to small, momentary fluctuations, which could be due to the task.

4. Looking at basic changes with passive tasks was viewed as a sensible starting point, as taking such measurements in this context is novel. It is accepted that as the tasks become more complex and the research moves into the field, then more complex feature detection may be necessary.

### *4.3.4 Equipment*

The equipment used throughout this research was the ProComp+, produced by Thought Technology[12]. It converts analog signals into a digital form and connects to a PC through a fiber-optic cable and adapter. The hardware is small: this was important, as participants may have been distracted or more nervous of having their signals measured by bulky equipment. The equipment is also portable, which was important because measurements were taken in different laboratories within UCL and also at Glasgow University. The cost for the specific setup used in the research reported in this thesis was £2105.

#### 4.3.4.1 Sensors

To measure SC, two sensors are placed on adjacent fingers (fingers 2 and 3, where finger 1 is the index finger). An imperceptible electrical current is passed between the two sensors and they measure the skin's capacity to conduct the current. HR and BVP were measured from the same sensor on the index finger: the photoplethysmograph. This sensor uses the light absorption characteristics of blood to measure the blood flow through capillary beds in the finger. These small capillaries contract when people experience stress, which cause the envelope to pinch inwards. The periodic component of this signal can also be used to measure HR in bpm. The sensors were configured to save at a rate of 20 samples per second. A picture of the sensors can be seen in figure 5 (when measurements are taken, the hand must be placed face down,

[12] http://www.thoughttechnology.com

however in this picture the hand is face up to show the sensors in more detail).

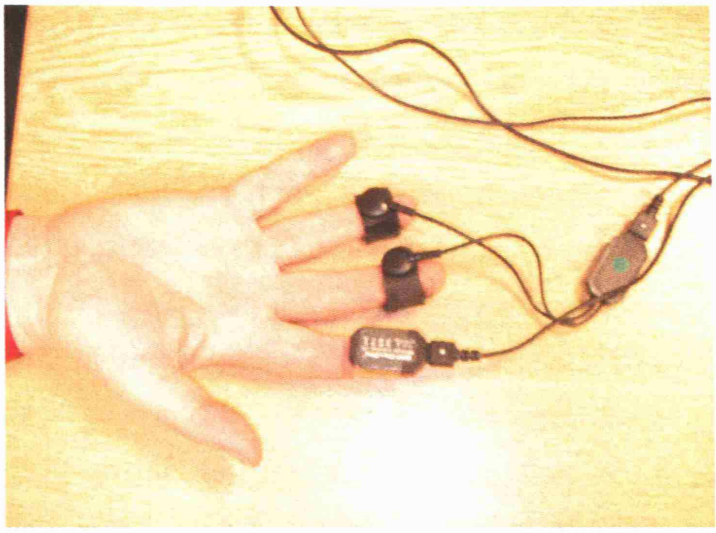


**Figure 5: Picture of physiological sensors**

In experiment 8, ElectroCardiogram (ECG) was measured because the photoplethysmograph sensor is extremely sensitive to movement. In experiments 1-7 this was not an issue, as each participant remained seated at a desk with his/her hand rested on the table to encourage him/her not to move it. However, in experiment 8 participants were required to physically walk around a VE, therefore using the photoplethysmograph sensor could have produced false readings. The measurement of ECG involved participants placing three adhesive sensors on their skin on the left side of their body. Two went on either side of the collarbone and one was placed on the bottom rib. From this signal, HR bpm can be measured.

### *4.3.5 Methodology for measuring physiological signals*

In experiments 2-5, SC, HR and BVP amplitude were measured. In experiment 1 BVP was incorrectly measured. Before the sensors were attached, the SC sensor was zeroed, which corrects any offsets. Each experiment began with the 15 minute baseline measuring session where participants were given a newspaper to read in order to get a resting baseline. The mean of the baseline was

then subtracted from the mean during the conditions. Participants were instructed to keep their hand still throughout the experiment.

Most of the SC data were not normally distributed. Performing a logarithmic transformation *"...produces a very marked fall in skewness and kurtosis"* (Venables & Christie 1980) in electrodermal signals, thus making the signal normally distributed and suitable for analysis with parametric statistics. Therefore in experiments 1–5, the log of the mean SC during the condition was calculated, from which the log of the mean SC baseline was subtracted. The log used was log to base 10 (x=log(x)). The HR and BVP data for all experiments was normally distributed (with the exception of 1 condition in experiment 3, which was excluded from analysis). Again, the mean of HR and BVP during the baseline was subtracted from the mean during conditions. Outliers identified in box plots in SPSS were removed.

All SC means and graphs in this thesis (with the exception of the raw data in appendix E) show the mean log on which a reverse transformation (x=power(10,x)) was performed because logging data can reverse the sign of data. The HR means and graphs shows the mean bpm minus the baseline and the BVP graphs show the mean % minus the baseline. The standard deviation graphs show the mean standard deviation.

SPSS version 11.5 was used to analyse the data. Ror experiments 1-5, repeated measures ANOVAs were performed, as all experiments were within-subjects designs. Sphericity is an assumption underlying the repeated measures ANOVA and refers to *"...the equality of variances of the differences between treatment levels"* (Field 2000). When sphericity is violated, a correction can be applied to produce a valid F-ratio. This was done in two cases in this thesis (SC experiment 1 for time and HR experiment 2 for degradation). In both of these cases, the Greenhouse-Geisser and

Huynh-Feldt estimates gave significant results, however to remain conservative the Greenhouse-Geisser corrections were used: *"if the two corrections give rise to the same conclusion it makes little difference which you choose to report (although if you accept the F-statistic as significant it is best to report the conservative Greenhouse-Geisser estimate)"* (Field 2000).

When there was a significant main effect of an independent variable with more than two levels, post-hoc pairwise comparison tests were performed. Variables with only two levels cannot be analysed further, as any significant effects can only reflect differences between the two levels (Field 2000), therefore the mean levels were reported. Pairwise comparisons compare all different combinations of the independent variable in SPSS. In doing these tests, a Bonferroni correction was used, which divides the probability value by the number of tests conducted and, thus ensures that the cumulative type 1 error (falsely rejecting the null hypothesis) is below 0.05. However, despite controlling the type 1 error rate well it is a conservative method (Field 2000). Where there were significant interactions, these were graphed. However, further analysis of significant interactions with variables with only two levels (of which all in this thesis were) is not necessary.

The data analysis reported in this thesis is more sophisticated than that conducted and reported in the papers published from the research reported in this thesis (see section 1.8), as they did not log SC, did not subtract the baseline, and did not remove outliers. In addition, the previous analysis did not take into account gender, repetition of conditions and order of presentation of conditions.

## Chapter Summary

This chapter introduced the evaluation approach utilised in the research reported in this thesis to measure the impact of media quality degradations on users, the 3-factor framework, which is

comprised of measures of task performance, user satisfaction and user cost. Task performance measures typically used in this area, such as measures of effectiveness and efficiency, are not suitable for most tasks performed using MMC, such as interviewing, which require a subjective opinion or judgement. Interviewing tasks, mainly University admission interviews, were used in 4 out of 5 of the MMC experiments reported in this thesis. Three of the tasks were passive, whereas two required the participants to actively perform an interview.

The ITU recommended subjective rating scales are typically used to assess quality in MMC. However, there are problems with these ratings scales in addition to a more fundamental problem with subjective assessment, which is that it is cognitively mediated. However, subjective assessment is an integral part of the 3-factor framework because if a user says the quality is not good enough, then it is likely that he/she will not use the application at such levels of quality. Subjective measures used in experiments 2-5 in this thesis were post-hoc rating scales, which were developed at UCL. In experiment 1, a basic questionnaire was used.

User cost has largely been neglected in HCI research, yet it is vital to gain an accurate impression of the impact of media quality degradations on users. Therefore, this dimension was focussed on in the research reported in this thesis. Due to drawbacks with subjective assessment, an objective method was used to measure user cost: physiological measures of perceptual strain (SC, HR and BVP). This method does not interfere with the user's task and provides a continuous stream of data throughout a condition as opposed to subjective data, which is mainly gathered at the end of a condition and consequently can be affected by primacy and recency effects.

The next two chapters report on empirical investigations that were conducted. Therefore, at this point it is timely to reiterate the fundamental questions that this research is tackling.

1. Can physiological responses to media quality degradations be detected?

2. Which media quality degradations have a negative impact on users, physiologically and subjectively?

3. How do physiological responses relate to subjective data?

# Chapter 5 The Impact of Audio and Video Degradations in Passive Tasks

## Chapter Aims

This chapter describes three experiments[13] that utilised passive tasks to examine the impact of low and high levels of video frame rate and a number of audio degradations with and without the video channel. Passive tasks were used as a starting point in this research because it was important to minimise the number of variables in operation to be able to conclude that any effects observed were due to the quality.

## 5.1 Experiment 1: Investigating the Impact of Low and High Video Frame Rates in a Recorded Interview Task

### *5.1.1 Introduction*

In experiment 1, the impact of low (5fps) and high (25fps) video frame rates were investigated[14]. The difference between the frame rates was extreme. This was done to test the sensitivity of the physiological signals and to determine if participants would subjectively notice the difference between the frame rates and, thus support the findings of Anderson et al. (2000) (see section 2.3.2.2).

[13] The raw physiological data for experiments 1-5 can be seen in appendix E. The raw subjective data for experiments 1 and 3 can also be seen in appendix E.

[14] The high frame rate was the same as that used by Anderson et al. (2000) and the lower frame rate was set at 5fps, as at the time this thesis research was conducted, rates as low as 5fps were common.

### *5.1.2 Design*

This experiment was performed in the Computer Science department of UCL. Twenty-four participants were randomly assigned to two groups. Both groups watched two recorded interviews between a candidate for the Computer Science degree course at UCL (candidate 1 was seen in interview 1 and candidate 2 was seen in interview 2) and an admissions tutor. The first interview that both groups saw started at 16fps and this frame rate lasted for 5 minutes. 16fps the rate at which lip synchronisation occurs (see section 2.3.2.1). This was done to allow participants to settle into their task and to give them a reference condition.

The interviews that group 1 saw followed the frame rate sequence of 5fps for five minutes, 25fps for five minutes, then back to 5fps for five minutes. The interviews that group 2 saw followed the frame rate sequence of 25fps for five minutes, 5fps for five minutes, then back to 25fps for five minutes. The frame rate changed within the interviews to allow a comparison between physiological responses to the two levels of frame rate within the same interview. The participants' task was to decide which of the two candidates should be offered a place at UCL.

### *5.1.3 Materials*

#### 5.1.3.1 Experimental materials

Participants watched the interviews on a PC and wore headsets through which they heard the audio. The four interviews were recorded over a standalone network. This measure was taken to ensure that there was no additional traffic that could affect the frame rates being sent and received. The audio quality of the interviews was good and did not vary. The tools used to record the audio and video were RAT and vic (see sections 2.2.1.1 and 2.2.1.2). The participants watched the image of the candidate on their machine to allow them focus solely on the candidate's video and the changes in it.

#### 5.1.3.2 Subjective assessment materials

After each interview, a questionnaire with 8 questions was administered (see appendix A), which was designed to gather opinions on the task, MMC in general, video quality, the sensors and to get an indication of the state of participants, for example how comfortable they felt. The questionnaire administered after interview 2 had three additional questions, regarding the overall task, the physiological sensors and the choice of candidate to be offered a place (see questions 9, 10 and 11 in appendix A). The standard applicant assessment forms used in the department of Computer Science at UCL (see appendix B) for such interviews were administered for completion after each interview. The evaluation measures used in this experiment can be seen in table 5.

| Element in 3-factor framework | Measure |
|---|---|
| Task Performance | Choice of candidate to get place |
| User Satisfaction | Post-hoc questions on quality |
| User Cost | • Physiological measures<br>• Post-hoc questions on stress/excitement/ comfort |

**Table 5: The 3-factor approach used in experiment 1**

### *5.1.4 Participants*

There were 24 participants, who were volunteers. Nine out of the 24 were female and 15 were male. In group 1 there were 7 males and 5 females, whereas in group 2 there were 8 males and 4 females.

### *5.1.5 Procedure*

Participants were told that they would be watching two recorded interviews and that their physiological responses would be measured. They were informed that they were free to withdraw from the experiment at any time. The physiological sensors were attached

to three fingers of each participant's non-dominant hand, thus leaving one hand free to complete the post-hoc questionnaires. An explanation of the signals being measured was offered. Baseline responses were gathered for fifteen minutes prior to the experiment beginning. In this time period participants were asked to read over the candidate assessment form in order to give them an idea of the qualities they should be looking for in the candidates. Once they had done this, they were offered a newspaper to read.

Once the baseline session was finished, the experimenter introduced the first candidate and played the interview. Once the interview was finished, the questionnaire and candidate assessment form were administered. The second candidate was then introduced, and the interview shown. Once the interview was over, the questionnaire and the candidate assessment form were again administered.

### *5.1.6 Hypotheses*

1. At 5fps, SC and HR will be higher than at 25fps, thus indicating perceptual strain. BVP was incorrectly measured.
2. There will not be a significant difference in responses between the groups at 16fps.
3. Subjectively, the difference between the two frame rates will be noticed, as it is more extreme than used by Anderson et al. (2000).

### *5.1.7 Results*[15]

#### 5.1.7.1 Physiological results

An independent groups t-test on the SC and HR 16fps data for group 1 (5-25-5fps) and group 2 (25-5-25fps) was performed to determine if the data from both groups were from the same population. Following this, a repeated measures ANOVA (one for SC and one for HR) with

---

[15] In Wilson & Sasse (2000a) the mean of each signal was calculated over both interviews and groups; thus, giving each participant a mean for 5fps and 25fps. This resulted in 5fps being significantly higher than 25fps in all signals.

the within-subjects variables time (3 levels) and interview (2 levels) and the between-subjects variables gender (2 levels) and group (2 levels) was performed to see if there were any significant differences in the data.

SC

A logarithmic transformation was performed on the SC data and the data were cleaned to remove outliers[16]. An independent groups t-test showed that there was not a significant difference between the two groups at 16fps (F=2.711, p=0.381).

The main effect of time was significant ($F_{(1.224,22.034)}$=3.954, p=0.05)[17]. Post-hoc pairwise comparisons showed that the difference between time 2 and 3 was significant (p = 0.031). Examination of the means shows that SC was higher at time 3 than time 2.

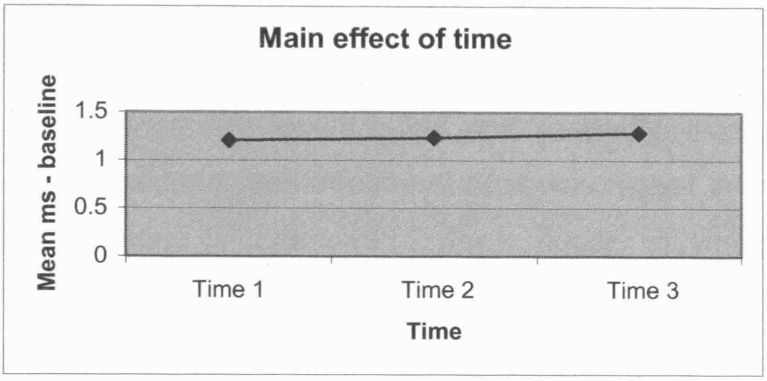



**Figure 6 Mean SC over time in experiment 1**

Figure 6 shows that SC at time 3 was significantly higher than at time 2.

There was also a significant interaction between interview and group ($F_{(1,18)}$ = 5.228, p=0.035).

---

[16] Two participants (7 and 10) were removed, which left 10 participants in group 1 and 12 in group 2. There were 14 males and 8 females.

[17] Mauchly's test of sphericity was significant, therefore the Greenhouse-Geisser correction was used.

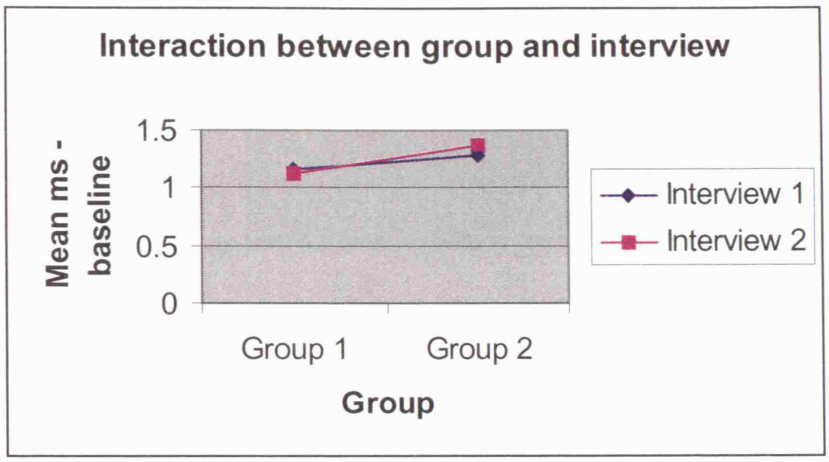



**Figure 7: Interaction between group and interview for SC in experiment 1**

Figure 7 shows that group 1's (5-25-5fps) SC is lower than group 2's (25-5-25fps) for both interviews. For group 1, interview 1 is higher than 2, whereas the opposite occurred for group 2.

In order to determine if there were significant differences between the frame rates in each interview, paired samples t-tests were performed. In group 1 interview 1 (see figure 8) there was a significant difference between the first presentation of 5fps and 25fps (p=.000), where 25fps was higher than 5fps. There was also a significant difference between the first and second presentations of 5fps, where the second presentation of 5fps was higher than the first (p=0.016). In group 1 interview 2 (see figure 9) there was a significant difference between the first and second presentations of 5fps (p=0.033), where the second presentation was higher than the first. There was also a significant difference between 25fps and the second presentation of 5fps, where 5fps was higher than 25fps. There were no significant differences between the frame rates in the SC responses of group 2.

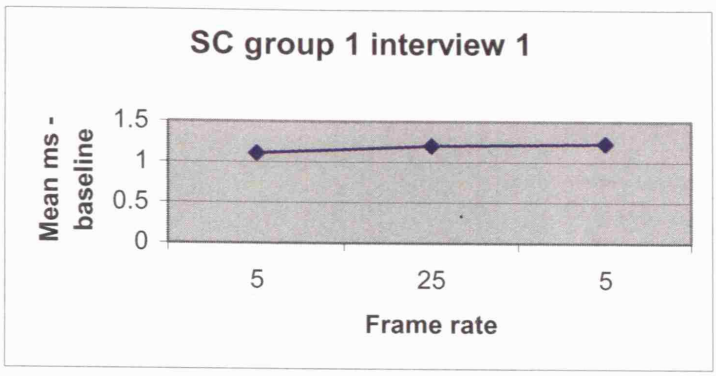



**Figure 8: Mean SC in group 1 interview 1**

Figure 8 shows that 25fps was significantly higher than the first presentation of 5fps and that the second presentation of 5fps was significantly higher than the first.

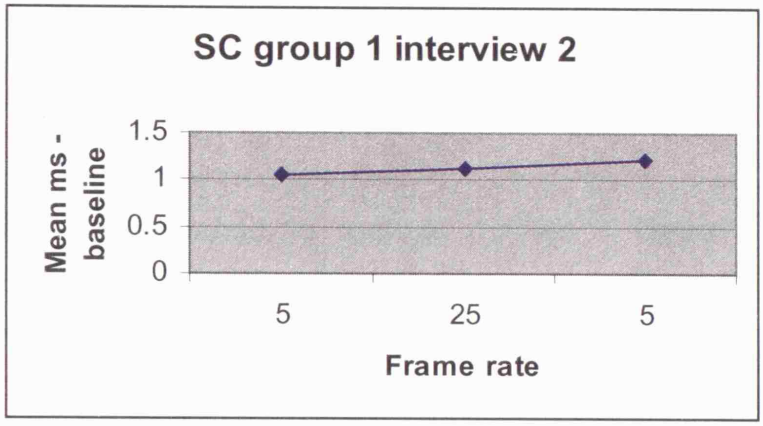



**Figure 9: Mean SC in group 1 interview 2**

Figure 9 shows that the second presentation of 5fps was significantly higher than 25fps and the first presentation of 5fps.

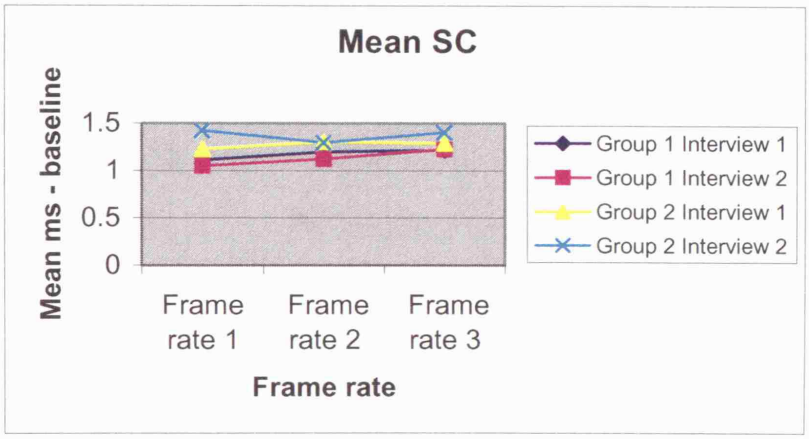



**Figure 10: Mean SC in experiment 1**

Figure 10 shows the mean SC of both groups for both interviews in experiment 1.

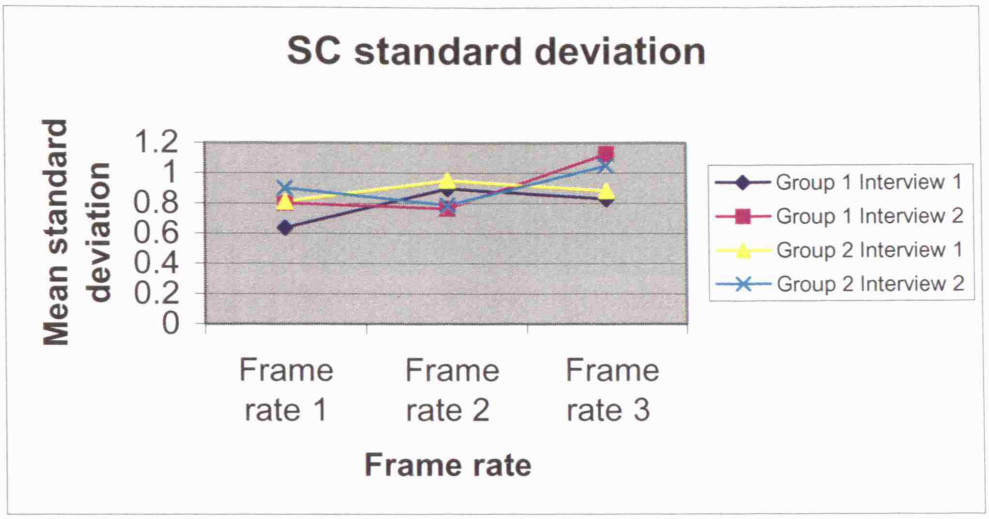



**Figure 11: Mean SC standard deviations in experiment 1**

Figure 11 shows the mean standard deviation of the SC signal of both groups for both interviews in experiment 1.

HR

The HR data was cleaned to remove outliers[18]. There was no significant difference between the two groups at 16fps (F=5.613, p=0.967). There was a significant main effect of interview ($F_{(1,16)}$ = 7.402, p=0.015). Examination of the means shows that HR was higher during interview 1. There were no significant interactions.

[18] Four participants (6, 9, 15 & 16) were removed, which left 10 participants in both groups and 13 males and 7 females.

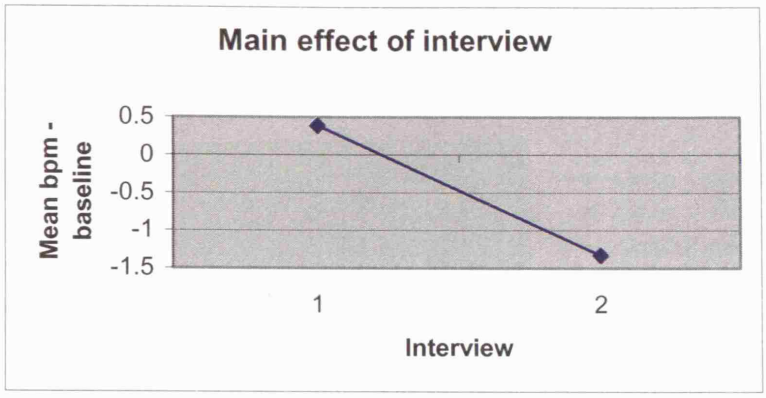



**Figure 12: Mean HR during interview 1 and interview 2 in experiment 1**

Figure 12 shows that HR was significantly higher in interview 1.

Paired sample t-tests for each interview showed no significant differences between the frame rates in either group.

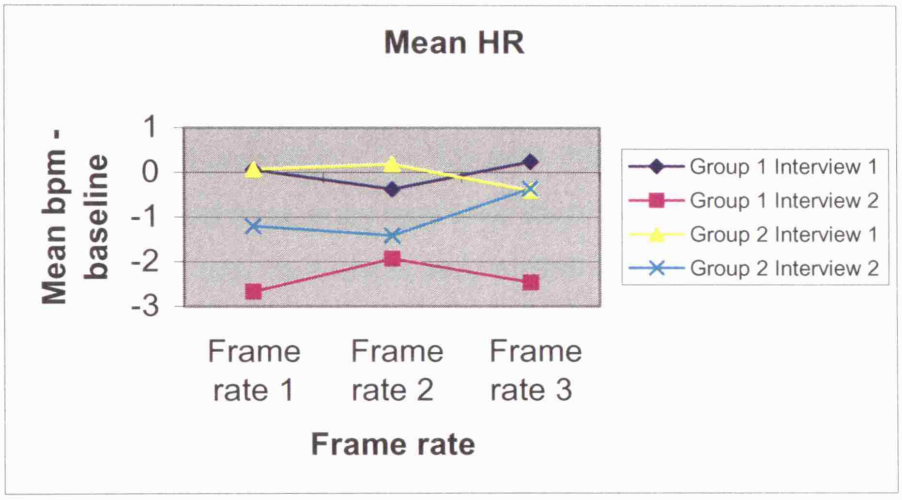



**Figure 13: Mean HR in experiment 1**

Figure 13 shows the mean HR for both groups and both interviews in experiment 1.

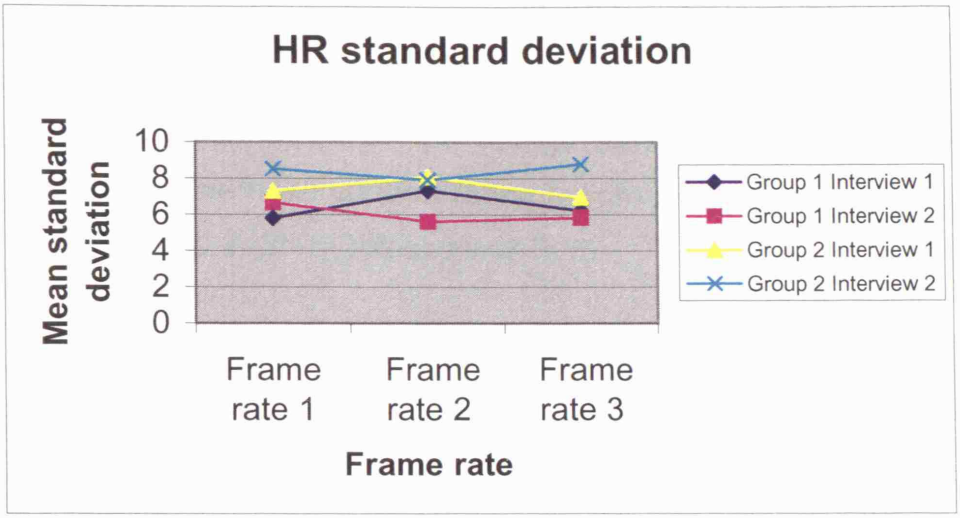



**Figure 14: Mean HR standard deviations in experiment 1**

Figure 14 shows the mean standard deviation of the HR signal of both groups for both interviews in experiment 1.

5.1.7.2 Subjective Results

The same 8 questions were asked after each interview (see appendix A). After both interviews had been seen, 3 further questions were asked. Chi-square tests were performed on all questions except question 3 (as this was numeric data) and question 5 (as the expected frequency was less than 5).

Question 1: Did you feel under any stress due to the quality at any point during this interview?

There were no significant differences in responses to question 1 in the 4 interviews. Overall, 13 participants in group 1 (5-25-5fps) and 14 in group 2 (25-5-25fps) said they were under stress.

Question 2: Did you feel excited at any point during this interview?

There were no significant differences in responses to question 2 in the 4 interviews. Overall, 8 participants in group 1 (5-25-5fps) and 16 participants in group 2 (25-5-25fps) said they were excited.

Question 3: On a scale of 1 to 5, how comfortable did you feel with the quality, with 1 being very comfortable and 5 being very uncomfortable?

For both groups, there were no significant differences in the comfort ratings over interview 1 and 2 (Wilcoxon). There were also no significant differences in the ratings between the groups (Mann-Whitney). The mean comfort rating for group 1 (5-25-5fps) was 3.31 and for group 2 (25-5-25fps) it was 3.15.

Question 4: Did you feel you could judge the character of the candidate well using this medium of communication?
Significantly more participants said they could judge the character of the candidate in group 1 (5-25-5fps) interview 1 (Chi-square with 1 degree of freedom = 8.33, p=0.004) and group 2 (25-5-25fps) interview 2 (Chi-square with 1 degree of freedom = 5.33, p=0.021).

Question 5: Did the quality of the video change at any point during the interview? If so, when?
The responses to this question, which asked participants if they noticed the frame rate change during the interview (see figures 12 and 13) were broken down into categories as an attempt to identify the participants that had noticed the frame rate change and those who were commenting on other aspects of the video.

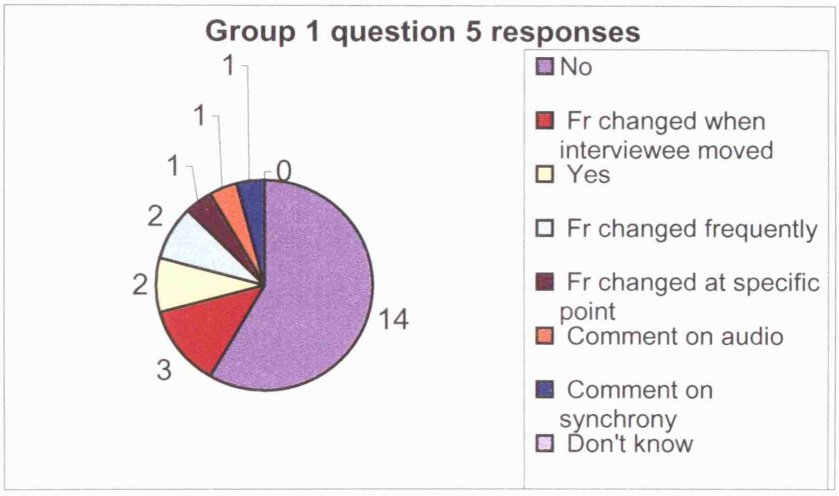



**Figure 15: Breakdown of group 1's responses to question 5 over both interviews**

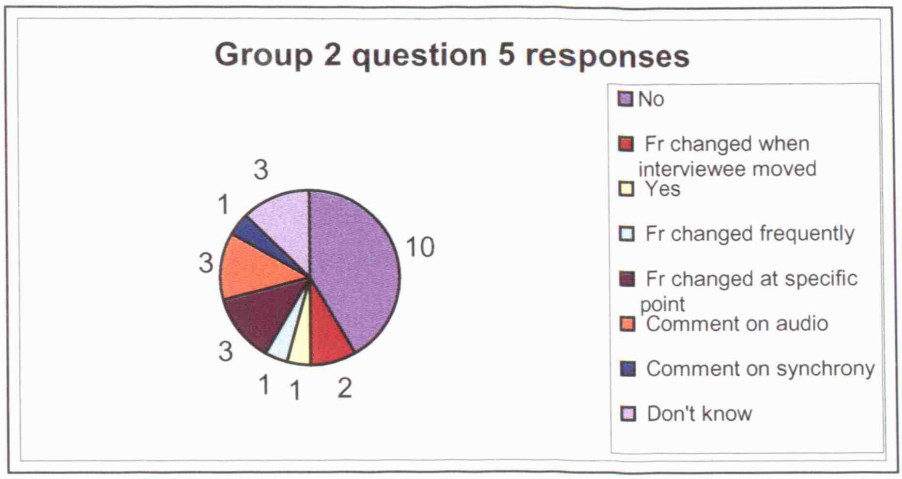



**Figure 16: Breakdown of group 2's responses to question 5 over both interviews**

The following two responses are classified as participants who noticed the frame rate change: 'frame rate changed at specific point' and 'yes'. It is possible that some participants who responded 'yes' may have been commenting on aspects of the video other than the frame rate change, however there is no way of teasing this out. Therefore, in group 1 over both interviews 12% (3/24) of participants noticed the frame rate change. In group 2 over both interviews 17% (4/24) of participants noticed the frame rate change.

Question 6: Overall, do you feel the quality of the video was good enough to support the interview?

Due to a misprint on the questionnaires, no useful data could be extracted.

Question 7: Do you prefer the video to be there as opposed to audio only?

There were no significant differences in the data for group 1 (5-25-5fps). In group 2 (25-5-25fps) significantly more participants said they preferred the video to be there in interview 1 (Chi-square with 1 degree of freedom = 8.33, p=0.004) and in interview 2 all participants said they preferred the video to be there.

Question 8: Did the video distract the audio at any point?

Significantly more participants in both groups said the video did not distract the audio in the first interview (group 1: Chi-square with 1 degree of freedom = 5.33, p=0.021, group 2: Chi-square with 1 degree of freedom = 8.33, p=0.004). There were no significant differences in either group to the second interviews.

Question 9: Did you feel the task put you under any pressure?

Significantly more participants in group 1 (5-25-5fps) (Chi-square with 1 degree of freedom = 5.33, p=0.021) said the task did not put them under pressure. The difference in group 2 (25-5-25fps) was not significant. Responses to this question can be seen in figure 17.

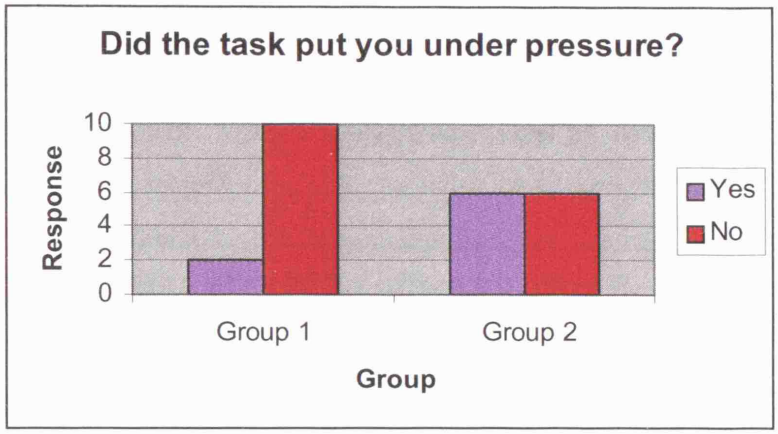



**Figure 17: Responses to question 9 in experiment 1**

Figure 17 shows that significantly more participants in group 1 said that the task did not put them under pressure.

Q10: Did having the sensors or your hands put you under any pressure?

The differences in both groups were not significant, as can be seen in figure 18.

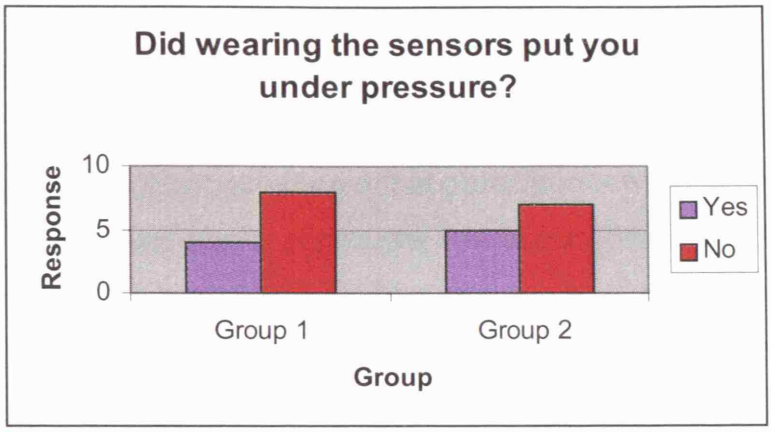



**Figure 18: Responses to question 10 in experiment 1**

Q11: Which candidate should be offered the place?

Significantly more participants in group 1 (5-25-5fps) said candidate 1 should be offered a place (Chi-square with 1 degree of freedom = 8.33, p=0.004). There was not a significant difference in the responses from group 2 (25-5-25fps). These results can be seen in figure 19.


**Figure 19: Responses to question 11 in experiment 1**

Figure 19 shows that significantly more participants in group 1 said that candidate 1 should be offered the place.

### *5.1.8 Discussion of results*

#### 5.1.8.1 Physiological results

There were no significant differences between groups 1 and 2 at 16fps in SC or HR, which shows that when the interviews started the groups were not significantly different. In SC, the third time period

was significantly higher than the second time period. This illustrates that the final 5 minutes of both interviews were more straining than the middle 5 minutes. This could be due to an accumulative effect of the degradations or could indicate that participants were frustrated by the task. There was also a significant interaction between group and interview in SC: group 2 (25-5-25fps) were under more perceptual strain during both interviews than group 1 and their SC increased from interview 1 to 2, whereas group 1's (5-25-5fps) SC decreased. This was not expected because group 2 had better overall quality than group 1. This result may indicate that the better quality caused participants to attend more to the task. Alternatively, the 5fps sections in group 2's interviews may have caused them more strain than group 1, as they began each interview with good quality

The only significant differences between the frame rates were in the SC of group 1. In interview 1, 25fps was significantly higher than the first presentation of 5fps, which may be due the negative effects of the 5fps persisting. The second presentation of the 5fps is not significantly higher than 25fps, which could indicate that experiencing the better quality has reduced the intensity of the response. However, it is significantly higher than the first presentation of 5fps, which may be due to a heightened level of strain after the initial increase between 5fps and 25fps. In the second interview there was no significant difference between the first presentation of 5fps and 25fps. This may be because participants have become used to the task and the experiment. However, the second presentation of 5fps was significantly higher than the first presentation of 5fps and 25fps. This may be due to an accumulation of the negative effects of the degraded quality or participants may be been frustrated or bored by the experiment or excited that it was about to end.

In HR, interview 2 was significantly less straining than interview 1. This shows that participants were more relaxed during the second

interview than the first, which is most likely due to them knowing what to expect and being more comfortable with the task.

The mean SC for both interviews and groups was above baseline, indicating that participants were less relaxed during the experiment than the baseline session, which was expected. The mean HR for the first interview for both groups was above baseline at two out of three time points, which again indicates that participants were less relaxed during the experiment than during the baseline session. However, during the second interview for both groups the mean HR was below baseline, which indicates a more relaxed state. This result shows a directional fractionation between SC and HR during the second interviews (see section 3.2.2.2). This can be interpreted as the task for the first interview being primarily cognitive, therefore HR increased to reject stimuli that would interfere with the performance of the task, whereas in the second interview the task that was now more familiar was perceptual, thus HR decreased to intake environmental stimuli. The results from the second interviews concur with the finding of Simons et al. (1999) that moving images cause HR to decelerate and that SC may be measuring arousal.

#### 5.1.8.2 Subjective results

Significantly more participants in group 1 (5-25-5fps) interview 1 and group 2 interview 2 (25-5-25fps) said they could judge the character of the candidate well using MMC. This indicates that the group with the poorer quality overall loses the ability to judge the candidate by the second interview, whereas the group with the better quality overall gains the ability by the second interview.

Significantly more participants said the video did not distract the audio in the first interview in both groups. This shows that participants may become less tolerant over time or that they were not paying as much attention to the task.

Significantly more participants in group 1 (5-25-5fps) said the task did not put them under pressure. It is surprising that the group with the poorer overall quality felt less pressure than group 2 (25-5-25fps). This may be because they were less engaged in the task as a result of the poor quality.

With regard to the assessment of the candidates, significantly more participants in group 1 (5-25-5fps) said that candidate 1 should be offered a place. The candidates were of a similar calibre, therefore this result was not expected. Candidate 1 was seen first, thus it may be that the negative effects of the poor quality accumulate so that by the second interview, participants are less tolerant and this is then attributed to the candidate, as opposed to the quality. Alternatively, participants may have stopped paying attention to the task by the second interview.

#### 5.1.8.3 Combining physiological and subjective responses

Because of the questionnaire not involving rating scales and the frame rate changing within the conditions, it is not possible to match the physiological responses with subjective results, with the exception of the question that asked whether participants noticed the frame rate change. Results showed that 12% of participants in group 1 and 17% in group 2 said they noticed that the frame rate had changed. This is not as many as were expected and indicates that participants may have been engaged in the task, therefore did not notice changes in the quality. This shows a discrepancy between the physiological and subjective results, as significant differences were found between the three frame rates in the SC of group 1.

A pattern between the groups emerged in both subjective and physiological measures. Group 1 said that in the first interview they could judge the character of the candidate better and that the candidate interviewed should be offered a place, whereas group 2 said that they could judge the second candidate better and that they should be offered the place (the last result was not significant). The

significant interaction between group and interview in SC shows that for group 1, SC was higher during interview 1 whereas for group 2 SC was higher for interview 2. Therefore, it appears that an increase in SC occurred when watching the favoured interviewee. This result may be reflecting engagement in the task as mentioned in section 5.1.7.1. Considering that both candidates were of a similar standard, this result indicates that the impression of the candidate may have been affected by the quality in the multimedia conference.

### *5.1.9 Addressing the hypotheses*

1. At 5fps, SC and HR will be higher than at 25fps, thus indicating perceptual strain. This hypothesis was only supported in the SC of group 1 interview 2, where the last presentation of 5fps was significantly higher than 25fps.
2. There will not be a significant difference in responses between the groups at 16fps. This hypothesis was supported.
3. Subjectively, the difference between the two frame rates will be noticed, as it is more extreme than used by Anderson et al. (2000), who found that the difference between 12 and 25fps was not noticed during an engaging task. This hypothesis was not supported. Combining the results over both groups and interviews, only under 15% of participants noticed the frame rate had changed. This result illustrates that in an engaging task, subjective measures should not be relied on in isolation because they miss more subtle changes, which were picked up by the SC signal.

### *5.1.10 Limitations of the experiment*

This experiment had limitations in its design, which may have impacted upon results. There was no complete control group (one group who completed the task at 16fps). Each group started off at this frame rate for five minutes, however having a control group would have allowed the impact of the high and low levels of frame rate to be more accurately determined. In addition, the conditions in each interview were unbalanced, as there were two five minute

sections of one level of frame rate and one five minute section of the other. This meant that in the ANOVAs it was not possible to directly compare one frame rate with another. It was a long experiment (in total around one hour) and participants may have become bored or frustrated by this. In addition, there was only a very small break between interviews, therefore responses to the second interview may have been influenced by responses to the first.

Finally, the post-hoc questionnaire used was basic. It was comprised of yes/no questions rather than rating scales, which give more information about the impression of the quality. The reason for this was that the scales used later in the thesis had not been finalised at this point.

### *5.1.11 Conclusions*

The lower frame rate (5fps) was expected to produce an increase in perceptual strain in comparison to the higher frame rate (25fps). This only occurred in the SC of group 1 interview 2, where the second presentation of 5fps was significantly higher than 25fps. Significant differences between the frame rates were only seen in group 1's SC. This may be because experiencing the higher frame rate first may have reduced the impact of the lower frame rate. In the SC signal across both groups, the third time period had significantly higher levels of SC than the second, which may either be due to boredom with the task or an accumulative effect of the lower frame rate.

The group with the better quality overall had higher levels of SC, which was not expected. However, the hypothesis was put forward that this may be reflecting an increased engagement in the task, which is supported by the positive reports of the second candidate.

The physiological results of this experiment were directionally fractionated (see section 3.2.2.2) during the second interview, which illustrates that SC and HR may be tapping into different states,

therefore providing support for the research reported in this thesis measuring more than one physiological signal.

The results from this experiment show most fundamentally that physiological responses to media quality degradations can be detected. This provides support for the use of the methodology in the research reported in this thesis. Furthermore, the directional fractionation and additional information gleaned by the subjective results show that it is important that more than one physiological signal is measured and that more than one factor of the 3-factor framework should be measured. Having investigated video frame rate, the next experiment investigated degradations in the audio channel.

## 5.2 Experiment 2: Investigating the impact of audio degradations in a passive listening task

### *5.2.1 Introduction*

It is well established that good audio quality is important in MMC (e.g. Sasse et al. 1994) and much effort has been expended to protect audio from network degradations. The networking research community has assumed that increasing the amount of bandwidth, thus reducing the amount of packet loss would ensure sufficient audio quality. Yet, in a large-scale field trial where sufficient bandwidth was available[19], users still reported problems in one out of three MMC sessions (Watson & Sasse 2000).

The three most commonly reported problems in this field trial were: missing words or incomplete sentences; variations in volume between participants; and variations in quality between participants. The first problem was most likely caused by packet loss, silence suppression or machine 'glitching'. The second problem was likely to be due to insufficient volume settings or a poor quality headset and the final problem was most probably caused by a lot of background noise, an open microphone or again, poor headset quality. The reception reports showed that generally audio packet loss was low (5% or lower). Short bursts of 20% or higher did occur, but this only happened occasionally. Given the low levels of packet loss, the reported problem of missing words cannot be explained solely by this. Therefore, an experiment was conducted to determine the subjective and physiological impact of a number of audio degradations caused by the network, end-user behaviour and hardware set-up.

### 5.2.2 Design

This experiment was performed in the Computer Science department at UCL. The author of the research reported in this thesis ran the physiological data collection part of the study and analysed the results. The co-experimenter, Anna Watson, planned and designed the experiment and administered and analysed the subjective data (Watson & Sasse 2000).

Twenty-four participants took part in the study. They all heard a volume test file of good quality first, which lasted for one minute. They then heard six two-minute audio files, of which the order was randomised for each participant. The volume test file was then played again and the same re-randomised six audio files were played. This was done to determine the consistency of the subjective ratings. See figure 20 for a photograph of the set-up for experiment 2.

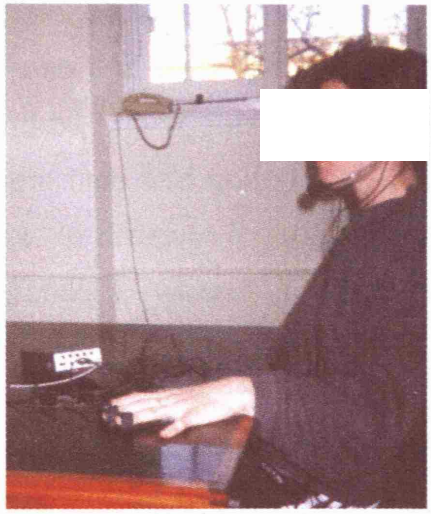


**Figure 20: Photograph showing the set-up for experiment 2.**

The task was one of passive listening, which is less stressful than being actively involved in a real-time task. This was done to eliminate the possibility of the responses to the task drowning out responses to the quality.

---

[19] The PIPVIC-2 (Piloting IP-based VideoConferencing) project involved 13 UK institutions and 150 participants in a range of educational activities from December 1998 to

### *5.2.3 Materials*

#### 5.2.3.1 Experimental material

The audio material used was a dialogue between two male speakers, which had been taken from previous project meetings conducted via MMC. The original material from the meeting could not be used as it was degraded, which meant there was no control over the quality, therefore it was recorded at good quality. Two male actors, who did not have regional accents, recorded the script via Sun Ultra workstations on the same local network in different rooms.

The material was recorded in RAT at 16-bit linear quality. In order to get the best quality possible, silence suppression was on, both microphones were open during the recording and both actors wore Canford DMH12OU headsets.

Once the whole dialogue had been recorded, the actors read parts of the dialogue again so that the experimenter could create the volume variation, echo and bad microphone conditions. The recordings were split into two-minute files and coded into DVI at 8kHz sampling rate and 40ms packets. In order to induce packet loss and repair (done with packet repetition) where it was required, the software programme test-repair was used. This is a component verification program that is included in RAT version 4. The codec and packet size used were the default settings of RAT version 3 that existed throughout the field trial. Seven conditions were generated for use in the experiment, which can be seen in table 6.

| Condition | Description |
|---|---|
| Reference | Degradation free |
| 5% packet loss (5% pl) | 5% packet loss on both voices repaired with packet repetition |
| 20% packet loss (20% pl) | 20% packet loss on both voices repaired with packet repetition |
| Echo | One speaker used an open microphone and speaker, which meant that the other speaker generated feedback |
| Quiet | One speaker recorded at a low volume, the other at a normal volume |
| Loud | One speaker recorded at a high volume, the other at a normal volume |
| Bad microphone (Bad mike) | One speaker used a bad quality microphone (Altia 087F) |

**Table 6: Description of conditions in experiment 2**

Due to the findings of the PIPVIC-2 field trial mentioned previously, 5% packet loss was chosen as the lower level of packet loss. The 20% packet loss condition was chosen as it is known that this is the level at which there is a sharp drop in the quality of repaired speech (Watson & Sasse (1997), Watson & Sasse (1998)).

Echo occurs when a user employs a speaker and an open microphone to communicate and forgets to mute their microphone when they are not speaking. It can also occur when using a 'leaky' headset, which is where the headphones leak sound into the microphone.

Many participants in the PIPVIC-2 trials complained of extreme volume differences between speakers. The listener can correct this by adjusting the volume of the incoming stream. However, this can result in the volume of the next person speaking being unsatisfactory and users can find themselves constantly altering the volume. The experimenters acknowledge that 'too loud' or 'too quiet' audio is subjective, however by piloting the material with Internet audio novices and experts, a more informed decision was made about representative levels.

The bad microphone condition was important to recreate, as during the PIPVIC-2 field trials, users complained about microphones being *"tinny"* or *"hummy"*. The experimenters acknowledge that the selection of a bad microphone is subjective and that a microphone that created 'bad' audio with one soundcard will not necessarily do so with another. Yet, it was viewed as being important to investigate, as many subjective comments make reference to how the voice sounds, and whether or not it is pleasant to listen to (Preminger & Tasell 1995).

Three Internet audio experts listened to the material and agreed that they contained the degradations they represented. In addition, the results from a pilot study with six participants (who were all Internet audio novices) where they had to rate the quality of the audio confirmed this.

#### 5.2.3.2 Subjective assessment materials

In order to gather subjective opinions of the quality, a 100-point scale was utilised, which was labelled at the end points with very poor quality (1) and very good quality (100). Participants were also asked to describe why they had awarded each rating. Table 7 shows the assessment methods used within the 3-factor approach in this experiment.

| Element in 3-factor framework | Measure |
|---|---|
| Task performance | Rate quality of audio |
| User Satisfaction | Rating scale after each condition and oral justification of rating |
| User cost | Physiological measures |

**Table 7: The 3-factor approach used in experiment 2**

### *5.2.4 Participants*

Twenty-four participants took part in the study. There were twelve males and twelve females. They all had good hearing and were between eighteen and twenty-eight years of age. They were all Internet audio and MMC novices.

### 5.2.5 *Procedure*

Participants were briefed about the physiological measures that would be taken (SC, HR and BVP) and were informed that they were free to withdraw from the experiment at any time. Participants were instructed to try and remain as still as possible throughout the experiment so as not to interfere with the signals. The sensors were then attached to their fingers and the baseline physiological responses were gathered for fifteen minutes whilst participants read a newspaper. Participants then listened to a one-minute volume reference file. Following that, they were asked if they were happy with the volume. If not, it was adjusted and that level was used throughout the experiment for that participant. Participants were told that the reference file should be regarded as being of the best quality that they would receive.

Six two-minute randomised order test files were then played, during which time the physiological responses were measured. After this. the reference condition was played again and was followed by the six test files again, which were repeated in a different order. After each test file participants gave a quality rating for the file as a whole on the 100-point scale. They were then asked to orally explain why they had given the rating they did and this was tape-recorded.

### 5.2.6 *Hypotheses*

1. It will be possible to detect significant changes in SC, HR and BVP due to audio degradations.
2. The reference condition will induce the least perceptual strain.
3. There will be an agreement between subjective responses and physiological results, as the task is not engaging.

### 5.2.7 *Results*[20]

#### 5.2.7.1 Physiological results

One participant had to be dropped from the analysis due to a fault with the recording equipment, which left data for 23 participants. For each signal, a two-way repeated measures ANOVA with within-subjects variables degradation (7 levels) and presentation (2 levels) and the between-subjects variable gender was performed.

SC

The data were cleaned for outliers[21]. There was a significant main effect of degradation ($F_{(6,114)}$ = 2.842, p=0.013). Post-hoc pairwise comparisons showed there was a significant difference between the reference and echo conditions (p=0.012) and the reference and loud conditions (p=0.008). Examination of the means (figure 21) shows that the reference condition was significantly lower than echo and loud.

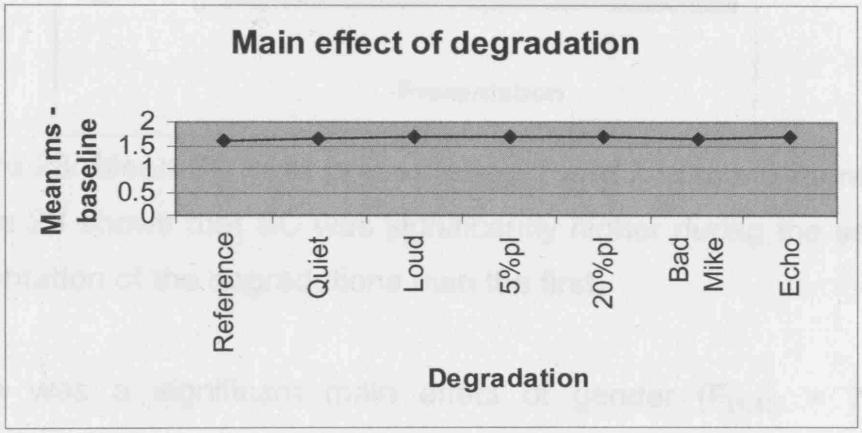



**Figure 21: Mean SC over conditions**

[20] In Wilson & Sasse (2000c) the mean response over both presentation of conditions was generated and showed a significant impact of audio degradation on HR and BVP.
[21] Two participants (3 & 12) were removed. This left 11 males and 10 females.

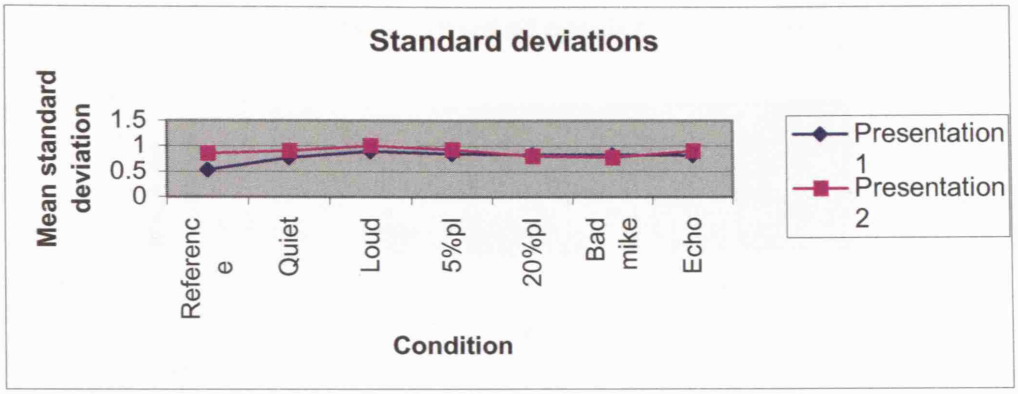



**Figure 22: Mean SC standard deviations over conditions in experiment 2**

There was a significant main effect of presentation of degradation ($F_{(1,19)}$ = 15.359, p=0.001).

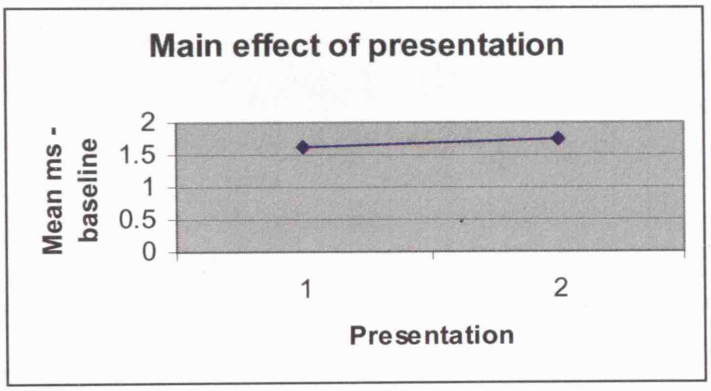



**Figure 23: Mean SC over presentation 1 and 2 in experiment 2**

Figure 23 shows that SC was significantly higher during the second presentation of the degradations than the first.

There was a significant main effect of gender ($F_{(1,19)}$ = 7.935, p=0.011). Examination of the means shows that males had a significantly higher SC than females.

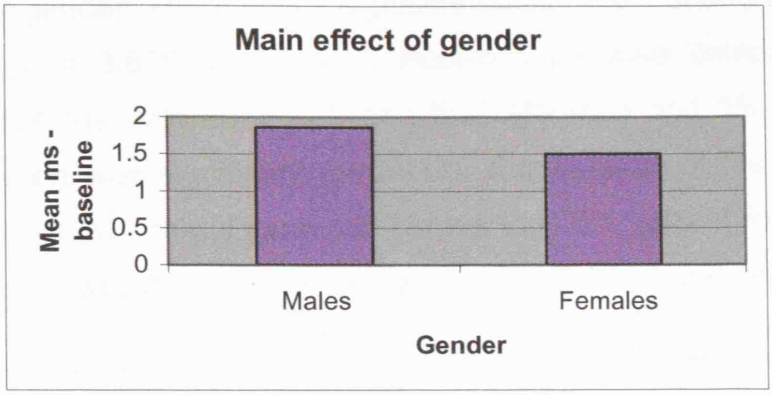



**Figure 24: Mean SC for males and females in experiment 2**

There was a significant interaction between presentation and gender ($F_{(1,19)}$ = 13.967, p=0.001).

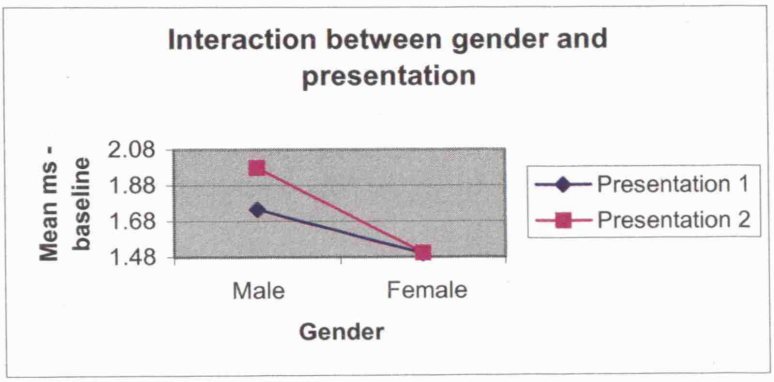



**Figure 25: Interaction between gender and presentation for SC in experiment 2**

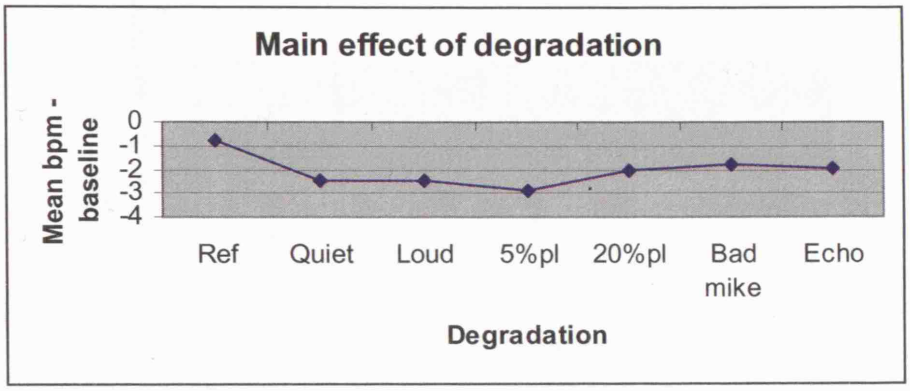
Figure 25 shows that during presentation, 2 males had a higher SC than during presentation 1, yet the SC of females was the same during both presentations, which was below that of the males for both presentations.

HR

The data were cleaned for outliers[22]. The two way repeated measures ANOVA revealed that there was not a significant effect of presentation of condition, therefore the mean responses to each condition were calculated and a one-way ANOVA was performed with the within-subjects variable degradation and a between-subjects

[22] Four participants (7, 8, 11 & 20) were removed. There were 11 males and 8 females.

variable of gender. There was a significant main effect of degradation ($F_{(3.060,52.026)}$ = 3.870, p=0.014)[23]. Post-hoc pairwise comparisons showed that the difference between the reference and 5% packet loss condition was significant (p=0.017). Examination of the means (figure 26) shows that Figure 26 shows that 5% packet loss was significantly lower than the reference condition, which indicates less




strain.

**Figure 26: Mean HR over conditions in experiment 2**

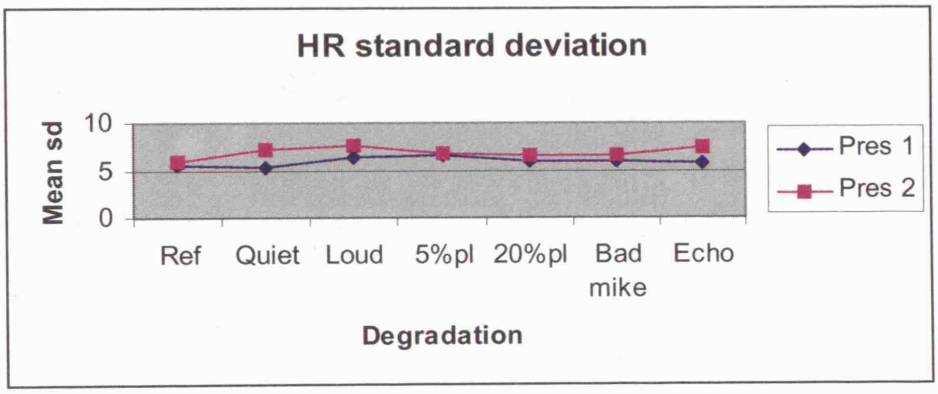



**Figure 27: Mean HR standard deviations over conditions in experiment 2**

BVP

The data were cleaned for outliers[24]. There was a significant main effect of degradation ($F_{(6,120)}$ = 2.767, p=0.015). Post-hoc pairwise

---

[23] Mauchly's test of sphericity was significant, therefore the Greenhouse-Geisser correction was used

[24] One participant (14) was removed. There were 12 males and 10 females.

comparisons showed no significant differences between conditions, however the difference between the bad microphone and loud condition was approaching significance (p=0.061). This lack of significant results is most likely due to the conservative nature of the Bonferroni correction (see section 4.3.5).

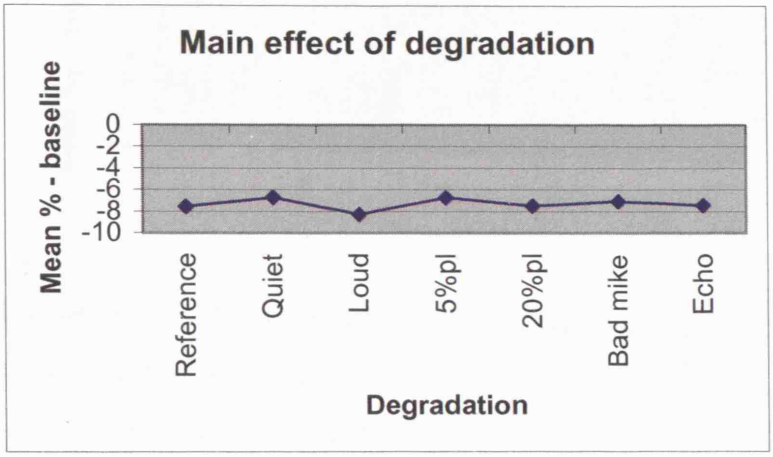



**Figure 28: Mean BVP over conditions in experiment 2**

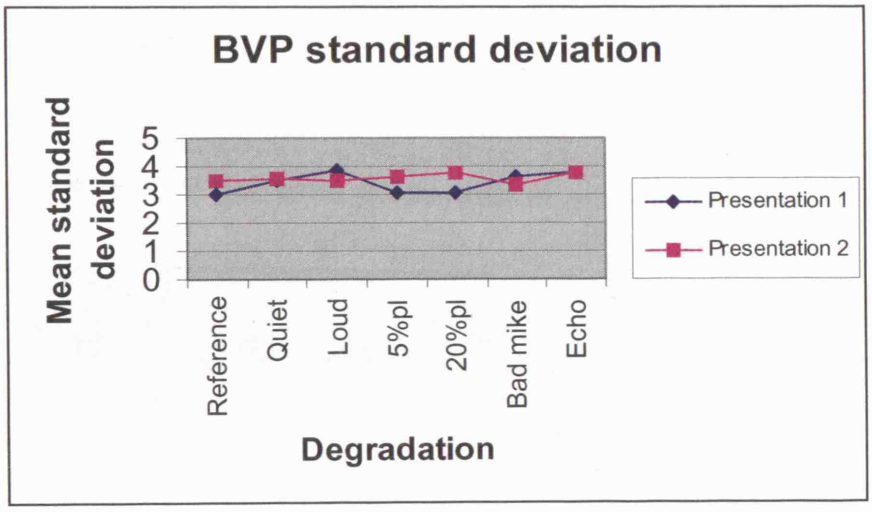



**Figure 29: Mean BVP standard deviations over conditions in experiment 2**

There was a significant main effect of presentation ($F_{(1,20)}$ = 6.821, p=0.017).

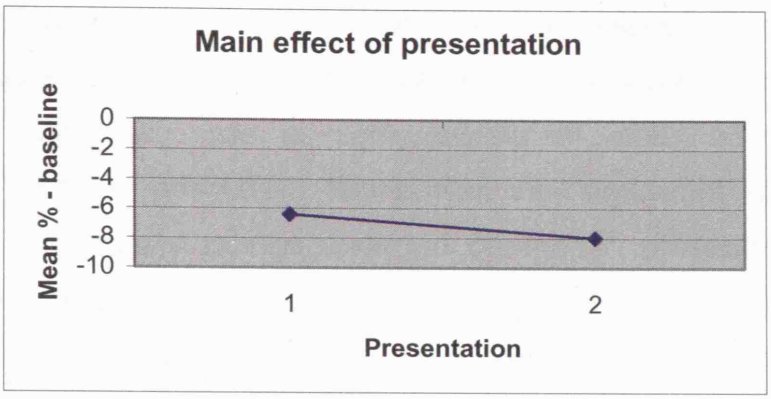



**Figure 30: Mean BVP over both presentations in experiment 2**

Figure 30 shows that BVP was significantly lower during presentation 2 than 1, which indicates more strain.

Finally, there was a significant interaction between degradation and presentation ($F_{(6,120)}$ = 3.398, p=0.004).

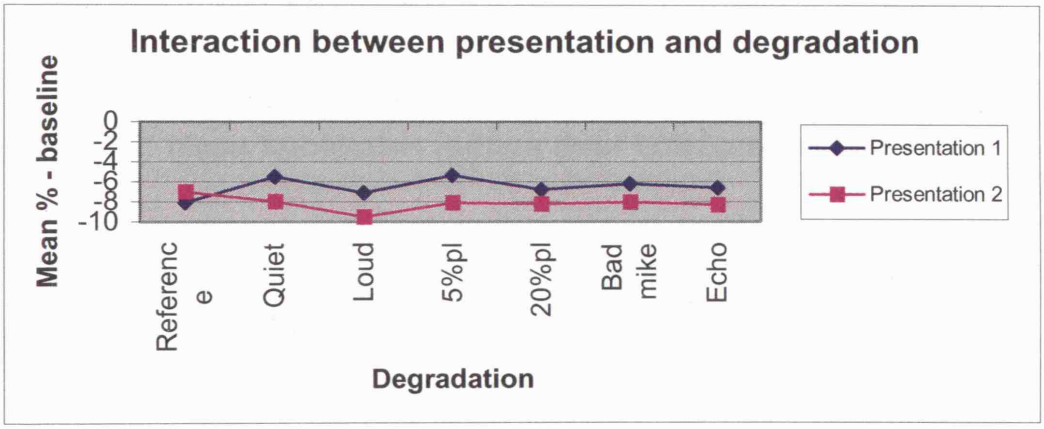



**Figure 31: Interaction between degradation and presentation for BVP in experiment 2**

Figure 31 shows that the interaction between degradation and presentation occurs at the reference conditions, which is the only point where presentation 1 has a lower BVP than presentation 2.

### 5.2.7.2 Subjective results

Analysis of the subjective data was carried out by Anna Watson and published in Watson & Sasse (2000). The results are summarised here to allow comparison with the physiological results.

A two-factor with replication ANOVA at the 1% level of probability showed that there was a highly significant effect of condition ($F(6,322)=62.25$, $p=<0.01$). There was no significant difference between the quality ratings awarded on the 1st presentation and those awarded at the second time of hearing ($F(1,322)=0.799$). Therefore, the mean responses were taken for each participant. Figure 32 shows the mean ratings.

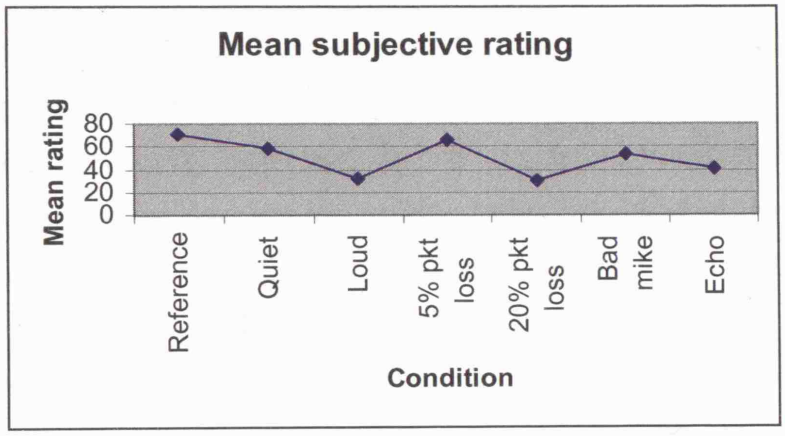



**Figure 32: Mean subjective ratings over both presentations**

An ANOVA on the combined means showed there was a highly significant main effect of condition ($F(6,161)=36.598$, $p<0.01$). Tukey HSD tests revealed the following differences.

- The differences between the reference condition and all conditions with the exception of 5% packet loss are significant. The reference condition is rated as the best quality.
- There is no significant difference between the 5% packet loss condition and the quiet condition (Qobt = 4.36).
- The 5% packet loss condition is rated as being of significantly better quality than: echo (Qobt = 9, $p<0.01$); loud (Qobt =

12.41, $p<0.01$); 20% packet loss (Qobt = 13.43, $p<0.01$); and bad microphone (Qcrit = 4.17, Qobt = 4.33, $p<0.05$).

### *5.2.8 Discussion of results*

#### 5.2.8.1 Physiological results

There was a significant impact of audio degradation on all three signals. SC responses were above baseline, whereas HR and BVP were below baseline. Thus, as in experiment 1 directional fractionation (see section 3.2.2.2) is observed. This indicates that the task was primarily perceptual. This was expected because the nature of the task meant that participants were solely focused on the quality.

The reference condition was expected to be the least straining and this was found in SC, as it had significantly less perceptual strain than echo and loud. However, in HR the reference condition had significantly more perceptual strain than 5% packet loss. This unexpected result in HR may have been because it was heard at the beginning of the experiment and at the start of the second presentation of degradations. Therefore, SC may be reflecting the quality of the condition, whereas HR may be reflecting anxiety due to the beginning of the experiment.

The second presentation of the signals was found to be significantly more straining than the first in both SC and BVP. This result may be indicative of either boredom or frustration at hearing exactly the same clips twice, or it could be due to an accumulative effect of the degraded audio quality. In BVP, the only condition in presentation 1 to cause more strain than in presentation 2 was the reference condition. This is most likely reflecting anxiety due to the experiment beginning.

There was a significant effect of gender in SC, where males had a higher SC than females and the interaction between gender and presentation showed that females had around the same level of SC

for both presentations, whereas males had a higher SC for the second presentation. There were 11 males and 10 females in the SC analysis, therefore this result cannot be explained by a gender imbalance. It may be due to the male participants getting more bored or frustrated with the experiment generally and especially at hearing the same clips twice, which could explain their SC being higher during the second presentation. Alternatively, this result may be due to both voices in all the clips being male and male participants responding more strongly to this.

#### 5.2.8.2 Subjective results

The subjective results illustrate that 20% packet loss, which occurred occasionally in the field trial mentioned in section 5.2.1, was rated as being of the worst quality. However, 5% loss, which was the more common rate of packet loss experienced in the trial, was rated as being of the second best quality. This result illustrates that degradations, such as loud volume differences between speakers and echo can have more of an impact on subjective opinions of the quality than the common level of audio packet loss experienced. Consequently, the impact they can have should be considered by network providers and application designers, who typically focus on packet loss.

#### 5.2.8.3 Combining physiological and subjective results

When examining physiological responses to the conditions, loud and echo caused significantly more strain than the reference condition in SC and 5% packet loss caused significantly less perceptual strain than the reference condition in HR. Therefore, the SC signal concurs to a degree with subjective responses in that echo and loud were 2 of the 3 worst rated conditions and were rated as being significantly worse than the reference condition. Interestingly, 20% packet loss did not induce any significant increases in perceptual strain in any of the signals. Therefore, despite it being rated as the worst it does not have a significant physiological impact on participants. This may be

because the quality is so poor that participants subjectively rate it as being poor almost immediately then stop attending to it. This would require further investigation in a task with more ecological validity. If 20% packet loss occurred in, for example a distance learning scenario, users could not stop attending to the channel and would have to expend effort to decode the information, which could cause an increase in perceptual strain. In addition, in HR 5% packet loss was significantly less straining than the reference condition, and this was rated as the second best condition. This agreement between subjective and physiological responses is most likely due to the non-engaging nature of the task, thus participants were solely focused on rating the quality.

### *5.2.9 Addressing the hypotheses*

1. It will be possible to detect significant changes in SC, HR and BVP due to audio degradations. This hypothesis was supported in all signals.
2. The reference condition will cause the least perceptual strain. This hypothesis was partly supported. It caused significantly less perceptual strain than echo and loud in SC but more than 5% packet loss in HR.
3. There will be a correlation between subjective responses and physiological results, as the task is not engaging. This hypothesis was partly supported. Echo and loud were among the three worst rated conditions and were significantly more straining than the reference condition in SC and 5% packet loss was rated as being of the second best quality and significant less straining than the reference condition in HR.

### *5.2.10 Limitations*

The main limitation of this experiment is that listening to and rating the quality of audio is not ecologically valid. However, it was important to determine if differences to audio degradations could be detected in a passive task with minimal variables in operation before adding more variables, such as the video channel.

### *5.2.11 Conclusions*

The results of this experiment show that in a passive perceptual task, physiological responses to audio degradations can be detected. Therefore, the results of this experiment provide support for the use of this method in MMC evaluation. The significant physiological results were in agreement with results from subjective rating scales, with the exception of the reference condition in HR (see section 5.2.8.1), which illustrates that in passive, non-engaging tasks subjective methods can give an accurate indication of the user's opinion of the quality.

The results also demonstrate the additional information than can be gained from the physiological data, which remains untapped by subjective data. For example, there were significant differences between the 2 presentations of conditions in SC and BVP, yet there were no significant differences between the subjective ratings given during the first and second conditions. Thus, even though participants can accurately perform their task (rating the quality), physiological results indicate more strain during the second presentation of conditions. Added information is also obtained from observing that SC and HR were fractionated, which indicates that the task was perceptual. This was expected because the task was one of passive listening and rating, thus had a minimal cognitive element.

## 5.3 Experiment 3: Investigating the Impact of Audio Degradations in a Recorded Interview Task

### *5.3.1 Introduction*

The results of experiment 2 showed that physiological responses to audio degradations can be detected. In experiment 3, four audio degradations from experiment 2 (20% packet loss, 5% packet loss, audio recorded using a bad microphone and loud volume) were examined with the video channel present in the context of a more engaging and longer task to determine if the same results would be found.

### *5.3.2 Design*

This experiment was performed in the Computer Science department of UCL. Twenty-three participants watched 4 recorded interviews. The interviews were between admission tutors and candidates for the degree course at UCL. There were 4 candidates in total and each participant saw each candidate once. The interviews were 10 minutes in length. Group 1 (in which there were 12 participants) experienced 5 minutes of normal quality followed by 5 minutes of degraded quality in all four interviews. Group 2 (in which there were 11 participants) experienced 5 minutes of degraded quality followed by 5 minutes of normal quality in all four interviews. The order of presentation of the conditions was randomised. The video quality was good (25fps) to minimise the impact of this channel and as in experiment 1, the participants watched solely the video of the candidate on the screen.

### *5.3.3 Materials*

#### 5.3.3.1 Experiment materials

A pilot trial was carried out with six experts in the Computer Science department at UCL to ensure that the conditions were representative of the degradations. All experts were in agreement that they were. The interviews were between university admissions tutors and candidates (3 males, 1 female) for the Computer Science degree

course at UCL. The interviews were recorded, scripted and acted, as in experiment 1. In order to generate the recordings, three machines were used (one unix and 2 PCs) that communicated via a hub. They were standalone, to ensure that network traffic did not interfere with the files. There were a total of five experimental files listed below, which were recorded and played out on a PC through Rat and vic.

1. An interview with good audio quality, which was recorded with 2 good Gamma[25] headsets. A reflector[26] was used once the playback had started to induce high levels of audio packet loss manually by the experimenter at the required point in the interview.
2. An interview with good audio quality, which was recorded with 2 good Gamma headsets. Again the reflector was used to induce low levels of packet loss manually at the set time.
3. An interview with good audio quality, recorded with 2 good Gamma headsets. The PC was connected to an amplifier that controlled the volume. The loud volume was manually induced at the required point by the experimenter to a level that the six experts agreed could be classed as 'loud'.
4. An interview recorded using a bad microphone for the first five minutes and a good microphone for the second five minutes. To allow the interviewee to change his microphone half way through the interview, it was scripted that he had dropped something on the floor that he had to pick up. When he was out of view of the camera, he changed his headset. The bad microphone used was the same one used in experiment 2 and was established through pilot testing as being of poor quality.
5. An interview recorded with a good microphone for the first five minutes and the same bad microphone for the second five minutes.

---

[25] Manufacturer of headsets.

26 A reflector is an application that listens on configured port(s) for packets and then sends those packets either back to the sender, or to a configured address or port. Dependent on the configuration, the reflector may be also set-up to drop or delay packets that pass through it.

#### 5.3.3.2 Subjective assessment materials

After each interview, a questionnaire was administered, which was comprised of 6 questions (see appendix C). The standard candidate assessment forms used in the computer science department at UCL were administered for completion after each interview (see Appendix B). In addition, after each interview the participants were asked if they thought the candidate should be offered a place on the Computer Science degree course at UCL. Table 8 shows the assessment measures used as part of the 3-factor evaluation approach in experiment 3.

| Element in 3-factor framework | Measure |
|---|---|
| Task performance | Decide if each candidate should be offered a place |
| User Satisfaction | Questionnaire after each interview on quality and adequacy of audio and video |
| User cost | Physiological measures |

**Table 8: The 3-factor approach used in experiment 3**

### *5.3.4 Participants*

There were 23 participants in this experiment (one had to be dropped due to problems with the experimental material). Twelve were female and eleven were male. They were paid £10 and were students recruited from within UCL.

### *5.3.5 Procedure*

Participants were told they would be watching four recorded interviews and that their physiological responses would be measured. They were informed that they were free to withdraw from the experiment at any time. Once the baseline session was finished, the experimenter introduced the first candidate and the interview was shown. When it finished, a questionnaire and candidate assessment form were administered and the participant was asked if the candidate should be offered a place. The same procedure occurred for the remaining three interviews.

### 5.3.6 *Hypotheses*

1. There will be a significant impact of audio degradations on the physiological signals.
2. The conditions that will cause most perceptual strain will be loud and 20% packet loss, whereas 5% packet loss and the bad microphone condition will cause the least perceptual strain.
3. There will be less agreement between the physiological and subjective results than in experiment 2, due to the engaging nature of the task.

### 5.3.7 *Results*

#### 5.3.7.1 Physiological results

Data for 1 participant (number 18) had to be excluded due to problems with the recording equipment. In all signals and for each participant, the mean of the normal quality section was subtracted from that of the degraded audio section. Therefore, in this experiment a task baseline as opposed to a resting baseline was utilised. For each signal a repeated measures ANOVA was performed with degradation as a within-subjects variable with 4 levels (20% packet loss, 5% packet loss, loud volume, audio recorded using a bad microphone) and group (2 levels) and gender as between-subjects variables.

SC

The data were cleaned to remove outliers[27]. No significant differences were found in the ANOVA. The mean SC and standard deviation of the signal can be seen in figures 33 and 34 respectively.

[27] Four participants (4, 9, 11 & 22) were removed, leaving 9 participants in group 1 and 9 in group 2 and 9 males and 9 females.

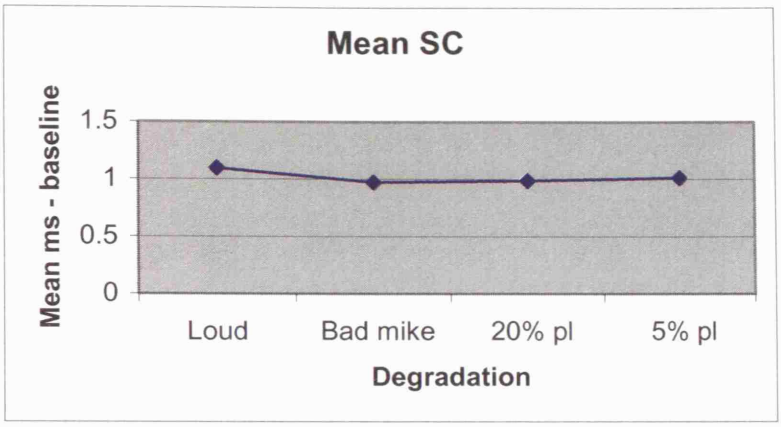



**Figure 33: Mean SC responses to degradations in experiment 3**

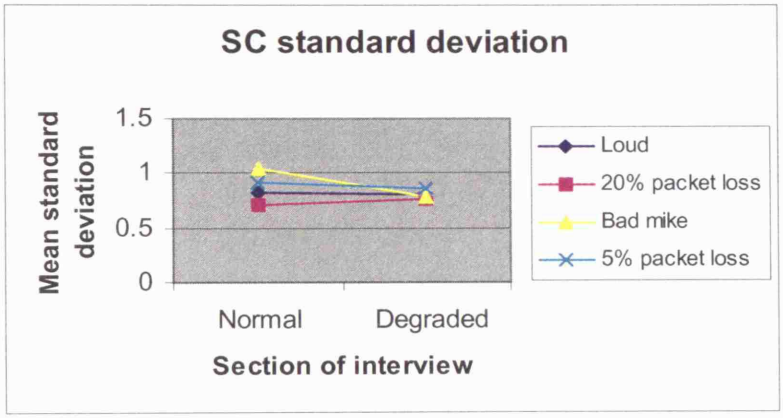



**Figure 34: Mean SC standard deviations in experiment 3**

HR

The data were cleaned to remove outliers[28]. The 5% packet loss condition was not normally distributed, as shown by the Shapiro-Wilk test in SPSS (p=0.01), therefore it was excluded from the ANOVA. No significant differences were found in the ANOVA. The mean HR and standard deviation of the signal can be seen in figures 35 and 36 respectively.

[28] Two participants (8 & 21) were removed, leaving 11 participants in group 1 and 9 in group 2 and 11 males and 9 females.

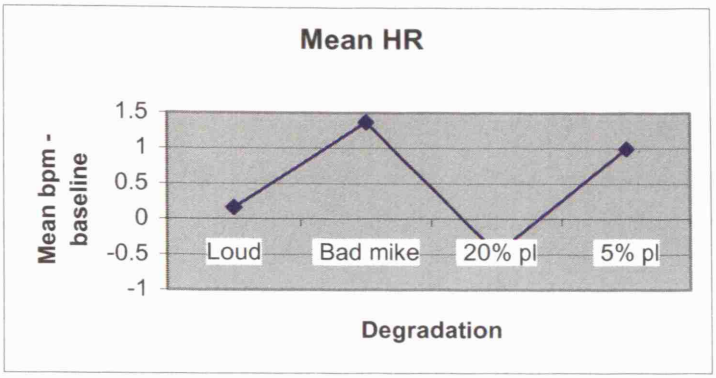



**Figure 35: Mean HR responses to degradations in experiment 3**

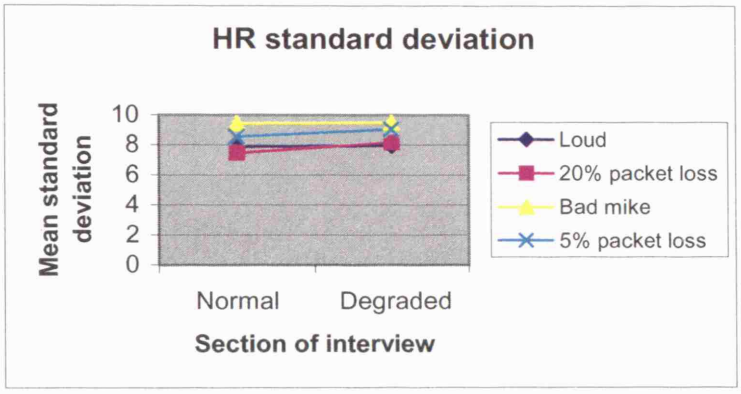



**Figure 36: Mean HR standard deviations in experiment 3**

BVP

The data were cleaned to remove outliers[29]. The 5% packet loss condition was not normally distributed, as shown by the Shapiro-Wilk test in SPSS (p=0.001), therefore it was excluded from the ANOVA. The main effect of group was approaching significance ($F_{(1,12)}$ = 4.424, p=0.057).

[29] Six participants (1, 2, 7, 13, 14 & 17) were removed, leaving 9 participants in group 1 and 7 in group 2 and 7 males and 9 females.

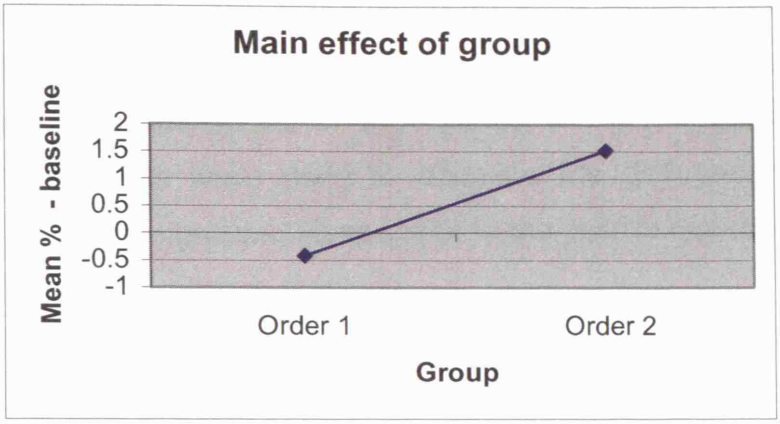



**Figure 37: Mean BVP for both groups in experiment 3**

The mean BVP and standard deviation for the signal can be seen in figures 38 and 39 respectively.

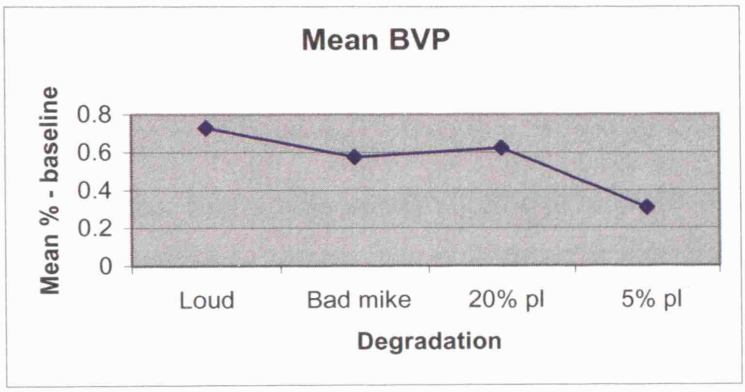



**Figure 38: Mean BVP responses to degradations in experiment 3**

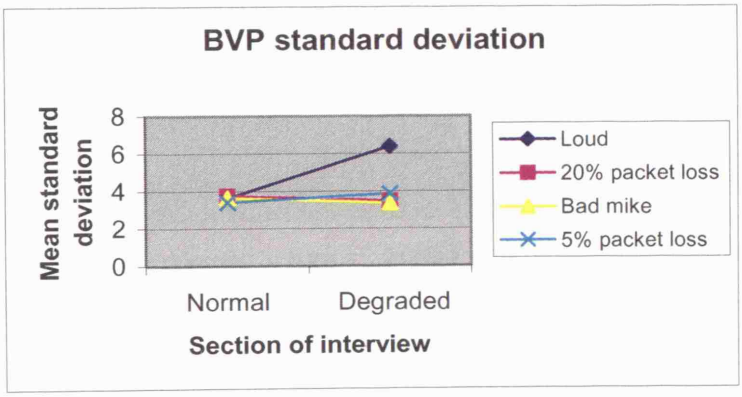



**Figure 39: Mean BVP standard deviations in experiment 3**

### 5.3.7.2 Subjective results

A short questionnaire was administered to participants after each interview (see appendix C). As there were yes/no questions in these questionnaires that have to be analysed using non-parametric

statistics, responses to the rating questions were also analysed using non-parametric statistics for consistency. Friedman tests were used to examine the differences in the ratings within the groups and Kruskal-Wallis tests were used to examine the differences in the ratings between the groups. Responses to the yes/no questions were analysed using Chi-square tests.

Question 1: What did you think of the quality of the audio?
One participant omitted to rate one condition, therefore this was replaced with the overall group mean for the condition. The differences between the ratings in group 1 were not significant, however they were in group 2 (Chi square with 3 degrees of freedom = 20.83, p=0.00). In group 2 the loud condition was rated as being of the best quality, followed by bad microphone, 5% packet loss and 20% packet loss.

Between the groups there was a significant difference between the 20% packet loss ratings, where group 2 rated it significantly lower than group 1 (1 degree of freedom = 6.30, p=0.011).

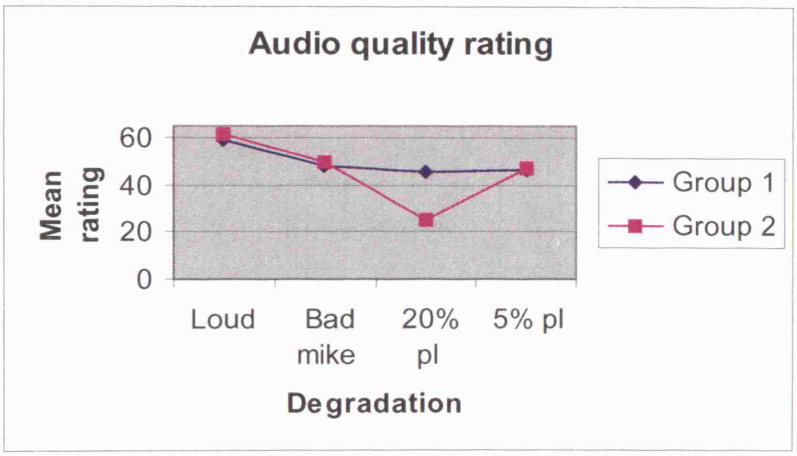



**Figure 40: Mean audio quality ratings combined over both groups in experiment 3**

Figure 40 shows that over both groups the loud condition was rated highest, followed by bad microphone, 5% packet loss and 20%

packet loss. The means for group 1 and 2 are similar with the exception of 20% packet loss.

Question 2: How adequate was the audio quality for the purposes of the interview?

Two participants omitted to rate one condition, therefore these were replaced with the overall group mean for the condition. In group 1 there was a significant difference in the ratings (3 degrees of freedom = 14.16, p=0.003). Figure 41 shows that the order of rating from best to worst was loud, 5% packet loss, 20% packet loss and bad microphone. In group 2 there was a significant difference in the ratings (3 degrees of freedom = 20.01, p=0.000). Figure 41 shows that the order of rating from best to worst was loud, bad microphone, 5% packet loss and 20% packet loss.

There were no significant differences in ratings between the groups. Figure 41 shows that the overall order of ratings from best to worst was loud, 5% packet loss, bad microphone then 20% packet loss.

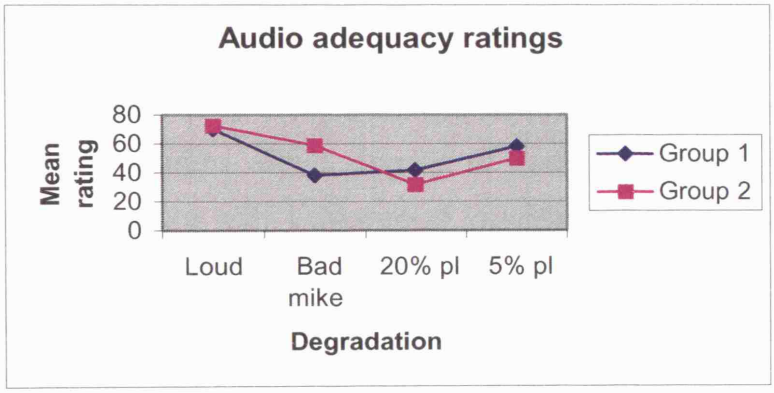



**Figure 41: Mean audio adequacy ratings in experiment 3**

Question 3: Did the quality of the audio change throughout the interview?

One participant in group 1 did not answer the question for the 5% packet loss condition and one in group 2 did not answer the question for the 20% packet loss condition. Therefore, these participants were excluded from the analysis on this question.

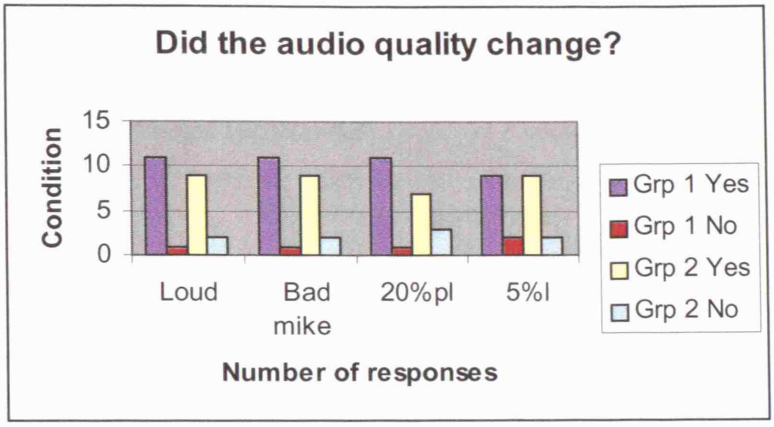



**Figure 42: Responses to question 3 in experiment 3**

Figure 42 shows responses to question 3 for group 1 and 2. Chi-square tests were performed on the data for group 1. These showed that significantly more participants in group 1 noticed that the audio had changed in all conditions: loud (Chi-square with 1 degree of freedom = 8.33, p=0.004); 20% packet loss (Chi-square with 1 degree of freedom = 8.33, p=0.004); bad mike (Chi-square with 1 degree of freedom = 8.33, p=0.004); and 5% packet loss (Chi-square with 1 degree of freedom = 4.45, p=0.035).

Chi-square tests were performed on the data for group 2. These showed that significantly more participants noticed the audio change in the loud (Chi-square with 1 degree of freedom = 4.45, p=0.035), bad mike (Chi-square with 1 degree of freedom = 4.45, p=0.035) and 5% packet loss conditions (Chi-square with 1 degree of freedom = 4.45, p=0.035). The difference in the 20% packet loss condition was not significant.

Question 4: What did you think of the quality of the video?

Two participants omitted to rate one condition, therefore these were replaced with the overall group mean for the condition. The ratings can be seen in figure 43. There were significant differences in group 1's ratings to the conditions (3 degrees of freedom = 9.027, p=0.029). The order of ratings from best to worst was 5% packet loss, 20% packet loss, bad microphone then loud. There were no significant differences in group 2's ratings. There were no significant

differences in ratings between the groups. The overall order of ratings from best to worst was 5% packet loss, 20% packet loss, bad microphone then loud (figure 43).

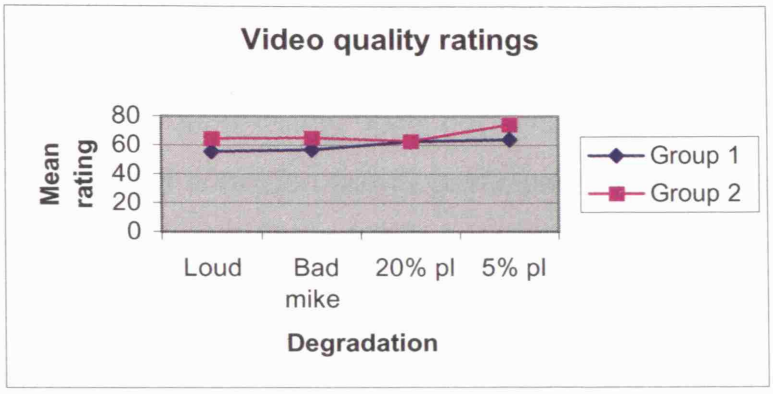



**Figure 43: Mean video quality ratings in experiment 3**

Question 5: How adequate was the video quality for the purposes of the interview?

Three participants omitted to rate one condition, therefore these were replaced with the overall group mean for the condition. There were no significant differences in group 1 or 2 or between the groups. Figure 44 shows that the overall order of ratings from best to worst was 5% packet loss, 20% packet loss, loud then bad microphone.

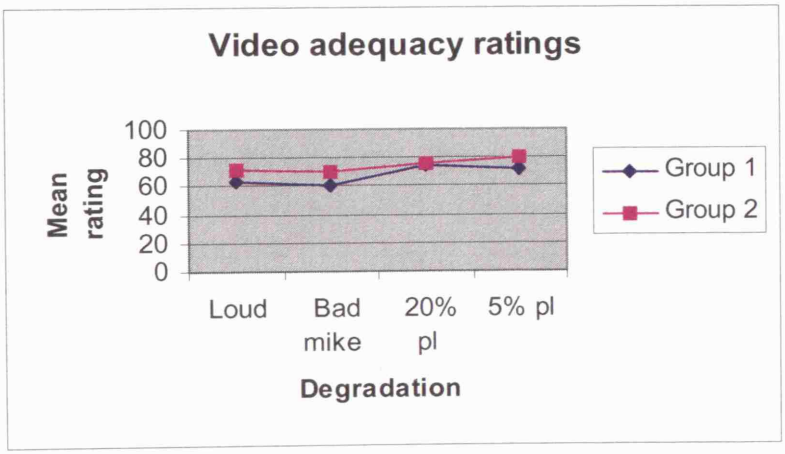



**Figure 44: Mean video adequacy ratings in experiment 3**

Question 6: Did the quality of the video change throughout the interview?

One participant in group 2 did not answer this question for the 20% packet loss condition, therefore this participant was excluded from the analysis on this question. Responses can be seen in figure 45. Significantly more participants said the video quality did not vary throughout the loud condition (all 12 participants said it did not vary), the 20% packet loss condition (Chi-square with 1 degree of freedom = 8.33, p=0.004) and the 5% packet loss condition (Chi-square with 1 degree of freedom = 8.33, p=0.035). The difference in the bad microphone condition was not significant.

In group 2, the only significant difference was in the 20% packet loss condition (Chi-square with 1 degree of freedom = 8.33, p=0.035).

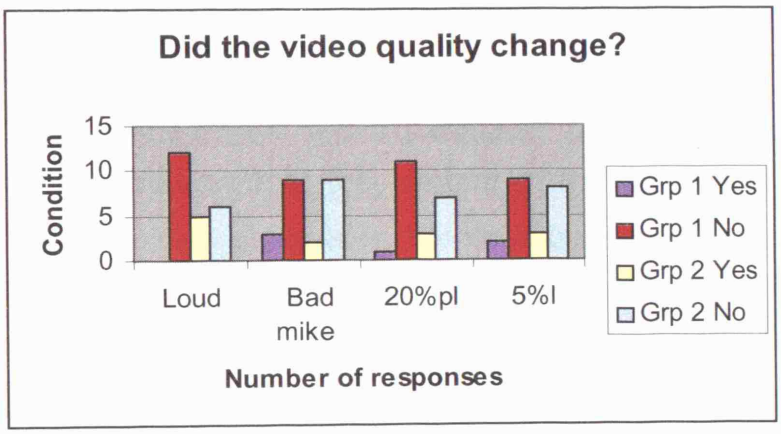



**Figure 45: Responses to question 6 in experiment 3**

Question 7: Should the candidate be offered a place on the computing degree course at UCL?

1 participant in group 1 and 1 in group 2 did not answer this question, therefore these participants were excluded from the analysis on this question. The only candidate whom more participants thought should not be offered a place was the candidate in the 5% packet loss condition (2 participants in group 1 and 0 in group 2 thought he should be offered a place).

### *5.3.8 Discussion of results*

#### 5.3.8.1 Physiological results

There were no significant differences in the physiological data. This may be for a number of reasons. Hearing audio degradations with the video channel may be less straining than hearing them alone (as in experiment 2, where significant differences were obtained). Alternatively, in this experiment the baseline used was the five minutes of normal quality. It may be that in this area utilising what is effectively a task baseline as opposed to a resting baseline is not effective, as the range of responses will be smaller. In the loud condition, the entire interview was loud, whereas in experiment 2 there were loud volume differences between speakers. Therefore, the lack of a significant result in the loud condition (which was significantly more straining than the reference condition in SC in experiment 2) could be due to differences in volume being more straining than loud volume throughout.

In addition, it could be that the content of the conditions affected participants more than the quality. For example, the candidate in the 5% loss condition, who was the only candidate that more participants thought should not be offered a place, did not adhere to the script as much as the other candidates and as a result their responses to him may have affected results. This and the fact that the 5% packet loss condition was not normally distributed in the HR and BVP signals indicate that the results of this condition should be interpreted with caution. Finally, the crude way of changing from a good to bad microphone may have also affected participants more than the degradation itself.

#### 5.3.8.2 Subjective results

The overall audio quality ratings from best to worst were loud, bad microphone, 5% packet loss and 20% packet loss. The overall audio adequacy ratings from best to worst conditions were loud, 5% packet

loss, bad microphone then 20% packet loss. Therefore, there is an agreement between audio quality and adequacy ratings as 20% was rated as the worst quality and least adequate and loud was rated as the best quality and most adequate. In experiment 2, loud audio was rated as having the second worst quality. These results indicate that when loud audio is heard as part of a multimedia conference, in an engaging task and on all of the audio stream, as opposed to volume differences between speakers, it is rated as being of a better quality than when it is heard as part of a rating task and without the video channel. There is also an agreement with the subjective results of experiment 2, where 20% packet loss was rated as being of the worst quality.

The video quality remained constant throughout all interviews and was at a high frame rate. Despite this, it was interesting to gather views on video quality to determine if audio degradations had an impact on video quality ratings. Group 2 rated the video quality higher than group 1, however group 1 rated the conditions significantly differently whereas group 2 did not. The overall order of ratings from best to worst was 5% packet loss, 20% packet loss, bad microphone and loud. Group 2 also rated the video adequacy higher than group 1. The overall order of ratings from best to worst was 5% packet loss, 20% packet loss, loud then bad microphone. Therefore, there is an agreement between video quality and adequacy ratings because 5% packet loss was rated as best quality and most adequate, followed by 20% packet loss.

The results indicate that when audio quality is rated as being good (loud) the video quality is rated as being poor. But when audio quality is rated as being poor (20% packet loss) the video quality is rated as being second best to 5% packet loss. Investigating this further with an eye-tracking machine would determine if the video is looked at more or less when the audio is poor.

Significantly more participants in group 1 noticed that the audio had changed in all conditions. In group 2 the only condition where the change was not noticed was the 20% packet loss condition. This condition was rated by group 2 as the lowest and was significantly lower than the rating given by group 1. An explanation for this result could be that starting an interview with 20% packet loss leads to overall impressions of the interview, even when there has been normal quality, being lowered and subsequent changes not being noticed as users may disregard the audio and focus more on the video, which is supported by video quality being rated as the second best for the 20% packet loss condition. This result can be compared to that in experiment 1 where only around 15% of participants overall noticed that the frame rate had changed, therefore it appears that changes in the quality of the audio channel are more noticeable than changes in the video channel, as they were noticed by over 80% of participants (when averaged across all conditions and both groups).

The video quality did not change during this experiment. Significantly more participants in group 1 said that the video quality did not vary in all conditions except the bad microphone condition, whereas this was the only condition in group 2 that was significant. This result may have been affected by the crude way of changing the microphone from good to bad. This result may also indicate that changing audio quality in an interview from bad to good may make participants less able to judge the video quality.

With regard to the decision about who should be offered a place, more participants in both groups said that candidates should be offered a place than those who said they should not, with the exception of the candidate in the 5% packet loss condition. This result provides support for the theory that responses to this candidate, who did not adhere to the script, may have affected responses to the quality.

### 5.3.9 *Addressing the hypotheses*

1. There will be a significant impact of audio degradations on the physiological signals. This hypothesis was not supported.
2. The conditions that will cause most perceptual strain will be loud and 20% packet loss, whereas 5% packet loss and the bad microphone conditions will cause less perceptual strain. Due to no significant differences being obtained, no comment can be made here.
3. There will be less agreement between physiological and subjective results than in experiment 2, due to the engaging nature of the task. Due to no significant differences being obtained in the physiological data, no comment can be made here.

### 5.3.10 *Limitations of the experiment:*

There were three limitations of this experiment, which may have influenced results. Firstly, there was no control condition, where participants watched ten minutes of good quality. This would have allowed responses under degraded interviews to be analysed and also to determine whether the change from good to degraded quality or vice versa was having an impact on results. Secondly, changing the audio within the condition may have affected responses, therefore it may have been better to have four conditions with the same quality level throughout. Thirdly, the candidate in the 5% packet loss condition did not adhere to the script as much as the other candidates, which may have influenced results. Using actors should control for this, therefore this was done in the next two experiments.

### 5.3.11 *Conclusions*

The results show that there was no significant impact of audio degradation on the physiological signals. This is in opposition to experiment 2, which found a significant impact of audio degradation on all signals in a passive task (see section 5.2.7.1). Therefore, it appears that adding the video channel compensates for the

perceptual strain induced by audio degradations. This fits in with previous research showing an increase in intelligibility of the audio channel when video is added.

An order effect was shown in the subjective results, where the group that started with good quality had better overall opinions of the candidates than the group who started with poor quality. This may be a primacy effect. In addition, the changes in the audio conditions were noticed in all good to bad conditions but only in the loud and 20% packet loss conditions (which were rated as being of the worst quality) where the degraded quality was seen first. Thus, changes in quality may be less noticeable if they occur after degraded quality, which again may be a primacy effect.

From the means it can be seen that all signals are above baseline (with the exception of the HR for the 20% packet loss condition). This is at odds with the data from the previous two experiments. From looking at the overall experiment means (see table 9), it can be seen that out of experiments 1-3, experiment 3 has the lowest SC, the highest HR and the highest BVP. However, it must be pointed out that this experiment used a task baseline as opposed to a resting baseline and this may have affected results. Alternatively, it may be that audio degradations in an engaging task with the video channel cause less perceptual strain than hearing them in isolation, as the video channel can assist in working out what participants are saying.

This is the only experiment where a fractionation between SC and HR has not occurred. Following the directional fractionation hypothesis, this result would be explained by the HR increasing to reject stimuli that would be disruptive to the performance of a cognitive function. The exception is 20% packet loss, which was below the HR baseline. This could indicate that this condition caused participants to focus on the video channel to gain additional information, due to the severity of the degradation on the audio

channel. The task in this experiment was to decide if four candidates should be offered a place. There were more decisions to make than in experiment 1, therefore it is possible that this task induced a greater cognitive load than the preceding two. An alternative explanation for this result is that participants were focussing on the audio channel, as it was degraded, and not paying as much attention to the video channel. The results of Simons et al. (1999) showed that HR decelerated when watching moving images, thus video in this experiment may have been glanced at but not focussed on, which may have caused HR to increase.

From table 9 it seems that experiment 2 had the most perceptual strain because it had the highest SC and the lowest HR and BVP. This may be because the task was not engaging, therefore the poor quality was focused on. In experiments 1 and 3 the task was engaging so less attention may have been paid to the quality degradations. In addition, experiment 2 was a shorter experiment, thus it may be that changes in the mean levels of physiological signals are easier to detect in shorter experiments. Finally, experiment 2 concerned solely audio degradations, which are more stressful than video degradations and may be more stressful than audio degradations with the video channel added because the video may help participants to decipher what participants are saying.

| Experiment | SC | HR | BVP |
|---|---|---|---|
| 1 | 1.23 | -0.85 | 0.01 |
| 2 | 3.85 | -2 | -7.16 |
| 3 | 0.9 | 0.64 | 0.77 |

**Table 9: Overall means of physiological signals in experiments 1-3**

## 5.4 Chapter conclusions

This chapter has presented the results of three experiments, which have utilised the 3-factor framework (task performance, user satisfaction and user cost) to measure the impact of media quality degradations on users. Experiment 1 used a recorded interview to

investigate the impact of 2 levels of video frame rate. Experiment 2 utilised a passive listening task to investigate the impact of 6 audio degradations on participants. Finally, experiment 3 used a recorded interview to measure the impact of four audio degradations that were also investigated in experiment 2.

The results from experiments 1 and 2 show that physiological responses to media quality degradations can be detected. Experiment 1 showed some significant differences in SC between frame rates when seeing interviews at 5-25-5fps. Experiment 2 showed that echo and loud volume differences between speakers cause significant increases in SC in comparison to a reference condition and in HR the 5% packet loss condition was significantly less straining than the reference condition. There were no significant differences in the physiological data in experiment 3 – the possible reasons for this are discussed in section 5.3.8.1.

The subjective results of experiment 1 showed that, most fundamentally, fewer than 15% of participants did not notice that the frame rate had changed, whereas differences were registered in the SC of group 1. There was a close association between the physiological and subjective results in experiment 2, which was expected because the task only required participants to rate the quality. In experiment 3 the subjective and physiological results cannot be compared, as there were no significant results in the physiological signals.

The direction of physiological responses can give some information about the task being performed and whether it involves mainly perceptual or cognitive functions. The results from experiment 1 showed a fractionation between SC and HR in the second interviews, which indicates that the first interview was cognitive in nature, whereas the second was perceptual. This may have been due to participants being more used to the task and less anxious about the

experiment. The results from experiment 2 showed that the task was primarily perceptual, whereas experiment 3 was cognitive. This was attributed to four decisions being made about interview candidates, as opposed to the decision between two candidates in experiment 1. In addition, the fact that a task baseline as opposed to a resting baseline was used may have influenced results.

The results from these experiments also show the additional information that physiological data can offer, which remains untapped by subjective assessment, for example the impact of presentation of degradations in SC and BVP (experiment 2) and the differences between the frame rates in the SC of group 1 in experiment 1. Moreover, measuring task performance (experiment 1) can offer illuminating information, for example in experiment 1, group 1 participants favoured the first candidate whereas group 2 participants favoured the second. Therefore, these results provide support for the 3-factor approach to media quality evaluation. Having investigated physiological responses in passive tasks, two experiments were performed to investigate whether the impact of media quality degradations could be detected in interactive tasks.

# Chapter 6 The Impact of Audio and Video Degradations in Interactive Tasks

## Chapter Aims

This chapter describes two experiments that utilised interactive tasks to examine the impact of video frame rate and audio packet loss. It is important to determine the viability of using physiological measures in more ecologically valid scenarios, as this will determine whether such measurements can be used in field studies.

## 6.1 Experiment 4: Investigating the impact of low and high video frame rates in a real-time interactive interviewing task

### *6.1.1 Introduction*

Having investigated the impact of low and high video frame rate in a passive task, the next logical step was to examine the same two frame rates in an interactive task to determine if they would have a similar effect or if the influence of an interactive task would override them. This was a joint experiment with Glasgow University as part of the ETNA project (see section 2.3.4) and took place in the Psychology department at Glasgow University. The project group jointly designed the experiment. The author of this thesis gathered and analysed the physiological data and administered the media quality questionnaires after each condition. The co-experimenter from Glasgow University gathered and analysed the eye-tracking data, analysed the subjective responses and analysed the length of time it took participants to complete the task.

### *6.1.2 Design*

Each participant played the role of a bank employee. They had to interview two applicants (one male and one female) for a loan, collect personal information from them and decide if they should be offered

a loan. This task was selected because it would encourage participants to focus on the video channel to judge whether the applicant was trustworthy. Each participant saw one interview at 5fps and one at 25fps. The audio and video were sent over a dedicated link between two computers to ensure that network traffic did not affect the quality received. The applicants for the loan were situated in the same building as the participants, however the participants were told they were in another venue in the city. The two applicants were actors recruited from Glasgow University's Drama department. They were briefed in the responses they should give to the questions the participants would ask. The order of presentation of frame rate and applicant interviewed were randomised. Thirteen participants saw 5fps followed by 25fps, whereas 11 participants saw the opposite. Eleven participants interviewed the female candidate then the male candidate, whereas 13 participants saw the opposite.

The interviewers were allowed to finish the interview when they felt they had enough information, yet were informed that they should aim for around five minutes. The audio quality was good and did not vary. The co-experimenter collected eye-tracking data from participants. The author of the research reported in this thesis stayed in the room with the participants, behind a partition, to monitor their physiological responses and to administer questionnaires.

### *6.1.3 Materials*

#### 6.1.3.1 Experiment materials

Participants watched the interviews on a PC and wore a headset with a microphone in order to hear the audio and to speak with the applicant. They also had their eye movements tracked, which involved them sitting in a chair with a headrest to encourage them to keep their heads still. The interviews were conducted over a dedicated link to ensure that network traffic did not interfere with the quality received.

#### 6.1.3.2 Subjective assessment materials

After each interview, video and audio adequacy and quality scales were administered (see Appendix C for an example). In addition the participants were asked if the applicant should be offered a loan[30]. Table 10 shows the 3-factor approach used in experiment 4.

| Element in 3-factor framework | Measure |
|---|---|
| Task performance | Social judgement on applicant<br>Time taken to complete interview |
| User satisfaction | Post-hoc questions on audio and video quality and adequacy |
| User cost | Physiological measures |

**Table 10: The 3-factor approach used in experiment 4**

### *6.1.4 Participants*

Twenty-four participants took part in this experiment. They were recruited from Glasgow University and were paid for their time. There were 5 males and 19 females.

### *6.1.5 Procedure*

Participants entered the room and were informed of the nature of the experiment and what their task was. They were informed that they were free to withdraw from the experiment at any time and were given a printout describing the task and the information they were required to find out from the applicant, such as what the loan was for. Once the baseline physiological signals had been measured, the calibration of the eye-tracker took place. Participants then interviewed the first applicant, after which they completed the audio and video quality and adequacy scales and made a decision about whether the applicant should be offered the loan. The same procedure occurred for the second interview.

### *6.1.6 Hypotheses*

1. There will be a significant impact of video frame rate.
2. 5fps will cause more perceptual strain than 25fps.

---

[30] This data was not analysed.

3. Participants will not subjectively notice the difference between the two frame rates, as they will be engaged in the task.

### *6.1.7 Results*

#### 6.1.7.1 Physiological results

Separate ANOVAs were performed for SC, HR and BVP. Frame rate was a within-subjects variable (2 levels: 5 and 25fps) and there were three between subjects variables:

- gender
- frame rate order (2 levels: frame rate order 1 was 5fps then 25fps; frame rate order 2 was 25fps then 5fps)
- actor order (2 levels: order 1 was female then male applicant; order 2 was male then female applicant).

SC

There were no outliers. No significant differences were found in the ANOVA. The mean SC and standard deviation of the signal can be seen in figures 46 and 47 respectively.

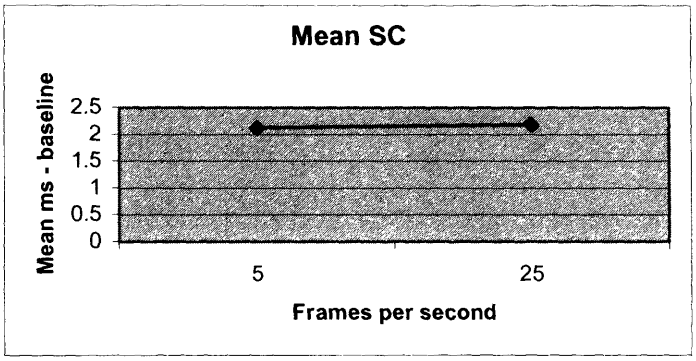



**Figure 46: Mean SC in experiment 4**

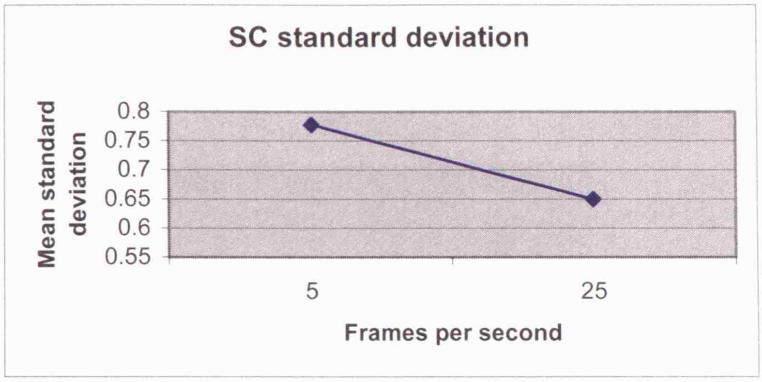



**Figure 47: Mean SC standard deviation in experiment 4**

HR

The data were cleaned to remove outliers[31]. There was a significant main effect of gender ($F_{(1,17)}$ = 5.707, p=0.029) and significant interactions between frame rate and frame rate order ($F_{(1,17)}$ = 10.764, p=0.004) and frame rate, frame rate order and actor order ($F_{(1,17)}$ = 5.911, p=0.026).

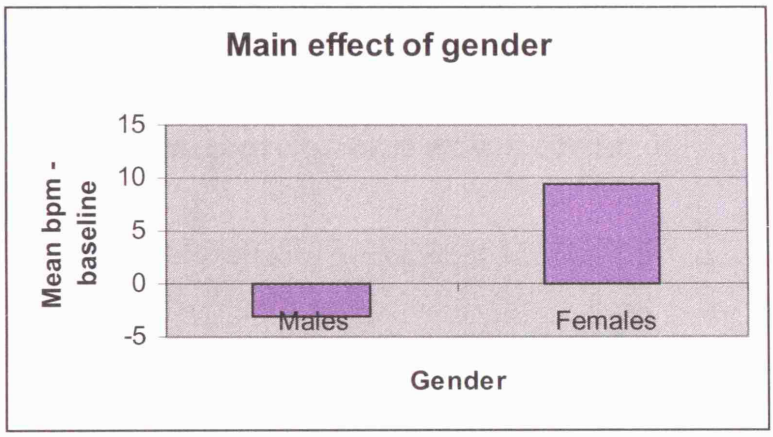



**Figure 48: Mean HR for males and females in experiment 4**

Figure 48 shows that males had significantly lower HR than females during the experiment. However, there were 5 males compared with 18 females, therefore no meaningful results can be obtained from this.

[31] The data from one participant (24) was removed, which left 5 males and 18 females. 12 participants had frame rate order 1 and 11 had frame rate order 2. 11 participants had actor order 1 and 12 had actor order 2.

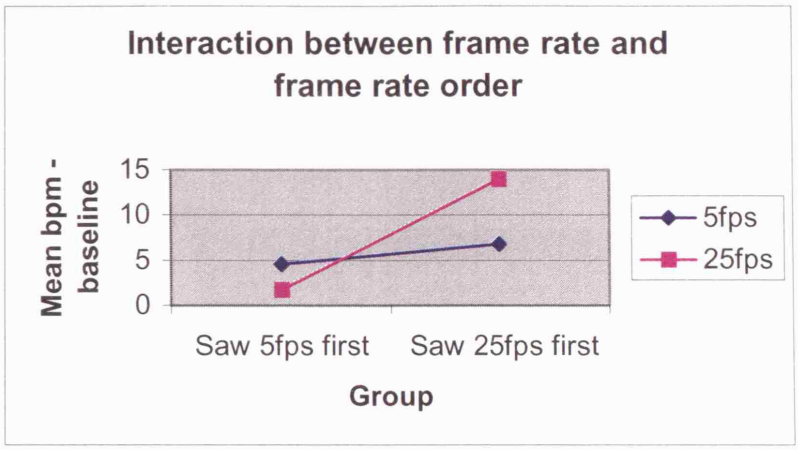



**Figure 49: Interaction between frame rate and frame rate order for HR in experiment 4**

Figure 49 shows that in the group who saw 5fps first, 5fps caused more perceptual strain than 25fps. However, for the group who saw 25fps first, 25fps caused more perceptual strain than 5fps. This shows that the frame rate seen first caused more perceptual strain and indicates that responses to the task were drowning out responses to the quality because participants were likely to be more anxious at the beginning of the experiment.

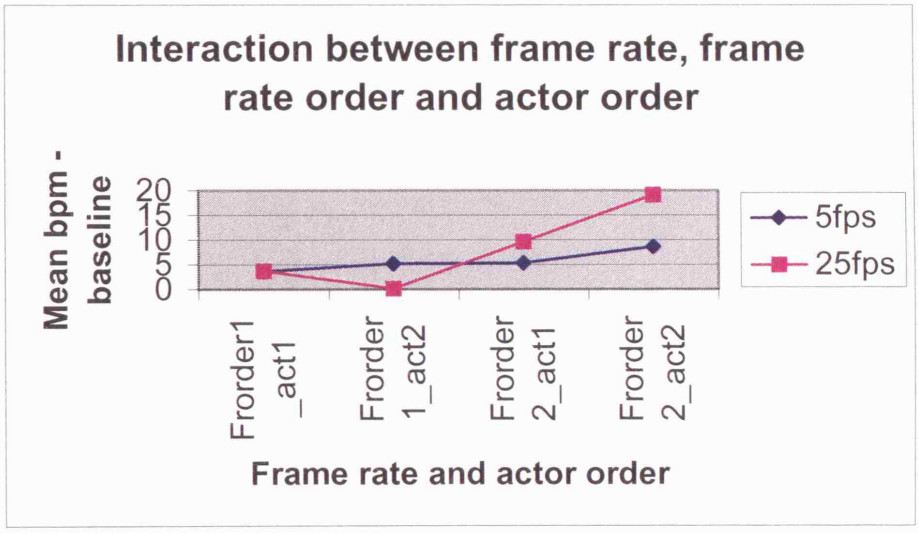



**Figure 50: Interaction between frame rate, frame rate order and actor order for HR in experiment 4**

Figure 50 shows that the only point when 5fps is more straining that 25fps is during frame rate order 1 (saw 5fps first) when the male applicant is interviewed.

The mean HR and standard deviation of the signal can be seen in figures 51 and 52 respectively.

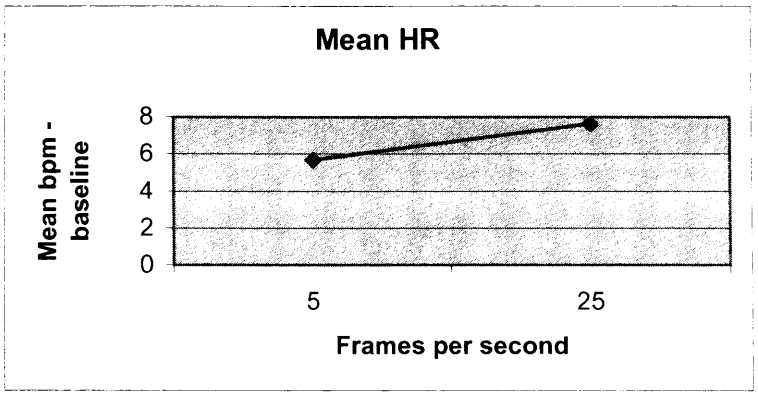



**Figure 51: Mean HR in experiment 4**

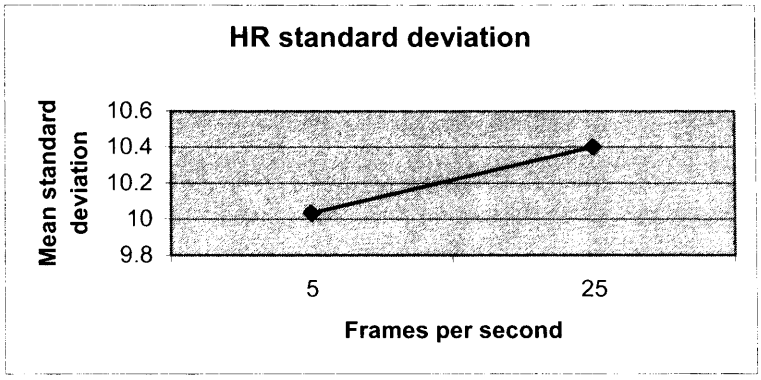



**Figure 52: Mean HR standard deviation in experiment 4**

BVP

The data were cleaned to remove outliers[32]. There was a significant main effect of actor order ($F_{(1,16)} = 4.507$, $p=0.05$).

[32] There were two outliers (10 & 19). This left 5 males and 17 females. 12 participants had frame rate order 1 and 10 had frame rate order 2. 10 participants had actor order 1 and 12 had actor order 2.

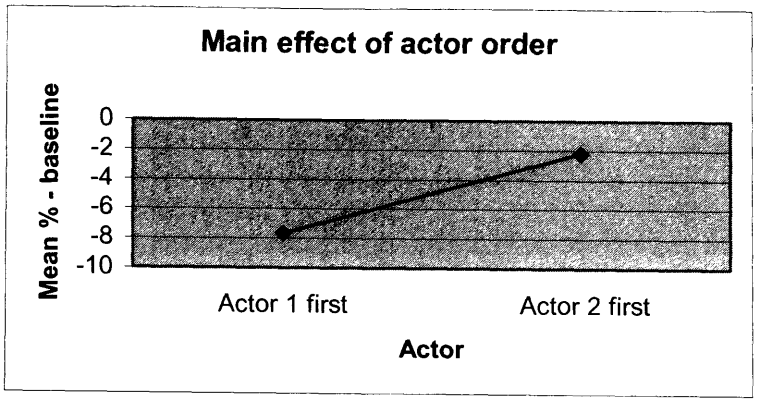



**Figure 53: Mean BVP for each actor in experiment 4**

Figure 53 shows that participants who saw the female candidate first had more perceptual strain than those who saw the male candidate first.

The mean BVP and standard deviation of the signal can be seen in figures 54 and 55 respectively.

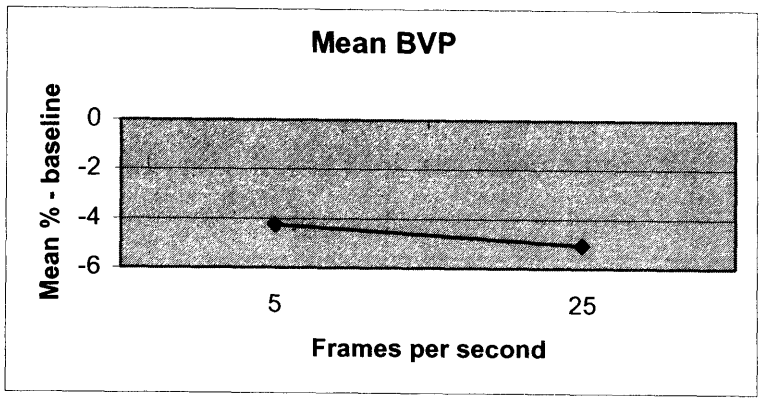



**Figure 54: Mean BVP in experiment 4**

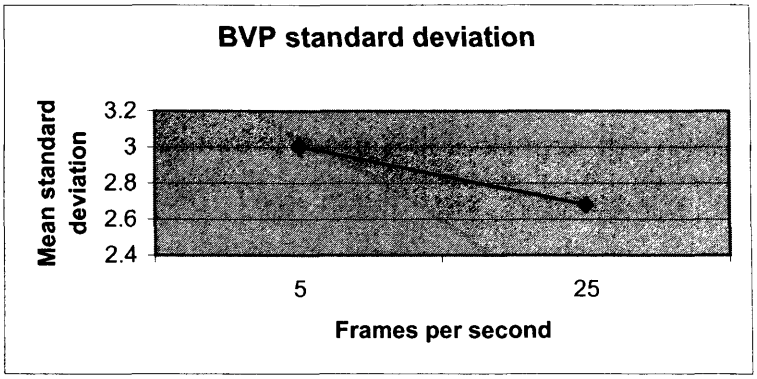



**Figure 55: Mean BVP standard deviation in experiment 4**

### 6.1.7.2 Subjective results

These data were analysed by the co-experimenter and will be summarised here. Audio quality and adequacy ratings showed no significant differences. This was expected because the audio quality was good and did not change through the interviews. There were no significant results for video adequacy. However, video quality had a significant result: high frame rate video was rated higher than low frame rate video ($F(1,17)=4.421$, $p=0.0491$). Table 11 shows that when 25fps is seen first, it and 5fps are rated higher than when 5fps is seen first.

| | 5fps first | | 25fps first | |
|---|---|---|---|---|
| | **5fps** | **25fps** | **5fps** | **25fps** |
| Audio quality | 78.77 | 80.15 | 73.64 | 77.18 |
| Audio adequacy | 82.62 | 84.23 | 68.27 | 79.82 |
| Video quality | 65.15 | 74.08 | 73.18 | 76 |
| Video adequacy | 65.92 | 70.23 | 58.46 | 59.27 |

**Table 11: Mean subjective ratings in experiment 4**

### 6.1.7.3 Time taken to complete task

These data were analysed by the co-experimenter. This result uses all participants with the exception of participant 1, as she spent an extremely long time on interview 1 because the interviewee was trying to get her to ask relevant questions. The time taken to complete the task was significantly different ($F(3,21)=7.975$, $p=0.0102$) with participants taking longer to complete the task in the 5fps condition.

#### 6.1.7.4 Eye tracking data

These data were analysed by the co-experimenter. There was a tendency towards spending more time looking at the video window in the 25fps condition. This result was approaching significance ($F(1,17)=3.744$, $p=0.0698$). A simple main effects analysis shows that this is almost entirely accounted for by participants who saw 5fps first ($F(1,17)=7.604$, $p=0.0135$).

The average length of the glances at the video window, although not significant, showed a tendency towards longer glances being directed towards an area picking out the applicant's head in the 25fps condition for fixations over 400ms ($F(1,18)=3.210$, $p=0.0900$). For glances at the larger area of the video window, this result was further subdued ($F(1,18)=2.811$, $p=0.1109$). However, in both analyses a further simple main effects test showed that this was again almost completely accounted for by the subjects who received 5fps first.

An analysis of average gaze length, which included all fixations on the video window over 100ms showed the same tendency ($F(1,19)=3.239$, $p=0.0887$) and repeated the conclusion that the differences were almost entirely based on those receiving 5fps first (simple main effects ($F(1,18)=6.132$, $p=0.0234$)).

### *6.1.8 Discussion of results*

#### 6.1.8.1 Physiological results

There were no significant differences in the SC signal. The HR signal showed that the frame rate seen first caused more perceptual strain. In addition, there was one point where 5fps had more perceptual strain than 25fps, which was when the male candidate was interviewed by the participants who saw 5fps first. BVP showed that seeing the female candidate first caused more perceptual strain than seeing the male candidate first.

In all 3 signals, there was no significant effect of frame rate. This may be explained by the stressful nature of the task, which could have drowned out any differences in responses due to the quality. This is supported by this experiment having the highest mean HR, second highest SC and second lowest BVP out of the first 5 experiments reported in this thesis (see table 16). Support for this also comes from the order result shown in HR, where whichever frame rate was seen first had more perceptual strain. The participants were novice users of the technology and novices at the task, thus this result may indicate them getting used to the technology or the task.

The BVP signal shows that interviewing the female candidate first caused significantly more perceptual strain than interviewing the male candidate first. Both the female candidate and the male candidate were actors, so the differences cannot be due to differences in their responses. Therefore, it may be an effect of the majority of the participants being female and interviewing a female.

In this experiment, SC and HR were above baseline, whereas BVP was below baseline. Thus, there is not the same fractionation between SC and HR as was seen in experiments 1 and 2. This mirrors the result from experiment 3 and the same conclusion regarding the task can be made here: that the task, which was unfamiliar to participants and was stressful, resulted in the task being primarily cognitive, as opposed to perceptual.

#### 6.1.8.2 Subjective results

Subjectively the difference between low and high video frame rates was noticed and this was a significant result. In addition, when 25fps was seen first, it and 5fps were rated higher than when 5fps was seen first. Therefore, experiencing good quality first seems to give participants a better overall impression of that quality and subsequent quality levels. The implication from this is that in such interviewing tasks, it is important to give good quality to begin with to increase user satisfaction.

The eye tracking data did not produce any significant results. This is not surprising as the task encouraged participants to focus on the video to generate an impression of the candidate and to determine if they should be offered a loan, therefore it is likely that even when the video was poor, they focussed on it. The physiological cost of doing so did not come out of the results, though it may have increased user cost.

### *6.1.9 Addressing the hypotheses*

1. There will be a significant impact of video frame rate. This hypothesis was not supported. This may be because responses to the task drowned out responses to the quality.
2. 5fps will cause more perceptual strain than 25fps. This hypothesis was not supported.
3. Participants will not subjectively notice the difference between the two interviews, as they will be engaged in the task. This hypothesis was not supported because subjective ratings did reveal significant differences to the two frame rates.

### *6.1.10 Limitations*

There are some limitations of this experiment. Firstly, the task was stressful and participants were concerned that they had to remember all relevant questions to ask because (due to the eye tracker) they could not read them on the question sheet. As a result of having to keep their heads and hands still due to the eye tracker and the physiological sensors respectively, some participants reported feeling trapped, which may have interfered with the results. This may have heightened responses due to participants feeling uncomfortable or people may have paid less attention to the task and been focussed on completing it as quickly as possible. There was no 16fps to allow the participants to acclimatise to the quality and task. The gender of participants was not balanced and finally, the participants were not experienced interviewers and the task was novel for them, which means that responses to the task may have drowned out responses to the quality.

### *6.1.11 Conclusions*

This experiment showed that video frame rate does not have a significant impact upon physiological signals in a stressful, interactive task that participants are not familiar with. In addition, there was a significant order effect in HR, which showed that whichever frame rate was seen first was more straining. This indicates that responses to the task were drowning out responses to the quality. Therefore, it appears that in a stressful task like this, using basic analysis on physiological signals is not effective.

The lack of fractionation between SC and HR indicates that this task was cognitive, which was expected, as it required participants to actively interview and make a judgement on the candidate, as opposed to passively watching interviews. Finally, additional data gathered as part of the 3-factor approach, showed that participants subjectively registered the difference between the frame rates and that high frame rates allow participants to complete their task quicker. Therefore, in such tasks there may be minimal benefit of measuring the means of physiological signals to investigate the impact of media quality degradations.

## 6.2 Experiment 5: Investigating the impact of audio and video degradations in an interactive task

### *6.2.1 Introduction*

Having found that video frame rate and audio degradations have an impact on participants in passive perceptual tasks in experiments 1 and 2, the next step for the research reported in this thesis was to induce degradations on both the audio and video channels in the same interview to investigate the interactive influence of one upon the other. This was a joint experiment between UCL and Glasgow University as part of the ETNA project (see section 2.3.4). Instead of having naïve participants performing an unfamiliar task, as in experiment 4, it was decided to use experienced participants performing a familiar task to see if responses to the quality would be discernable from responses to the task. The experiment was designed and planned by the project group. The author of this thesis gathered and analysed the physiological data at UCL. A co-experimenter at UCL (Anna Watson) set-up and ran the experiment with a co-experimenter at Glasgow University (Rachel McEwan) and gathered and analysed the subjective data.

### *6.2.2 Design*

This experiment was performed in the Computer Science department of UCL. Thirteen experienced undergraduate applicant interviewers were the participants in this experiment. They were all lecturers at UCL, which increased the ecological validity of the experiment. They conducted four interactive interviews with candidates (situated at Glasgow University) for the Computing degree course at UCL. The interviews were conducted over the network and in real-time. The interviewers only saw a picture of the candidate on their screen, as occurred in experiments 1, 3 and 4.

In order to induce the packet loss onto the audio stream, a reflector (see section 5.3.3.1) was used at UCL. The encoding scheme was DVI and repaired packet loss was used. The video frame rate was

set at the required level by the co-experimenter at Glasgow University before the interviewers began.

Participants were told that the candidates were real applicants and that the interviews were to determine the feasibility of utilising MMC as an interviewing tool. However, the candidates were actors who had been given training on their role by the co-experimenter. This measure was taken to ensure consistency of responses. They all played the role of a strong candidate to minimise the risk of any differences in results being due to the candidates as opposed to the quality. There was no limit set on the time that the interview could last for. There were four conditions and the order of presentation to participants was randomised.

1. Good video quality (~25fps) with good audio quality (~0% packet loss) (GVGA)
2. Good video quality (~25fps) with bad audio quality (~15% packet loss) (GVBA)
3. Bad video quality (~5fps) with good audio quality (~0% packet loss) (BVGA)
4. Bad video quality (~5fps) with bad audio quality (~15% packet loss) (BVBA)

The frame rate and packet loss levels are all stated as being around the level, as the experiment was conducted over the network, therefore there was traffic.

### *6.2.3 Materials*

#### 6.2.3.1 Experiment materials

During the baseline gathering session, the participants were given the application forms of the four interviewees to read and familiarise themselves with. These were created by the co-experimenter and also given to the actors, so that they could familiarise themselves with the role they were playing.

#### 6.2.3.2 Subjective assessment materials

Questionnaires were administered after each interview on the quality and adequacy of the audio and video channels (see Appendix C for an example of the rating scale used), as were candidate assessment forms (see appendix B). The video and audio quality and adequacy scales were on a scale of 1-100, where 1 was very poor quality/totally inadequate and 100 was very good quality/completely adequate. Participants were also given the opportunity to comment at the end of the experiment on their impression of the experience. Table 12 shows the 3-factor approach used in this experiment.

| Element in 3-factor framework | Measure |
|---|---|
| Task performance | Judgement on applicant and technology<br>Time |
| User satisfaction | Post-hoc questions on audio and video quality and adequacy |
| User cost | Physiological measures |

**Table 12: The 3-factor approach used in experiment 5**

### *6.2.4 Participants*

There were 13 participants. They were paid £20 in book tokens for their time. Twelve participants were male and one was female. This gender imbalance reflected the make-up of the Computer Science department at UCL when this experiment was conducted.

### *6.2.5 Procedure*

Participants were informed that they would be performing four interviews with four candidates for the undergraduate computing degree course at UCL. They were told that the aim of the experiment was to determine the feasibility of using MMC to perform undergraduate admissions interviews in the future, as this would save candidates who live far away from having to travel to London. They were informed that they were free to withdraw from the experiment at any time.

Once the physiological baseline measuring session was finished, the experimenter at UCL checked (via a phone call) with the co-

experimenter at Glasgow University that the session was ready to proceed and then introduced the first candidate to the interviewer. The physiological recording began when the interviewer started to speak to the candidate. After the interview was completed, the physiological recording was stopped and the questionnaire and candidate assessment form were administered. This procedure was identical for the remaining three interviews.

### *6.2.6 Hypotheses*

1. There will be a significant impact of media quality degradations on physiological signals.
2. The interview with bad audio and video quality will be the most straining and the interview with good audio and video quality will be the least straining. The interview with bad audio quality and good video quality will cause more perceptual strain than that with good audio quality and poor video quality, as audio is generally regarded as being more important than video quality (Sasse et al., 1994). Therefore, the order (from most to least perceptually straining) will be: BVBA, GVBA, BVGA then GVGA.
3. Participants will not subjectively notice changes in the audio and video quality, as they will be engaged in the task.

### *6.2.7 Results*

The objective statistics gathered showed that eleven participants received the conditions that were intended (0% or 15% packet loss; 5fps or >20fps). However, the first two participants experienced very poor audio quality. The 'no loss' condition was 15% packet loss for these participants and the level of loss in the high loss condition was 25-30%. Therefore, these participants were excluded from the analysis, which left a total of eleven participants.

### 6.2.7.1 Physiological results

A repeated measures ANOVA with the independent variable degradation with four levels (BVBA, GVGA, BVGA, GVBA) was performed on each signal[33].

SC

Data was only available for 10 participants, due to a problem with the recording equipment for participant number 1. There were no outliers. There were no significant differences in the data. The mean SC and standard deviation for this signal can be seen in figures 56 and 57 respectively.

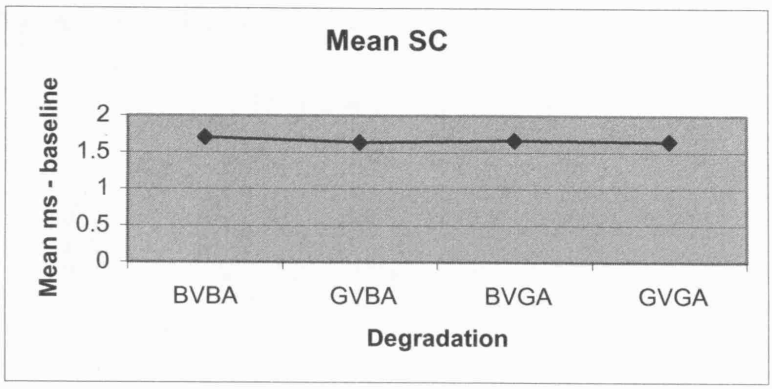



**Figure 56: Mean SC in experiment 5**

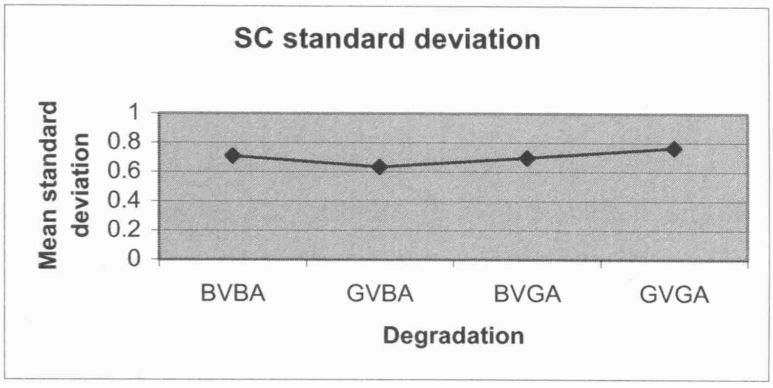



**Figure 57: Mean SC standard deviation in experiment 5**

HR

---

[33] As there was only one female, gender could not be included as a between-subjects variable.

The data were cleaned for outliers[34]. There were no significant differences in the data. The mean HR and standard deviation for this signal can be seen in figures 58 and 59 respectively.

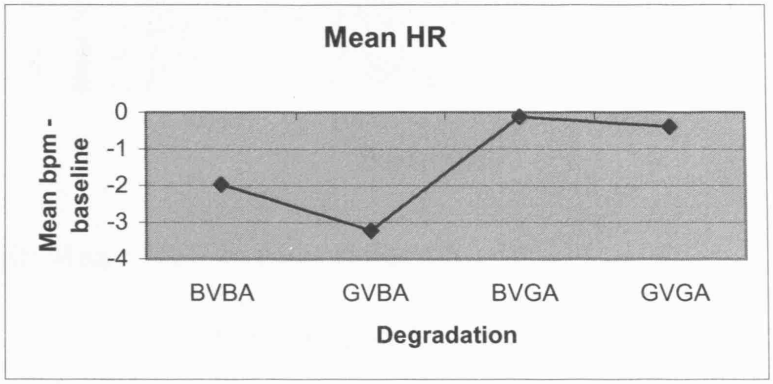



**Figure 58: Mean HR in experiment 5**

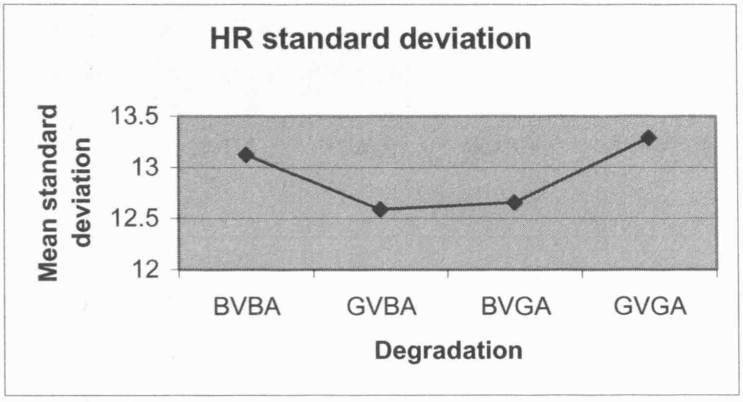



**Figure 59: Mean HR standard deviations in experiment 5**

BVP

The data were cleaned for outliers[35]. There was a significant main effect of degradation ($F_{(3,24)}$ = 4.469, p=0.013). Post-hoc pairwise comparisons did not reveal any significant differences between the conditions. This is likely to be due to the conservative nature of the Bonferroni correction (see section 4.3.5). The mean BVP and standard deviation for this signal can be seen in figures 60 and 61 respectively.

---

[34] Two participants (10 & 11) were removed.
[35] Two participants (1 & 9) were removed.

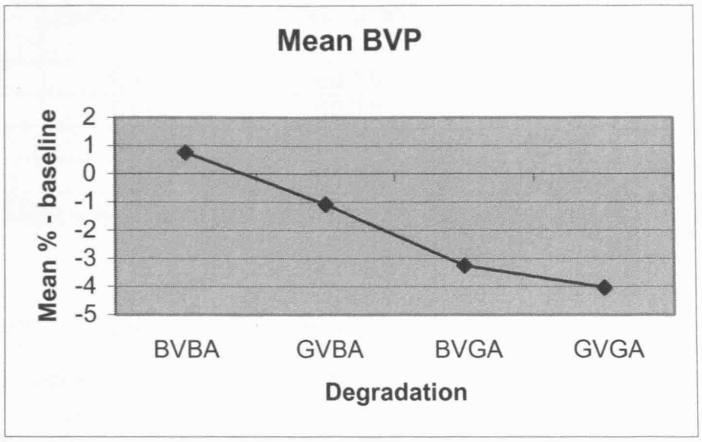



**Figure 60: Mean BVP in experiment 5**

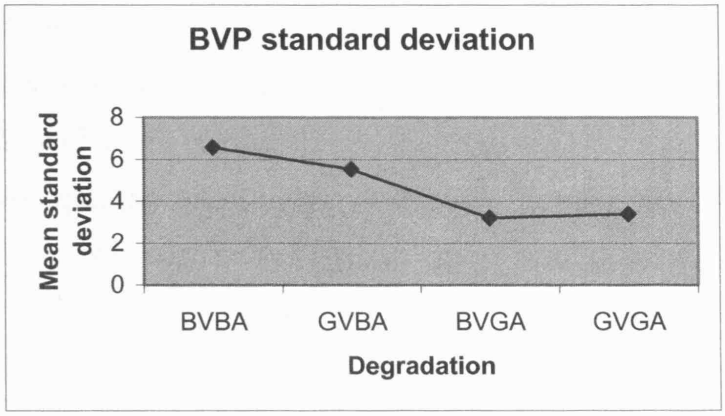



**Figure 61: Mean BVP standard deviations in experiment 5**

### 6.2.7.2 Subjective results

The subjective results were analysed by Anna Watson (UCL) and Lucy Smallwood (Glasgow University). Table 13 shows the mean subjective ratings given to each condition by participants. A two way within-groups ANOVA was performed using the data from each rating scale. A main effect was found of audio in the ratings for audio quality ($F=26.033$, $p<0.001$). No interactions were found. There was a main effect of video frame rate ($F=9.891$, $p<0.05$). This appears to be mainly accounted for by the video quality ratings at low levels of audio quality although there was a trend towards a difference in video quality ratings at high quality audio. No interactions were found. There was a main effect of audio in the ratings for audio adequacy ($F=13.168$, $p<0.01$). No interactions were found. Finally, no main effects or interactions were found for the video adequacy ratings.

| Condition | Audio quality | Video quality | Audio adequacy | Video adequacy |
|---|---|---|---|---|
| GVGA | 63.91 | 66.55 | 74 | 67.64 |
| BVGA | 64.64 | 52.73 | 74.91 | 62.09 |
| GVBA | 40.64 | 67.36 | 55.45 | 65 |
| BVBA | 38.27 | 51.73 | 53.46 | 59.27 |

**Table 13: Mean subjective ratings in experiment 5**

Comparisons between the means indicate that the following differences are significant.

- Audio quality: good video and poor audio quality (GVBA) is rated significantly lower than good video and audio quality (GVGA) ($p<0.01$). Poor video and audio quality (BVBA) is rated significantly lower than poor video and good audio quality (BVGA) ($p>0.01$).
- Video quality: good video and audio quality (GVGA) is rated significantly higher than poor video and good audio quality (BVGA) ($p<0.05$).
- Audio adequacy: good video and poor audio quality (GVBA) is rated significantly lower than good video and audio quality (GVGA) ($p<0.05$). Poor video and good audio quality (BVGA) is rated significantly higher than poor video and audio quality (BVBA) ($p<0.05$).

At the end of the experiment, the interviewers were asked to give their impressions of the experience. Most felt that the technology made such remote interview feasible, however they would prefer to interview face-to-face: *"You can tell a lot from a handshake".* In addition, none of the candidates were viewed as being borderline in their application. It is possible that the quality may have had more of an impact in such a situation. See table 14 for comments made by interviewers on the audio and video quality received.

| Condition | Comments |
|---|---|
| Audio in good video and poor audio condition | *"Handover was difficult."*<br>*"The only problems are with things like conversational turn-taking and interjections."* |
| Video in good video and poor audio quality condition | *"Again, the slow updating lends to a feeling of detachment."*<br>*"Quite small, definitely requires full attention to understand body language."*<br>*"Improvement mainly due to my own adaptation."*<br>*"No 'sense of person'."*<br>*"I think you get used to it and begin to accommodate."*<br>*"Lack of eye contact was not as off-putting as I would have expected...In fact, it made me feel more comfortable!"* |
| Audio in good video and audio quality condition | *"Some drop outs at critical moments."*<br>*Sound kept breaking up, leading to a lot of repetition. Broke up flow of conversation and was disruptive."*<br>*"Some echo to start but I almost didn't notice later on."* |
| Video in good video and audio quality condition | *"Still felt detached."*<br>*"No engagement with the interviewee."* |
| Audio in poor video and audio quality condition | *"Sometimes quality is only just enough to understand what is being said."*<br>*"I think you get used to it – but still the odd overlap and interruption – hard to give cues to cut sentences."*<br>*"Loss of the audio in the form of 'gaps' forces me to use slow speech and occasionally repeat requests."* |
| Video in poor video and audio quality condition: | *"Total loss of synch with the audio at some points. I would have preferred to just use the telephone."*<br>*"Very difficult to use my body language to help calm his nerves."*<br>*"Picture size and frame rate too small / slow for 'involvement'."*<br>*"It was fine for the job. Beginning to feel more comfortable."* (during interview 3) |
| Audio in poor video and good audio quality condition: | *"Only just possible to follow the conversation. Disrupted flow and discouraged conversation. Not adequate for doing a proper job."*<br>*"Broke up too much. Hard to get continuity – it resulted in too many interruptions and gaps in my understanding (I didn't want to keep asking him to repeat)."* |
| Video in poor video and good audio quality condition: | *"Very slow frame rate, couldn't distinguish facial expression."*<br>*"Very poor – broken images continuously – he was moving about too much. It distracted me from the task."*<br>*"Slow update, but just about good enough. However, feedback from facial expressions was lost. Would have been OK if sound was better."* |

**Table 14: Comments made by interviewers on the audio and video quality experienced in experiment 5.**

### 6.2.8 *Time to complete task*

The author of this thesis analysed the time participants took to complete the task. A repeated measures ANOVA with the within-subjects factor of time did not show any significant differences.

| Condition | Time (minutes) |
|---|---|
| BVGA | 12.4 |
| GVGA | 11 |
| BVBA | 11.54 |
| GVBA | 11.18 |

**Table 15: Mean time to complete the interviews experiment 5**

Table 15 shows that the interviews with poor video and good audio quality took the longest to complete and the interviews with good video and audio quality were the fastest to be completed. Therefore, better quality allowed participants to complete the interviews quicker, whereas video with a low frame rate and good audio takes the longest amount of time. The interviews with poor video and audio quality took the third longest to complete, which may be because it makes communication more difficult. The finding that the condition with the poor video frame rate took the longest to complete may mean that participants struggled to get a good impression of the candidates due to the poor video. However, this condition was rated as having the highest audio quality and being most adequate with regard to audio, therefore it is possible that participants spoke more in this condition because they found it relatively easy to communicate in.

### 6.2.9 *Discussion of results*

#### 6.2.9.1 Physiological results

The lack of significant impact on the physiological signals can be explained by the long and stressful task that the participants had to perform. Therefore, it appears that, as in experiment 4, responses to the task drowned out responses to the quality. Despite this, it was important to use the physiological signals to determine if any differences would be discernable in an experiment like this.

#### 6.2.9.2 Subjective results

In the audio quality and adequacy scales, the difference between poor audio and video quality (BVBA) and the condition rated as best (BVGA) was significant. This shows that when rating audio where the video quality is poor, the difference between good and bad audio quality is registered. In addition, the differences between good video and poor audio (GVBA) and good video and good audio (GVGA) were significant. Thus, when video quality is good the difference between good and poor audio quality is noticeable.

In the video quality scales, the difference between good and bad video was noticed when the audio quality was poor and not when it was good. Therefore, good audio quality may in some way compensate for a poor video quality.

### *6.2.10 Addressing the hypotheses*

1. There will be a significant impact of media quality degradations on the physiological signals. This hypothesis was not supported.
2. The interview with bad audio and video quality will be the most straining and the interview with god audio and video quality will be the least straining. The interview with bad audio quality and good video quality will cause more perceptual strain than that with good audio quality and poor video quality, as audio is generally regarded as being more important than video quality (Sasse et al., 1994). Therefore, the order (from most to least perceptually straining) will be: BVBA, GVBA, BVGA then GVGA. There were no significant differences in the physiological signals, therefore no comment can be made on this hypothesis.
3. Participants will not subjectively notice changes in the audio and video quality, as they will be engaged in the task. This hypothesis was not supported, as significant differences in subjective ratings were obtained.

### 6.2.11 *Limitations of experiment*

There are two limitations to this experiment. Firstly, on average the length of the task was an hour. This was a long time for participants to interview for and they may have become frustrated or fatigued. Secondly, the task was stressful, so responses to the task seemed to drown out responses to the quality.

### 6.2.12 *Conclusions*

These results show that the task used in this experiment, which participants are familiar with is mainly perceptual, however there was not a significant impact of degradation on physiological responses, except on BVP (yet, post-hoc tests did not reveal any significant differences between conditions). This may have been due the small number of participants or the stressful nature of the task. However, subjective data and data on the time taken to complete the task illustrated the more complete picture on user experience that can be gained by the 3-factor approach to evaluation and showed that the interview with best quality was completed quickest and interviews with poor audio and video quality were rated as being of the worst quality.

## 6.3 Chapter conclusions

This chapter has described the results of two experiments, both of which used interactive tasks, however experiment 4 used participants who were naïve to the task and experiment 5 used participants who were experienced in the task. Table 16 shows that experiment 4 had the highest SC, HR and lowest BVP, thus it can be concluded that experiment 4 caused more perceptual strain than experiment 5. This is not surprising because the participants in experiment 4 were naïve interviewers, whereas the participants in experiment 5 were experienced interviewers. In addition, the participants in experiment 4 had less time to get used to the task, as their two interviews were around 5 minutes in length whereas the four interviews in experiment 5 were around 11 minutes long.

| Experiment | SC | HR | BVP |
|---|---|---|---|
| 1 | 1.23 | -0.85 | 0.01 |
| 2 | 3.85 | -2 | -7.16 |
| 3 | 0.9 | 0.64 | 0.77 |
| 4 | 2.16 | 6.67 | -4.65 |
| 5 | 1.67 | -1.42 | -1.89 |

**Table 16: Overall means of physiological signals in experiments 1 to 5**

Another observation is that in experiment 4 there is not a fractionation between SC and HR, whereas in experiment 5 there is. This indicates that the task in experiment 4 is cognitive, due to the participants being inexperienced in it and the task in experiment 5 is perceptual, as the participants were experienced interviewers.

There was no significant impact of frame rate in experiment 4. This is likely to be because responses to the task drowned out responses to the quality. This hypothesis is supported by the finding that the frame rate seen induced a higher level of HR. The order of actor seen also had a significant impact. Experiment 5, which was the longer and more complex of the two interactive experiments showed no significant differences in physiological responses (with the exception of BVP) and this may again be due to the stressful nature of the task and the large number of variables in operation.

# Chapter 7 Conclusions and Recommendations

## Chapter Aims

This chapter presents and discusses the conclusions and implications of the research reported in this thesis. It describes how HCI researchers and network providers can use the findings and discusses potential future research. This thesis has undertaken exploratory research in laboratory-based settings and contributed substantive and methodological findings with regard to media quality degradations in MMC. A novel measure of user cost as part of a 3-factor evaluation framework has shown significant and meaningful results, from which HCI researchers and network providers can benefit.

## 7.1 The research problem restated

Computer workstations and high bandwidth networks can deliver high quality audio and video, however this is expensive. Most users do not want to pay more than necessary for their communications. In addition, there will always be a market for lower quality at a lower financial cost (Podolsky 1998). Thus, the minimum levels of media quality that support users undertaking specific tasks need to be determined. It is always important to take the task being performed into consideration, as different tasks will require different levels of quality. For example, an important business meeting over a multimedia conference link will require higher audio and video quality than two friends using MMC for an informal chat.

The establishment of such quality thresholds is essential information for network providers and application designers. Yet, in the areas of HCI and Computer Networking there are no methods or guidelines on the most appropriate way to do this. In telecommunications, subjective assessment methods have been traditionally used,

however the commonly used ITU recommended rating scales have limitations, for example the scale labels do not mean the same in differences languages, therefore they are not the international scales they claim to be. Moreover, the scales are commonly used in isolation, which has drawbacks because subjective assessment is cognitively mediated. This means that external variables such as budget or task difficulty can influence the ratings given. Thus, results obtained may give a misleading impression about the impact of the quality on the user.

This thesis has reported the investigation of a novel, objective method of determining the user cost of media quality degradations in MMC: the measurement of physiological indicators of perceptual strain. User cost measures were taken along with measures of task performance and user satisfaction to create a 3-factor evaluation framework that was utilised in five MMC experiments.

## 7.2 Research questions

The three main research questions addressed in this thesis are:

1. **Can physiological responses to media quality degradations be detected?**

Results showed that differences in mean physiological responses can be detected. Significant main effects of media quality degradations were found in experiments 1 and 2. However, results from experiments 3, 4 and 5 (with the exception of BVP, which did show a significant impact of degradation but post-hoc pairwise comparisons showed no significant differences between conditions) were not significant.

The results from the research reported in this thesis show that there are many variables that can impact upon physiological results, of which the task being performed is one. All the tasks used in this thesis were designed to be engaging, with the exception of the

passive listening task used in experiment 2. Experiments 1 and 3 were passive tasks, whereas experiments 4 and 5 were interactive tasks. The physiological results shed more light on the nature of these tasks.

Directional fractionation was found between SC and HR in experiments 1, 2 and 5. This indicates that these tasks were perceptual in nature, where HR decreased in order to facilitate an intake from the environment. However, it was not found in experiments 3 and 4, thus these tasks can be concluded as being primarily cognitive. Participants in experiment 4 were naïve interviewers, whereas participants in experiment 5 were experienced interviewers. This explains the task being cognitive in experiment 4, as participants had to concentrate on the task and ask questions. Experiment 3 was not designed to be a stressful task and aimed to be similar to that in experiment 1. However, the results from experiment 1 indicated it was a perceptual task (for the second interview), whereas experiment 3 was cognitive. In experiment 1, two interviews were seen and a decision on which candidate should be offered a place at UCL had to be made. However, in experiment 3 four candidates were seen and for each candidate a yes or no had to be given with regard to whether they should be offered a place. Therefore, experiment 3 involved more comparisons between candidates and a decision to be made on each, which appeared to place a cognitive load on the participants.

The physiological results also provide information on other variables, such as gender, repetition of conditions and the order of presentation of conditions. There were significant main effects of gender in experiments 2 (SC) and 4 (HR). In experiment 2 there was not a gender imbalance, therefore it may be that the males in this experiment (who had a higher SC) were more affected by the degradations or it may be because the clips heard were of two males. Their SC for the second presentation of conditions was higher than the first, whereas the SC of females was almost the same for

both presentations, so males may have been bored or frustrated during the second presentation of conditions. In experiment 4, there was a gender imbalance (18 females and 5 men), thus it is likely that this interfered with the gender results.

There was a significant main effect of interview in HR (experiment 1), where participants were more relaxed in interview 2. This may be because they were more relaxed and familiar with the task by the second interview. In addition, there was a significant interaction between interview and group, where the group who experienced 5-25-5fps had a higher level of SC during interview 1, whereas the group who experienced 25-5-25fps had a higher level of SC during interview 2. Overall, group 1 had a lower SC level than group 2.

There was a significant main effect of the repetition of conditions in SC and BVP (experiment 2): SC was higher during the second presentation of conditions, whereas BVP was lower, both of which indicate more perceptual strain during the second presentation of the conditions. There were also significant interactions between presentation and gender (SC) and degradation and presentation (BVP) in this experiment.

There were 3 significant interactions involving the order of presentation of conditions in experiments 4. The results show that the order of frame rate and the frame rate seen and the order that the interviewees were seen, the order of frame rate and frame rate had a significant impact on HR. The order the interviewees were seen also had a significant impact on BVP.

**2. Which media quality degradations have a negative physiological impact on users?**

Experiment 1 showed that in a passive, engaging perceptual task where the frame rate changed from low to high then back to low (5-25-5fps), significant differences between the frame rates were found in SC. SC was significantly higher the second time 5fps was seen in

both interviews. In the first interview, 25fps was significantly higher than the first presentation of 5fps and in the second interview the last presentation of 5fps was significantly higher than 25fps.

Experiment 2 showed that in a passive perceptual rating task, audio degradations (low and high levels of packet loss, loud and quiet volume differences between speakers, echo, audio recorded using a bad microphone and a good quality reference condition) had a significant impact on all signals. The echo and loud conditions were significantly more straining than the reference condition in SC and in HR, the 5% packet loss condition was significantly less straining than the reference condition. Post-hoc comparisons showed no significant differences between conditions in BVP.

3. **How do physiological responses relate to subjective data: do they correlate?**

In experiment 1, tying together the subjective and physiological responses was difficult because the frame rate changed from good to bad quality and vice versa in the same interview, the questionnaire was not sophisticated and it did not include rating scales. However, one of the main findings was that participants did not subjectively notice that the frame rate had changed, whereas physiological differences to the frame rates were observed in the SC data from group 1.

In experiment 2, echo and loud were significantly more straining than the reference condition in SC and were among the three worst rated conditions subjectively. In addition, in HR the 5% packet loss condition was significantly less straining than the reference condition and this was the second best rated condition. It is likely that this concurrence between the two factors occurred because the task was one of passive rating and participants were not being distracted from their ratings by performing an engaging task.

In experiments 3, 4 and 5 it was not possible to compare subjective and physiological responses, due to the lack of significant physiological results.

## 7.3 Contributions of the thesis

The investigation of the impact of MMC quality on the end-user is of the highest importance, as they ultimately determine the success, or otherwise of an application. In the area of HCI, there are no specified criteria or established methods to assist in quality thresholds and the methods that are traditionally used have drawbacks when used in isolation. Therefore, in tackling this issue, this research makes a number of methodological and substantive contributions.

### *7.3.1 Methodological Contributions*

1. A critical evaluation of existing assessment methods for MMC quality was performed and highlighted the drawbacks of utilising subjective assessment methods in isolation (chapter 2).

2. The method of utilising physiological signals to assess the impact of MMC quality was developed in five experiments. The use of the method was then extended to applications other than MMC in three experiments (in the areas of web delay and pricing, web page design and layout and in an area other than HCI: VR (see appendix D)).

3. An analysis of the limitations of measuring physiological signals in this area is given below.

    i. The main problem with measuring physiological signals is that even if a significant difference is detected, it is difficult to attribute valence to it, for example is the participant feeling under perceptual strain or could the response be indicative of something else, such as more

engagement in the task? The way round this that was adopted in the research reported in this thesis was to utilise physiological responses as part of a 3-factor approach, as the other types of data can offer explanations for the physiological results. For example, the SC in experiment 1 showed increases when watching the interview of the candidate that more participants chose to be offered a place on the computing degree course.

ii. It takes a substantial amount of time to measure and analyse the physiological responses due to the volume of data produced. When adding to this the number of strands of data that should be measured in addition to the physiological signals, such as subjective data, this illustrates that there must be good reasons for using them.

iii. Data can be lost due to problems with the sensors or the signals. It can be difficult to pick this up during the experiments, thus making it less easy to recruit another participant. This was found by the research reported in this thesis and also in the research from the Affective computing group at MIT (see section 3.2.5).

iv. Some participants feel uncomfortable wearing the sensors and reported that they were afraid to move their hand for fear of interfering with the results.

v. The content of the task during conditions should be controlled carefully. For example, the candidate in the 5% packet loss condition in experiment 3 (see section 5.3.8.1) was different to the other 3 candidates. Responses to him could have affected the physiological results.

#### 7.3.1.1 Guidelines for measuring physiological signals in HCI

From the research reported in this thesis, guidelines for the use of physiological signals in HCI can be given. These are separated into guidelines for HCI researchers and practitioners, the latter of whom may have less time to collect and analyse results.

**Guidelines for HCI researchers**

The research reported in this thesis benefited from measuring three physiological signals, therefore more than one signal should be measured. The signals often do not concur, for example in experiment 2 the reference condition was the least perceptually straining in SC and the most in HR. In addition, fractionation between SC and HR, as was observed in experiments 1, 2 and 5, gave critical information about the nature of the tasks that the participants were performing.

As mentioned in section 7.3.1, measuring physiological signals is not a quick method. In the first instance, adding a 15-minute baseline session to an experiment can make some experiments unduly long. Having a within-subjects design, as in the research reported in this thesis, is the most common design in experiments utilising physiological signals (see table 4), however this also takes time because each participant has to complete each condition. Finally, in the research reported in this thesis, the analysis of the physiological signals took longer than that of the subjective responses. These time factors must be considered in the planning of an experiment to ensure it is not too long for participants and to ensure that the experimenter has sufficient time for the analysis.

Careful consideration must be given to the type of task that the participants will perform. At the most basic level, the measurements will be easier to interpret if movement is kept to a minimum. If movement is required, as was in experiment 8 (see appendix D), an ECG sensor should be used to measure HR and BVP should not be measured, as it is very sensitive to motion artefacts. As has been

discussed, in experiments 1 and 2 there were significant differences in physiological responses to media quality degradations. The task in experiment 2 was solely rating, and the task in experiment 1 was passive and involved making a simple decision between two candidates. However, as the tasks became more complex (for example, in experiment 3 giving a decision about 4 candidates) and interactive, it became more difficult to obtain significant differences in the physiological signals. Therefore, researchers must make a trade-off between the ecological validity of tasks and obtaining significant results. From the results of the research reported in this thesis, a minimum of 24 participants are needed to obtain significant differences in results. However, around 5 more than this should be measured to account for the loss of data as a result of problems with the equipment.

The experience of participants with the task should also be taken into account. This was illustrated in experiments 4 and 5, where experiment 5 (involving experienced interviewers) showed a fractionation between signals but experiment 4 (involving naïve interviewers) did not. This indicates that the task in experiment 4 was cognitive, yet in experiment 5 it was perceptual.

The environment of the room in which the experiment takes place must be controlled. The temperature should be comfortable and any possible distractions, such as the telephone ringing, should be eliminated. Where this is not possible, such events should be recorded so that they can be matched up in the data. Controlling and accounting for such external variables assists in the interpretation of the physiological data.

Physiological signals should not be measured in isolation. Results from the experiments carried out as part of the research reported in this thesis showed that gathering user satisfaction and task performance data (where relevant) all add together to give a rich set of data about the user experience. For example, in experiment 4

there were no significant differences in the physiological signals, yet when the subjective data and information about how long it took participants to complete the interviews under each frame rate were considered, it was found that the higher levels of frame rate allowed participants to complete their task quicker and participants said they preferred it.

The order in which conditions are presented can have an impact on physiological signals. For example, significant interactions involving order were found in the HR data in experiment 4. Therefore, it must be ensured that order is counterbalanced, as was in the research reported in this thesis. In addition, a gender balance should be obtained in experiments, as it can impact on results, such as in experiment 2.

Repeating conditions has an impact on physiological responses, for example in experiment 2 the second presentation of conditions was significantly more perceptually straining than the first in SC and BVP. Therefore, repetition of conditions must be accounted for in the analysis.

The analysis performed on the signals depends on what is being measured. If simple changes in levels over conditions longer than one minute are being measured, as in the research reported in this thesis, they can be detected by using means minus the baseline. If more complex states, such as emotions are being detected, then the use of a number of features, such as the number of ORs may be more appropriate, however specific recommendations for the features to be measured is outwith the scope of this thesis.

**Guidelines for practitioners**

If time limits are in operation, then it may be more effective to use physiological signals in an in-depth case study with a few participants. By initially graphing the data and matching it up with events, where relevant, this will allow the practitioner to determine if

any basic changes in response to the variable have occurred, as opposed to performing statistical analysis. However, measures of task performance and subjective assessment should still be taken in order to get an accurate indication of the user experience.

If measurements are being taken over a longer period of time, or in a field trial, then the physiological measuring equipment used in the research reported in this thesis may be too restrictive, thus alternatives should be explored. An example of a recent system devised for long-term, unobtrusive monitoring is the BodyMedia SenseWear Pro2 Armband[36]. This is a system that is worn on the arm and collects many forms of data, such as physiological signals (including temperature and SC), physical activity, number of steps taken, sleep/wake states and energy expenditure. It also allows the wearer to time stamp the data for important events, which can then be matched up with the physiological data. It is currently used by clinicians, researchers and also by caregivers to monitor patients remotely through a web based application and could be used to measure the long-term impact of interacting with MMC on a daily basis.

The LifeShirt could also be used for such monitoring (Wilhelm et al. 2003) This system has been developed to monitor changes in respiration in people with medical or psychological disorders (such as measuring hyperventilation in panic) out of a lab-based setting. It measures signals such as respiration, ECG and postural changes. The signals are displayed and saved on a small computer. There is also the function for the wearer to record their symptoms and mood. Therefore, this system and the BodyMedia armband with their ambulatory functionality and capacity to account for physical events and gather subjective data offer a solid proposition for use in long-term trials. Finally, research by Goulev (2005) has extended that of

---

[36] www.bodymedia.com

the Affective Computing research group (see section 3.2.5) by developing a system that detects emotion from the SC signal in real-time, thus offering the potential for an application to adapt to the state of the user.

### *7.3.2 Substantive Contributions*

1. Guidelines on the minimum levels of media quality (for some parameters) that users require for completing a number of specific tasks are now given.

i. When an engaging task is being performed that does not place a cognitive load on participants, seeing 5-25-5fps for the first time can cause SC to significantly increase between the first presentation of 5fps and 25fps and between the two presentations of 5fps. In a second interview, SC again significantly increases between the two presentations of 5fps and the second presentation of 5fps is significantly higher than 25fps. This result implies that applications should avoid providing frame rates as low as 5fps at the beginning of such a task, due to the increases in strain they induce.

ii. In a passive listening task, echo and loud volume differences between speakers caused increases in SC from a normal quality condition, whereas HR showed a significant decrease to 5% packet loss from a reference condition. Therefore, applications could be designed to provide a feature that tests the audio stream for degradations caused by the hardware set-up and end user behaviour and alerts the user when such degradations are present (Watson & Sasse, 2000). Such a feature could also give the user assistance to find the cause of the degradation so it can be rectified. With Voice over IP applications like Skype (see section 2.2.1.1) being increasingly used, this is extremely important.

iii. In active interviewing tasks, responses to the task can drown out responses to the quality.

2. A fuller appreciation of the impact that media quality degradations caused by the network, the hardware set-up and end-user behaviour have on users' physiology and subjective ratings has been gained, which again appears to depend on the task.

i. Passive tasks

In the passive task in experiment 1, beginning the interviews at 5fps (caused by the network) induced significant differences between the frame rates in SC. In the passive task in experiment 2, loud volume differences between speakers (caused by end user behaviour) and echo (caused by end user behaviour) caused significant increases in SC.

ii. Interactive tasks

In the two interactive tasks, the physiological responses to the quality may have been drowned out by responses to the task.

3. A demonstration of effects that subjective data does not pick up was given, such as the impact of repetition of conditions in SC and BVP in experiment 2. In addition, occasions where the physiological responses did not produce significant results and the subjective results have to be focussed on was shown, such as in experiment 4. This means that both streams of data, in addition to task performance, should be gathered to aid their interpretation.

Overall, the use of physiological measures in determining quality thresholds is suited to passive perceptual tasks that do not place a cognitive load on the user, such as the passive listening task in experiment 2 or watching recorded interviews, as in experiment 1.

However, they should not be replaced by or be a replacement for subjective measures. In HCI generally, the measures have great potential. The finding that the signals gave additional information about the nature of the tasks participants were performing through their fractionation, or otherwise, was of importance. In addition, they are useful when a variable changes within a condition, as they give continuous data. Subjective responses may not be able to pick up differences due to engagement in the task (such as whether participants noticed the frame rate change in experiment 1) and may also be affected by primacy or recency effects.

The results from the research reported in this thesis have provided support for the use of the 3-factor evaluation approach in HCI. The wealth of measures taken provided additional information that would have remained uncovered if only one factor been adopted.

## 7.4 Limitations of the thesis research

There are four main limitations to the research reported in this thesis. Firstly, the lack of significant results in some experiments may be due to not having enough participants. Twenty-four participants were included in each experiment, except experiment 5, however due to outliers and problems with the equipment and signals the number used in the analysis was often less than this. Experiment 5 only had eleven participants, because experienced interviewers were needed, therefore it is unsurprising that this did not produce significant results. In a within-subjects design, 24 participants should be the aim, however more should be included initially to prepare for losing data.

The design of experiments 1 and 3 did not allow for direct comparisons between subjective and physiological responses as participants experienced both good and poor quality within the same conditions. However, it is experiments such as this that can benefit from the continuous stream of data that physiological signals offer. In

addition, the basic questionnaire used in experiment 1 did not use rating scales, which made it difficult to compare the subjective and physiological results.

Finally, there was a gender imbalance in some of the experiments, which may have influenced results. This mainly occurred because four out of five of the experiments took part in the Computer Science department at UCL, which had more males than females at the time the research was conducted.

## 7.5 Recommendations

### *7.5.1 Recommendations for HCI researchers*

1. The 3-factor approach should be used in media quality evaluation and none of the three factors should be measured in isolation.
2. When measuring user cost through physiological signals, researchers should use more than one signal. If only one signal had been measured in the research reported in this thesis, important effects would have been missed, such as the significant interactions in HR in experiment 4 involving order. In addition, measuring both SC and HR can provide information about the nature of the task being performed through directional fractionation or otherwise.
3. Careful design of experiments including the environment, tasks used, order of presentation of conditions, repetition of conditions and the number of participants is required if physiological responses are measured.
4. The use of physiological signals can be applied to other areas of HCI, such as web site evaluation (see appendix D).
5. Physiological signals can be used to give additional data in areas other than HCI, such as VR (see appendix D).
6. The use of physiological signals must be justified before carrying out an experiment, as they are not a 'quick fix'.

### *7.5.2 Recommendations for network providers*

1. Audio degradations and video frame rate have a significant impact on physiological responses in passive perceptual tasks.

i. Low video frame rates can cause increases in SC when experienced first in a multimedia conference.

ii. Problems due to end user behaviour and equipment, such as loud volume differences between speakers and echo can negatively affect users.

2. The 3-factor approach should be used when evaluating the impact of media quality in order to give a fuller picture of the user experience.

## 7.6 Agenda for future research

The results of the research reported in this thesis have posed many questions, which should be investigated in future research. The experiments in this thesis were all around an hour in length. The next stage would be to measure the signals in longer MMC tasks over hours, weeks or months, for example via a user having weekly videoconferencing meetings. Would users habituate to the quality or would the perceptual strain accumulate? If the latter occurred, then adverse effects on health could be seen when coupled with job stress. It would also be advantageous to investigate the use of alternative physiological signals, such as muscle tension in the face, to determine if they can be used as a measure of user cost in MMC.

Performing the research mentioned above would necessitate moving the research from the lab into the field. There are countless variables at operation in the field, which would make data interpretation difficult, however the role of physiological measurements in the field must be determined. At the very least they would be able to measure basic changes, which could be useful for people trying to reduce their stress levels.

In performing longer-term studies, it would be beneficial to use physiological measures that can give an indication of stress, such as the hormone cortisol. Looking at longer-term stress responses would be very different to the research conducted in this thesis, which looked at short-term perceptual strain responses. However, it is important to see what happens with longer-term interaction with media quality degradations and how deep the effects of the perceptual strain go. If increases were found in cortisol, it would indicate that the person's health could be affected in the long term.

Within longer-term trials, investigating the misattribution of affective responses is important. This hypothesises that when a user communicates with someone under poor quality, the negative effects of the quality may be attributed to the communicator, as opposed to the technology. This effect was hinted at in experiment 1, where the candidate that group 1 saw first was chosen by significantly more participants than the candidate seen second, whereas in group 2 there were no significant differences between the candidates chosen. This pattern of results indicates that in group 1 (who experienced poorer quality), the negative effects of the quality may have influenced the decision about which candidate should be offered a place to the detriment of the candidate seen second. Such effects could offer information on the minimum quality levels required for such tasks.

In areas where movement of participants is required, the use of wearable sensors and methods to account for movement in the analysis of the signals should be explored. Many systems have been developed to do this, such as the LifeShirt (see section 7.3.1.1). Another measuring device that would be useful in measuring long-term behavioural responses is the pressure mouse (Reynolds 2001). This mouse is equipped with eight pressure sensors, thus it can passively measure tension and frustration and the data gathered can be used to adapt interfaces in response to the user's state. In the

long-term measurement of responses, measuring behavioural states through such tools would provide a valuable stream of information with which to compare to physiological signals. In conclusion, the research presented in this thesis serves as a solid basis for the 3-factor approach to media quality evaluation, which has exciting potential in other areas.

# References


Aldridge R, Davidoff J, Ghanbari M, Hands D, Pearson D. 1995. Measurement of scene-dependent quality variations in digitally coded television pictures. *IEE Proceedings - Vision, Image and Signal Processing* 142: 149-154

Aldridge RP, Hands DS, Pearson DE, Lodge NK. 1998. Continuous assessment of digitally-coded television pictures. *IEE Proceedings - Vision, Image and Signal Processing* 145: 116-123

Allanson J, Wilson GM. 2001. *The role of electrophysiology in HCI.* Presented at IHM-HCI 2001, September 10-14, Lille, France, 251-252

Allanson J, Wilson GM. 2002a. *Physiological Computing.* Presented at CHI 2002, April 20-25, Minneapolis, Minnesota, USA

Allanson J, Wilson GM. 2002b. *Physiological Computing*, University of Lancaster, Computing Department, Technical Report, ISSN 1447-447X

Anderson A, Smallwood L, MacDonald R, Mullin J, Fleming A. 2000. Video Data and Video Links in Mediated Communication: What do Users Value? *International Journal of Human Computer Studies* 52: 165-187

Anderson AH, Newlands A, Mullin J, Fleming AM, Doherty-Sneddon G, Velden JVd. 1996. Impact of video-mediated interaction on simulated service encounters. *Interacting with Computers* 8: 193-209

Anderson AH, O'Malley C, Doherty-Sneddon G, Langton S, Newlands A, Mullin J, Fleming A, Velden JVd. 1997. The impact of VMC on collaborative problem solving: An analysis of task performance, communicative process, and user satisfaction. In *Video-Mediated Communication*, ed. K.E. Finn, AJ Sellen, S Wilbur, pp. 133-155: Lawrence Erlbaum Associates, Inc

Andreassi JL. 2000. *Psychophysiology. Human Behaviour and Physiological Response*: Lawrence Erlbaum Associates

Argyle M. 1990. *Bodily Communication.* London: Routledge

Barber P, Laws JV. 1994. Image quality and communication. In *Multimedia Technologies and Future Applications*, ed. RI Damper, W Hall, JW Richards, pp. 163-178. London: Pentech Press

Bersak D, McDarby G, Augenblick N, McDarby P, McDonnell D, McDonald B, Karkun R. 2001. Intelligent biofeedback using an immersive competitive environment. Available from http://medialabeurope.org/mindgames/publications/publicationsAtlanta2001rev3.pdf

Blokland A, Anderson AH. 1998. Effect of low frame-rate video on intelligibility of speech. *Speech Communication* 26: 97-103

Bouch A, Sasse MA. 1999. *Network Quality of Service: What do Users Need?* Presented at 4th International Distributed Conference, September 22-23, Madrid, 78-90

Bouch A, Watson A, Sasse MA. 1998. *QUASS - A tool for measuring the subjective quality of real-time multimedia audio and video.* Presented at HCI '98, 1-4 September, Sheffield, England

Boyle EA, Anderson AH, Newlands A. 1994. The effects of visibility on dialogue and performance in a co-operative problem solving task. *Language and Speech* 37: 1-20

Brown CC. 1967. The techniques of plethysmography. In *Methods in psychophysiology*, ed. CC Brown. Baltimore: Williams & Wilkins

Brown G, Anderson AH, Yule G, Shillcock R. 1984. *Teaching Talk*: Cambridge University Press

Cacioppo JT, Tassinary LG. 1990. Inferring psychological significance from physiological signals. *American Psychologist* 45: 16-28

Cacioppo JT, Tassinary LG, Berntson GG, eds. 2000. *Handbook of Psychophysiology*: Cambridge University Press

Cannon WB. 1915. *Bodily Changes in Pain, Hunger, Fear and Rage.* New York: Appleton

Chapanis A, Ochsman R, Parrish A, Weeks G. 1972. Studies in interactive communication: the effects of four communication modes on the behaviour of teams during co-operative problem solving. *Human Factors* 14: 487-509

Cockton G. 2002. From doing to being: bringing emotion into interaction (Editorial). *Interacting with Computers* 14: 89-92

Cook MR. 1974. Psychophysiology of peripheral vascular changes. In *Cardiovascular psychophysiology*, ed. PA Obrist, AH Black, J Breber, AV Dicara, pp. 60-84. Chicago: Aldine

Dillon C. 2002. *'It's been emotional': affect, physiology and presence.* Presented at Fifth International Workshop on Presence, 9-11 October, Universidade Fernando Pessoa, Porto

Dillon C, Keogh E, Freeman J, Davidoff J. 2001. *Presence: Is your heart in it?* Presented at 4th International Workshop on Presence, 21-23 May, Temple University, Philadelphia

Dix A, Finlay J, Abowd GD, Beale R. 2004. *Human-Computer Interaction*: Pearson Prentice Hall

Duncan-Johnson CC, Coles MGH. 1974. Heart rate and disjunctive reaction times: the effects of discrimination requirements. *Journal of Experimental Psychology* 103: 1160-1168

Field A. 2000. *Discovering statistics using SPSS for windows.* London: SAGE Publications

Fluckiger F. 1995. *Understanding networked multimedia*: Prentice Hall

Freeman J, Avons SE, Pearson DE, IJsselsteijn WA. 1999. Effects of sensory information and prior experience on direct subjective ratings of presence. *Presence: Teleoperators and Virtual Environments* 8: 1-13

Freeman J, Lodge N, Moss T. 2001. Taking the viewer there. Available from http://homepages.gold.ac.uk/immediate/immersivetv/Freeman_Lessiter%20-%20tile2001.pdf

Frokjaer E, Hertzum M, Hornbaek K. 2000. *Measuring usability: are effectiveness, efficiency and satisfaction really correlated?* Presented at ACM CHI 2000 Conference on Human Factors in Computing Systems, April 1-6, The Hague, The Netherlands, 345-352

Gale A, Christie B, eds. 1987. *Psychophysiology and the Electronic Workplace*: John Wiley & Sons

Garau M. 2003. *The Impact of Avatar Fidelity on Social Interactions in Virtual Environments*. Doctoral thesis. University of London

Goulev P. 2005. *An Investigation in to the use of AffectiveWare in interactive computer applications*. Doctoral thesis. Imperial College of Science, Technology and Medicine

Graham FK, Clifton RK. 1966. Heart rate change as a component of the Orienting Response. *Psychological Bulletin* 65: 305-320

Handley MJ. 1997. *An Examination of MBone Performance*, USC/ISI Research Report: ISI/RR-97-450.

Healey J. 2000. *Wearable and Automotive Systems for Affect Recognition from Physiology*. Doctoral thesis. Massachusetts Institute of Technology, Boston, USA

Healey J, Picard R, Dabek F. 1998. *A new affect-perceiving interface and its application to personalised music selection*. Presented at Workshop on Perceptual User Interfaces, November 4-6. San Francisco, CA

Healey J, Picard RW. 1998. *StartleCam: a cybernetic wearable camera*. Presented at International Symposium on Wearable Computers, October 19-20, Pittsburgh, PA

Hearnshaw D. 1999. *Desktop conferencing for tutorial support*. Doctoral thesis. University College London, University of London

Horn D. 2001. *Seeing is Believing: Video Quality and Lie Detection*. Doctoral thesis. University of Michigan

ITU-R B-. Methodology for the subjective assessment of the quality of television pictures. *Available from http://www.itu.int/publications/itu-t/iturec.htm*

ITU-T P. Interactive test methods for audio visual communications. *Available from http://www.itu.int/publications/itu-t/iturec.htm*

ITU-T P. Methods for subjective determination of transmission quality. *Available from http://www.itu.int/publications/itu-t/iturec.htm*

ITU-T P. Subjective video quality assessment methods for multimedia applications. *Available from http://www.itu.int/publications/itu-t/iturec.htm*

Jones BL, McManus PR. 1986. Graphic scaling of qualitative terms. *Society of Motion Picture Television Engineers Journal* November 1986: 1166-1171

Kahneman D. 1973. *Attention and Effort*: Englewood Cliffs, NJ; Prentice-Hall

Kahneman D, Tursky B, Shapiro D, Crider A. 1969. Pupillary, heart rate and skin resistance changes during a mental task. *Journal of Experimental Psychology* 79: 166-167

Kaiser DN, Sandman CA. 1975. Physiological patterns accompanying complex problem solving during warning and

non-warning conditions. *Journal of Comparative and Physiological Psychology* 89: 357-363

Kies JK, Williges RC, Rossin MB. 1996. *Controlled laboratory experimentation and field study evaluation of video conference for distance learning applications. Rep. HCIL-96-02*, Virginia Tech

Kies JK, Williges RC, Rosson MB. 1997. Evaluating desktop videoconferencing for distance learning. *Computers in Education* 28: 79-91

Kitawaki N, Nagabuchi H. Quality assessment of speech coding and speech synthesis systems. In *IEEE Communications Magazine*, pp. 36-44

Knoche H, DeMeer HG, Kirsch D. 1999. *Utility Curves: Mean Opinion Scores Considered Biased.* Presented at 7th International Workshop on Quality of Service, June 1-4, London, UK

Lacey JI. 1959. Psychophysiological approaches to the evaluation of psychotherapeutic process and outcome. In *Research in psychotherapy*, ed. EA Rubenstein, MB Parloff. Washington, D.C.: American Psychological Association

Lacey JI. 1967. Somatic response patterning and stress: some revisions of activation theory. In *Psychological stress: issues in research*, ed. MH Appley, R Trumbull, pp. 14-42. New York: Appleton-Century-Crofts

Lacey JI, Kagan J, Lacey BC, Moss HA. 1963. The visceral level: situational determinant and behavioural correlates of autonomic response patterns. In *Expression of the emotions*

*in man*, ed. PH Knapp. New York: International Universities Press

Lang PJ, Greenwald MK, Bradley MM. 1993. Looking at pictures: affective, facial, visceral and behavioural reactions. *Psychophysiology* 30: 261-273

Lewis C. 1982. *Using the thinking-aloud method in cognitive interface design. Rep. RC 9265*, Yorktown heights, NY: IBM T.J. Watson Research Centre

Lindsley DB. 1952. Psychological phenomena and the electroencephalogram. *Electroencephalography and Clinical Neurophysiology* 4: 443-456

Mandryk RL, Inkpen KM. 2004. *Physiological indicators for the evaluation of co-located collaborative play*. Presented at Conference on Computer Supported Co-operative Work, November 6-10, Chicago, USA, 102-111

Manzanaro JG, Escalada LJ, Lioreda MH, Szymanski M. 1991. *Subjective image quality assessment and prediction in digital video communications.* COST 212 HUFIS Report

McCanne S, Jacobson V. 1995. *Vic: A flexible framework for packet video*. Presented at ACM Multimedia '95, November 5-9, San Francisco, USA

McGurk H, MacDonald JW. 1976. Hearing lips and seeing voices. *Nature* 264: 126-130

Meehan M. 2001. *Physiological Reaction as an Objective Measure of Presence in Virtual Environments.* Doctoral thesis. University of North Carolina at Chapel Hill

Monk AF, Watts L. 1995. *A poor quality video link affects speech but not gaze*. Presented at CHI '95, May 7-11, Denver, Colorado, 274-275

Mullin J, Jackson M, Sasse MA, Watson A, Anderson AH, Smallwood L, Wilson G. 2002. The ETNA Taxonomy. *Available from http://www-mice.cs.ucl.ac.uk/multimedia/projects/etna/taxonomy.pdf*

Mullin J, Smallwood L, Watson A, Wilson GM. 2001. *New techniques for assessing audio and video in real-time interactive communication*. Presented at IHM-HCI 2001, September 10-14, Lille, France, 221-222

Nakazono K. 1998. Frame rate as a QoS parameter and its influence on speech perception. *Multimedia Systems* 6: 359-366

Nardi B, Schwarz H, Kuchinsky A, Leichner R, Whittaker S, Sclabassi R. 1993. *Turning away from talking heads: an analysis of "video-as-data"*. Presented at CHI '93 Human Factors in Computing Systems, April 24-29, Amsterdam, The Netherlands, 327-341

O'Conaill B, Whittaker S, Wilbur S. 1993. Conversations over video conferences: an evaluation of the spoken aspects of video-mediated communication. *Human Computer Interaction* 7: 398-428

Olson JS. 1994. *In a framework about task-technology fit, what are the task features?* Presented at Computer Supported Co-operative Work '94. Workshop on Video-Mediated

Communication: Testing, Evaluation and Design Implications, October 22-26, Chapel Hill, North Carolina, USA

Olson JS, Olson GM, Meader DK. 1994. *What mix of audio and video is useful for remote real-time work*. Presented at Computer Supported Co-operative Work '94. Workshop on Videomediated Communication: Testing, Evaluation and Design Implications, October 22-26, Chapel Hill, North Carolina, USA

O'Malley C, S Langton S, A Anderson A, G Doherty-Sneddon G, Bruce V. 1996. Comparison of face-to-face and video-mediated interaction. *Interacting with Computers* 8: 177-192

Pavlov IP. 1927. *Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex*. London: Oxford University Press

Picard R. 1997. *Affective Computing*. Cambridge: MIT Press

Picard R, Healey J. 1997. Affective Wearables. *Personal Technologies* 1: 231-240

Picard R, Scheirer J. 2001. *The Galvactivator: a glove that senses and communicates skin conductivity*. Presented at 9th International Conference on Human Computer Interaction, August 5-10, New Orleans

Picard R, Vyzas E, Healey J. 2001. Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Transactions Pattern Analysis and Machine Intelligence* 23: 1175-1191

Podolsky M, Romer, C. and McCanne, S. 1998. *Simulation of FEC-based error control for packet audio on the Internet*. Presented at IEEE INFOCOM '98 - The Conference of Computer Communications, March 29- April 2, San Francisco, USA, 505-515

Preece J, Rogers Y, Sharp H, Benyon D, Holland S, Carey T. 1994. *Human Computer Interaction*: Addison-Wesley

Preminger JE, Tasell DJV. 1995. Quantifying the relationship between speech quality and speech intelligibility. *Journal of Speech and Hearing Research* 38: 714-725

Reisberg D, McLean J, Goldfield A. 1987. A lipreading advantage with intact auditory stimuli. In *Hearing by Eye: The Psychology of Lipreading*, ed. B Dodd, R Campbell. London: Lawrence Erlbaum Associates

Reynolds C. 2001. *The sensing and measurement of frustration with computers*. Doctoral thesis, Massachusetts Institute of Technology, Boston, USA

Rowe DW, Sibert J, Irwin D. 1998. *Heart rate variability: indicator of user state as an aid to human computer interaction*. Presented at CHI 1998, April 18-23, Los Angeles, CA, USA

Roy RR. 1994. Networking constraints in multimedia conferencing and the role of ATM networks. *AT & T Technical Journal*: 97-108

Rudman C, Hertz R, Marshall C, Dykstra-Erickson E. 1997. Channel overload as a driver for adoption of desktop video for distributed group work. In *Video Mediated Communication*, ed.

KE Finn, AJ Sellen, SB Wilbur. Malwah, NJ: Lawrence Erlbaum Associates

Sasse MA, Bilting U, Schulz C-D, Turletti T. 1994. *Remote Seminars through Multimedia Conferencing: Experiences from the MICE project*. Presented at INET'94/JENC5 International Networking Conference, June 13-17, Prague, 251/1-251/8

Scheirer J, Fernandez R, Klein J, Picard RW. 2002. Frustrating the user on purpose: a step towards building an affective computer. *Interacting with Computers* 14: 93-118

Shackel B. 1981. *The Concept of Usability*. Presented at IBM Software and Information Usability Symposium, September 15-18, New York, USA, 1-29

Short J, Williams E, Christie B. 1976. *The Social Psychology of Telecommunications*: Wiley

Simons RF, Detenber BH, Roedema TM, E RJ. 1999. Emotion processing in three systems: the medium, and the message. *Psychophysiology* 36: 619-627

Sokolov EN. 1963. *Perception and the conditioned reflex*. New York: MacMillan

Summerfield Q. 1992. *Lipreading and audio-visual speech perception. Rep. B335*, Philosophical Transactions of the Royal Society of London

Tang JC, Issacs EA. 1993. Why do users like video? Studies of multimedia collaboration. *Computer-Supported Co-operative Work* 1: 163-196

Veinott ES, Olson J, Olson GM, Fu X. 1997. *Video matters! When communication ability is stressed, video helps*. Presented at CHI '97, March 22-27, Atlanta, GA, 315-316

Venables PH, Christie MJ. 1980. Electrodermal activity. In *Techniques in Psychophysiology*, ed. I Martin, PH Venables, pp. 3-69: John Wiley & Sons

Ward RD, Marsden PH. 2003. Physiological responses to different WEB page designs. *International Journal of Human Computer Studies* 59: 199-212

Ward RD, Marsden PH, Cahill B, Johnson CA. 2001. *Using skin conductivity to detect emotionally significant events in human-computer interaction*. Presented at IHM-HCI 2001, September 10-14, Lille, France, 25-28

Warren RM. 1970. Perceptual restoration of missing speech sounds. *Science* 167: 392-393

Wastell DG. 1990. *Mental effort and task performance: towards a psychophysiology of HCI*. Presented at INTERACT '90 - 3rd International Conference on HCI, August 27-31, Cambridge, UK, 107-112

Wastell DG, Brown ID, Copman AK. 1982. A psychophysiological investigation of system efficiency in public telephone switchrooms. *Ergonomics* 25: 1013-1040

Wastell DG, Cooper CL. 1996. Stress and technological innovation: a comparative study of design practices and implementation strategies. *European Journal of Work and Organisational Psychology* 5: 377-397

Watson A. 2001. *Assessing the Quality of Audio and Video Components in Desktop Multimedia Conferencing.* Doctoral thesis. University College London

Watson A, Sasse MA. 1997. *Multimedia conferencing via multicast: determining the quality of service required by the end user.* Presented at AVSPN '97 - International Workshop on Audio-Visual Services over Packet Networks, September 15-16, Aberdeen, Scotland, 189-194

Watson A, Sasse MA. 1998. *Measuring perceived quality of speech and video in multimedia conferencing applications.* Presented at ACM Multimedia '98, September 12-16, Bristol, England, 55-60

Watson A, Sasse MA. 2000. *The Good, the Bad and the Muffled: the Impact of Different Degradations on Internet Speech.* Presented at 8th ACM International Conference on Multimedia, October 30 - November 3, Marina Del Rey, California, 269-302

Whitefield A, Wilson, F. and Dowell, J. 1991. A framework for human factors evaluation. *Behaviour and Information Technology* 10: 65-79

Wilhelm F, Roth WT, Sackner MA. 2003. The LifeShirt: an advanced system for ambulatory measurement of respiratory and cardiac functions. *Behaviour Modification* 27: 671-691

Wilson F, Descamps PT. 1996. *Should We Accept Anything Less than TV Quality: Visual Communication.* Presented at

International Broadcasting Convention, September 12-16, Amsterdam

Wilson G. 1999a. The relationship between media quality and user cost in networked multimedia applications. In *Interfaces*, pp. 27-28

Wilson GM. 1999b. *The Relationship between Media Quality and User Cost in Networked Multimedia Applications*. Presented at Affective Computing Workshop, held by British HCI Group and University College London, April 10, London, UK

Wilson GM. 2000. *Neglected data: user cost in networked multimedia applications*. Presented at HCI 2000, September 5-8, Sunderland, UK, 133-134

Wilson GM. 2001. *Psychophysiological indicators of the impact of media quality on users*. Presented at CHI 2001, March 31 - April 5, Seattle, USA, 95-96

Wilson GM, Sasse MA. 1999. *Listen to your heart rate: counting the cost of media quality*. Presented at International Workshop on Affect in Interactions, October 21 - 22, Siena, Italy, 16-20

Wilson GM, Sasse MA. 2000a. *Do Users Always Know What's Good For Them? Utilising Physiological Responses to Assess Media Quality.* Presented at HCI 2000, September 5 - 8, Sunderland, UK. 327-339

Wilson GM, Sasse MA. 2000b. *The head or the heart? Measuring the impact of media quality*. Presented at CHI 2000, April 1- 6, The Hague, The Netherlands, 117-118

Wilson GM, Sasse MA. 2000c. *Investigating the Impact of Audio Degradations on Users: Subjective vs Objective Assessment Measures*. Presented at OZCHI 2000: Interfacing Reality in the New Millennium, December 4 - 8, Sydney, Australia, 135-142

Wilson GM, Sasse MA. 2000d. Listen to Your Heart Rate: Counting the Cost of Media Quality. In *Affective Interactions - Towards a New Generation of Computer Interfaces*, ed. A Paiva, pp. 9-20: Springer

Wilson GM, Sasse MA. 2000e. *Multimedia conferencing: what cost to users?* Presented at 6th Open European Summer School: Innovative Internet Applications, September 13-15, University of Twente, Enschede, The Netherlands, 173-180

Wilson GM, Sasse MA. 2004. From doing to being: getting closer to the user experience. *Interacting with Computers* 16: 697-705

Winton WM, Putnam L, Krauss R. 1984. Facial and autonomic manifestations of the dimensional structure of emotion. *Journal of Experimental Social Psychology* 20: 195-216

# Appendix A: Experiment 1 questionnaire

1. Did you feel under any stress due to the quality at any point during this interview?
2. Did you feel excited at any point throughout this interview?
3. On a scale of 1 to 5, how comfortable did you feel with the quality, with 1 being very comfortable and 5 being very uncomfortable?
4. Did you feel you could judge the character of the candidate well using this medium of communication?
5. Did the quality of the video change at any point during the interview? If so, when?
6. Overall, do you feel the quality of the video was good enough to support the interview?
7. Do you prefer the video to be there as opposed to audio only? If so, why?
8. Did the video distract the audio at any point?

Further comments on quality / any observations:

The following three questions were asked at the end of the experiment.

9. Did you feel the task put you under any pressure?
10. Did having the sensors on your hand put you under any pressure?
11. Which candidate should be offered the place?

# Appendix B: Candidate assessment form

| Quality | Very high | High | Dept. norm | Low | Very low | Unquality |
|---|---|---|---|---|---|---|
| **General Intelligence** Articulate, logical, can debate a point. | | | | | | Hesitant, boring, stereotyped. |
| **Motivation** Why Computer Science? Widely read, generally interested in life? | | | | | | Doing Computer Science because couldn't do anything else. |
| **Social Skills** Have they influenced or interested people? | | | | | | No interest in other people, societies etc. |
| **Creativity** Signs of originality. | | | | | | Stereotyped answers, has never done anything interesting. |
| **Enthusiasm** Wants to take the course because of real interest in areas offered. | | | | | | Only taking course as it is seen to be a guaranteed path for employment. |
| **Academic Intelligence** Based on predicted A level performance and interview, what chance does the student have of completing degree? | | | | | | The student is academically qualified but does not possess necessary staying power. |
| **Self Esteem** Despite predicted A level results, would an offer in a lower category be better? | | | | | | Regardless of the class of offer the student will be interested. |
| **Other Strengths** If low A level results predicted, is there evidence of other relevant factors? | | | | | | No evidence presents itself. |

Should the candidate be offered a place?

# Appendix C: Experiment 3 questionnaire

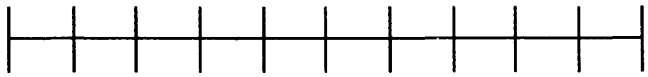
1. What did you think of the quality of the audio? (1=very poor, 100=very good)

1 10 20 30 40 50 60 70 80 90 100

Please describe any problems:

2. How adequate was the audio quality for the purposes of the interview? (1=totally inadequate, 100=completely adequate)

1 10 20 30 40 50 60 70 80 90 100



3. Did the quality of the audio change throughout the interview? Yes/No

4. What did you think of the quality of the video? (1=very poor, 100=very good)

1 10 20 30 40 50 60 70 80 90 100

Please describe any problems:

5. How adequate was the video quality for the purposes of the interview? (1=totally inadequate, 100=completely adequate)

1 10 20 30 40 50 60 70 80 90 100

6. Did the quality of the video change throughout the interview? Yes/No

# Appendix D: Extending the scope of the method

This appendix reports the results of three experiments outwith the area of MMC quality assessment to which the author of this thesis contributed. The purpose of this contribution was to examine the potential of physiological measurements in a different application and different area to the research reported in this thesis.

## Experiment 6: Using physiological measures to determine the impact of delay and pricing structure in a web-based library application

This experiment was designed and run by Anna Bouch (UCL Computer Science department). The analysis of the physiological data was performed by John McCarthy (UCL Computer Science department). The author of this thesis advised on the measuring of physiological signals and the equipment set-up. The results of this study were not published.

This experiment investigated the impact of QoS and pricing structure on 32 participants. Two types of delay were used as the QoS measure: fixed interval (6 seconds) and variable interval (2-14 seconds). There were 2 levels of pricing structure: fixed or variable. All participants were exposed to four conditions covering all combinations of manipulations.

The participants had to complete a series of web searches and download a number of files. The study was set up as a simulation, however all pages were cached locally and the response time was controlled.

| | | QoS | |
|---|---|---|---|
| | | **Stable** | **Variable** |
| Price | Stable | 1 | 2 |
| | Variable | 3 | 4 |

**Table 17: Conditions in experiment 6**

Task performance, user satisfaction, through the QUASS tool (see section 2.4.2), and user cost data, through SC[37], were collected. All participants completed the task successfully. The QUASS results showed no reliable effects of how participants positioned the slider across conditions: the participants' responses were highly variable, which potentially masked any effects of the conditions.

Reliable differences found in the physiological signals. Across all conditions, SC was significantly higher during the experiment than during the baseline session ($t(31)=-8.571$, $p=<0.001$) (see figure 62).

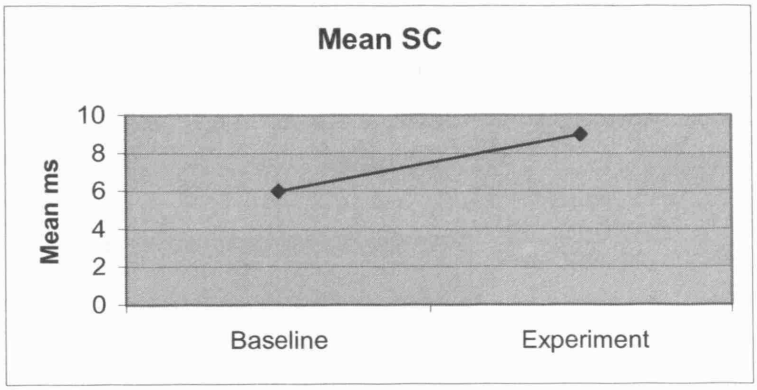



**Figure 62: Mean SC (not logged) in experiment and baseline session in experiment 6**

Within the experimental conditions, there was a significant difference in SC in response to fixed vs. variable QoS: when participants were exposed to a fixed delay to load web pages their SC was significantly lower compared to a variable delay [$F(1,31)=5.7$, $p<0.05$), as can be seen in figure 63.

[37] As the data analysis was performed by another experimenter, SC was not logged in this experiment. In addition, the baseline SC was not subtracted from responses in the conditions, as was done in experiments 1-5.

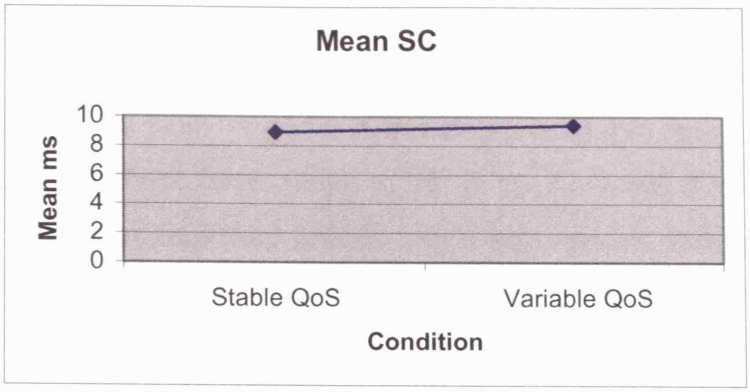



**Figure 63: Mean SC (not logged) in stable and variable QOS conditions in experiment 6**

There was a significant interaction between QoS and price [$F(1,31)=7.978$, $p<0.05$), as can be seen in figure 64. The direction of the interaction suggests that when the QoS is stable, the introduction of a stable pricing structure can increase SC. But under conditions of a variable QoS, a stable pricing structure has the effect of reducing SC.

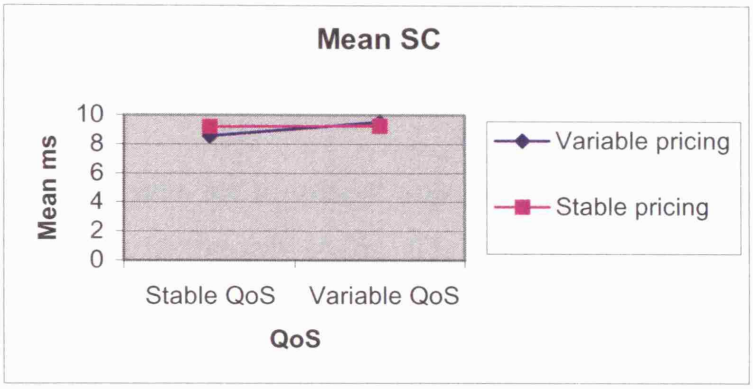



**Figure 64: Interaction between pricing and QoS for SC in experiment 6**

These results show that a variable delay increases SC and that a stable pricing structure can decrease SC when the delay is variable, but it can increase SC when the delay is stable. These results indicate that stable delays should be aimed for, regardless of the pricing structure as stable QoS always had the lower SC. With regard to pricing, consideration should be given to the types of delay users are experiencing. If they experience stable delay, then variable

pricing reduces SC, whereas if the delay is variable, stable pricing reduces SC. This effect may be because participants cannot cope with variability in both strands. If the delay is stable, they may feel that they can accept fluctuations in the price, however if the delay is variable they may feel they should pay a set amount. It may be they only think it is worth a set amount, whereas they may feel stable quality is worth more. However, the results of this study are limited, as only one physiological signal was measured.

## Experiment 7: Using physiological measures to evaluate web site designs and content

This experiment was planned and set-up by John McCarthy (UCL Computer Science department), the author of this thesis, Jens Riegelsberger (UCL Computer Science department) and Daniel Bruneau (UCL Computer Science department). The analysis of the experimental results was performed by John McCarthy. Jens Riegelsberger and Daniel Bruneau ran the experiment. The results of this study were not published.

This experiment was an empirical evaluation of the largest Internet service providers in the UK (AOL, BT, Freeserve and Tiscali). The evaluation technique used eye tracking and SC to examine web interface layout. It was hypothesised that the number of eye fixations would be positively correlated with the deviation of SC from the baseline, which would indicate participants experiencing difficulty finding the task target.

This experiment used a 4x4x4 mixed design. The between-subjects variable was order and the within-subjects variables were website and task. All participants performed all 4 tasks on each of the 4 sites. The tasks were:

1) to search for something
2) to read the latest news

3) to shop for a digital camera
4) to send an email.

The tasks did not have a time limit. The length of time to complete the task was not measured. The order of presentation of conditions was randomised. Twenty-four undergraduates from UCL took part in this experiment. There were 17 males and 5 females with a mean age of 22. All participants were experienced web users who used the Internet for at least 2 hours per day. They were paid £5 for taking part in the experiment.

There were no effects of website and task on SC. There were no correlations between the number of fixations for a trial and SC. As a second phase of investigative analysis, SC responses were filtered to leave the highest and lowest deviations from baseline[38]. These levels were then examined to determine if they were associated with a particular site-task combination. For both the highest and lowest scores, particular site-task combinations were observed with a frequency greater than expected by chance. The email task on AOL was associated with the highest SC responses and the news task on Freeserve was associated with the lowest SC responses, as can be seen in Table 18.

| | AOL | Tiscali | Freeserve | BT |
|---|---|---|---|---|
| Search | | | | |
| Email | Highest SC (4) $p<0.05$ | | | |
| News | | | Lowest SC (5) $p<0.05$ | |
| Shop | | | | |

**Table 18: Physiological results in experiment 7**

The email task on AOL received the fewest fixations of all sites, yet there was evidence of a large change in SC in nearly 20% of participants. This unexpected result was investigated further in the

eye movement data. Across all 4 tasks, the top navigation bars received the most fixations on the email task. However, for AOL the email login received the most fixations, accounting for 39% of fixations on the email task. This region also received a large number of fixations even when the task was not to write an email (accounting for 23% of fixations across all 4 tasks). It is highly likely that participants had already sampled the area extensively before having to find it. Therefore, the significant increase in SC could have been because participants knew where to look to complete the task and felt excited by this. This goes against the initial hypothesis that SC would increase with the number of eye fixations, as participants may experience difficulty finding the task target.

The results from this experiment illustrate that in some experiments examining the means may not be sophisticated enough to obtain significant results, therefore other analysis methods may have to be adopted. It also shows that an increase in the level of SC may not always be negative, as was also shown in experiment 1 where the group with better overall quality had a higher SC level. The main limitation of this experiment is that only one physiological signal was measured. If additional signals had been measured, more data would have allowed the hypothesis regarding the excitement in the AOL email task to be further investigated. Additionally, no subjective measures were taken. This would have allowed participants to report how they were feeling when performing the tasks.

There was a lost opportunity with regard to examining the length of time it took participants to complete the tasks, as this may have shown that the email task on AOL was the quickest to complete and would have tied in with the excitement hypothesis. Moreover, task performance was not measured. Occasions where participants could not complete the task could have been tied in with the physiological

---

[38] The data analysis was performed by another experimenter, therefore differs to that performed in the research reported in this thesis.

data and it would have been of interest to observe if they caused SC to increase. The next experiment moved away from web based applications to the area of VR.

## Experiment 8: Using physiological measures in a Virtual Reality application investigating the impact of avatar gaze

This experiment was undertaken as part of the PhD thesis of Maia Garau (Garau 2003). The author of this thesis advised on and assisted in the setting up of the physiological measuring equipment.

The aim of this experiment was to investigate the extent to which virtual agents are treated as social beings as they become more responsive in a VE. The task for participants was to observe their surroundings then report on what they experienced. The subjective measurements taken were presence, co-presence, the degree to which participants modified their behaviour to account for the agents being in the room and the degree of sentience attributed to the agents. SC and HR (from ECG) were the objective responses measured in this experiment.

Forty-one participants took part in the experiment: 24 were male and 17 were female. They were each paid £5. The experiment took place in a CAVE, which is *"...a room in which the user is presented with high-resolution images projected in real-time on 3 walls and the floor"* and produces *"...the illusion of 3-factor objects appearing both within and beyond the walls of the CAVE"*[39]. The experiment used a between-groups one-way design with the degree of agent responsiveness as a four level factor.

1. Agents were static, frozen in a reading pose.

[39] See http://www.cs.ucl.ac.uk/research/vr/Projects/Cave/

2. Agents were moving, as would be expected of people in a library, carrying out behaviours such as turning pages and occasionally looking around.
3. Agents were responsive to participants, depending on where the participants were in the space, and also displayed the behaviours in condition 2. The degree of each agent's visual engagement with the participant was a function of interpersonal distance: intimate, personal, social and public. In the public zone the agent was focussed on things on the table, whereas in the intimate zone the agent turned round in its chair and 'looked at' the participant.
4. The responsive behaviours were the same as in condition 3, but in this condition the agent would speak to the participant.

There were three stages to the experiment. The first involved the measuring of the baseline responses and the completion of a questionnaire on Social Avoidance and Stress. The second phase was when participants had to enter the CAVE. They first experienced the training room, which was to help participants to navigate in the CAVE. They then moved onto the virtual library and experienced one of the four conditions for a period of four minutes, during which the physiological data were marked by a researcher for significant events, such as participants moving close to agents. The final phase involved a questionnaire about presence and a semi-structured interview.

Overall, the subjective results indicated that *"...increasing agent responsiveness even on a simple level can impact on certain aspects of participant's social responses"* (Garau 2003). Due to problems with the physiological measuring equipment, data was only available for 33 participants.

| Condition | Number of participants | Training room | Library room | p value |
|---|---|---|---|---|
| Static | 8 | 75.4 (11.41) | 77.36 (12.4) | 0.450 |
| Moving | 7 | 80.37 (11.64) | 82.42 (12.33) | 0.216 |
| Responsive | 8 | 58.65 (17.88) | 64.93 (17.36) | 0.034 |
| Talking | 10 | 76.09 (10.39) | 79.49 (8.82) | 0.124 |

**Table 19: HR results in experiment 8**

| Condition | Number of participants | Training room | Library room | p value |
|---|---|---|---|---|
| Static | 8 | 7.64 (2.19) | 8.89 (2.65) | 0.008 |
| Moving | 7 | 10.89 (4.2) | 12.42 (5.43) | 0.064 |
| Responsive | 8 | 11.37 (5.52) | 13.64 (6.05) | 0.003 |
| Talking | 10 | 10.99 (5.5) | 13.13 (7.63) | 0.034 |

**Table 20: SC results in experiment 8**

Tables 19 and 20 show that HR and SC increased when participants went from the training room to the virtual library. For SC this is significant for all conditions except 'moving', although this condition was approaching significance. For HR it is only significant for 'responsive'. However, when taking the reported presence scores and details on previous computer usage into account it was found that this HR increase diminishes with computer use and with increasing reported presence in library. This is consistent with findings for co-presence and presence, therefore there was a correlation between subjective and physiological responses for HR. The marked data was not used in the analysis, although it did show some instances of SC increasing during events of importance, such as when the agent spoke to the participant.

This experiment showed that SC was an indicator of increased presence in the virtual library. It is possible that the moving condition was not significant as it had the least number of participants. HR was not found to be as effective as SC in this experiment.

## Conclusions

The results of these experiments have shown the benefits of measuring physiological responses in applications other than MMC and areas other than HCI. Experiment 6 showed significant

differences in SC, yet there were no significant differences in the subjective measure. This may be because using the QUASS slider added an additional level of difficulty in this interactive task. This gives more weight to the disadvantages of continuous subjective assessment in interactive tasks and support for utilising physiological measures in such situations. The analysis varied to that in experiments 1-5 in this thesis, as it was performed by a different researcher and consequently used raw mean data, yet this still produced significant differences. It would have been beneficial to measure additional physiological signals because experiments 1-5 in this thesis illustrate the fractionation that can occur between signals, which offers additional information on the task.

Experiment 7 again used a different method of analysis: highest and lowest deviations from the baseline were examined and produced a significant result, which could be indicative of excitement. This experiment illustrates that in some tasks, using mean physiological responses may not be fine-grained enough to produce significant differences, therefore the use of alternative data analysis techniques should be explored. This issue may be especially pertinent when the conditions are short, as opposed to the conditions in the research reported in this thesis, which were at least 2 minutes long. The measurement of additional physiological signals, such as HR, would have allowed further conclusions to be made about the nature of the task. In addition, the use of subjective data would have shown if participants were aware of where they were looking on the screen and made the evaluation approach more balanced by using subjective as well as objective measurements.

Finally, experiment 8 showed significant differences due to agent responsiveness in a VE. The difference between the design of this experiment and that of the experiments reported in this thesis is that it was between-subjects and did not use changes from the baseline. In order to use between-subjects designs in physiology, it is

recommended that the data are normalised (by subtracting the baseline and dividing by the standard deviation) so that data from each participant can be directly compared (Ward & Marsden 2003). Overall, SC seemed to be responding to the environment, thus may be useful as a between-subjects measure of presence, whereas HR did not. Examination of the raw means is not useful in this scenario, due to the problem stated above, thus no more results can be inferred from the means.

In conclusion, these experiments have shown that the use of physiological measurements has potential in area other than HCI. However, they also illustrate the need for a 3-factor evaluation approach using both objective and subjective methods and the issues that can arise in the analysis of physiological data.

# Appendix E: Raw data

## Experiment 1

| Participant | Gender | Group |
|---|---|---|
| 1 | Male | 1 |
| 2 | Female | 1 |
| 3 | Male | 1 |
| 4 | Male | 1 |
| 5 | Female | 1 |
| 6 | Female | 1 |
| 7 | Male | 1 |
| 8 | Male | 1 |
| 9 | Female | 1 |
| 10 | Female | 1 |
| 11 | Male | 1 |
| 12 | Male | 1 |
| 13 | Male | 2 |
| 14 | Male | 2 |
| 15 | Male | 2 |
| 16 | Male | 2 |
| 17 | Female | 2 |
| 18 | Female | 2 |
| 19 | Male | 2 |
| 20 | Male | 2 |
| 21 | Female | 2 |
| 22 | Female | 2 |
| 23 | Male | 2 |
| 24 | Male | 2 |

**Table 21: Gender of participants and grouping**

***Subjective data***

| Partic | Q1 | Q2 | Q3 | Q4 | Q7 | Q8 |
|---|---|---|---|---|---|---|
| 1 | 2 | 2 | 4 | 1 | 2 | 2 |
| 2 | 1 | 2 | 4 | 1 | 1 | 2 |
| 3 | 2 | 2 | 3 | 1 | 2 | 2 |
| 4 | 1 | 2 | 4 | 1 | 1 | 2 |
| 5 | 1 | 2 | 4 | 2 | 1 | 1 |
| 6 | 1 | 2 | 5 | 1 | 1 | 2 |
| 7 | 1 | 1 | 3 | 1 | 1 | 2 |
| 8 | 2 | 1 | 4 | 1 | 1 | 2 |
| 9 | 1 | 1 | 3 | 1 | 2 | 2 |
| 10 | 1 | 1 | 3 | 1 | 1 | 2 |
| 11 | 2 | 2 | 3 | 1 | 2 | 1 |
| 12 | 1 | 1 | 3 | 1 | 1 | 2 |
| 13 | 2 | 1 | 4 | 2 | 1 | 2 |
| 14 | 2 | 1 | 3 | 1 | 1 | 2 |
| 15 | 1 | 1 | 3 | 1 | 1 | 2 |
| 16 | 1 | 2 | 2 | 1 | 1 | 1 |
| 17 | 1 | 1 | 4 | 2 | 1 | 2 |
| 18 | 1 | 1 | 4 | 2 | 1 | 2 |
| 19 | 1 | 1 | 3 | 1 | 2 | 2 |
| 20 | 2 | 1 | 2 | 1 | 1 | 2 |
| 21 | 2 | 1 | 2.7 | 1 | 1 | 2 |
| 22 | 1 | 2 | 4 | 1 | 1 | 2 |
| 23 | 2 | 1 | 3 | 1 | 1 | 2 |
| 24 | 1 | 2 | 5 | 1 | 1 | 2 |

**Table 22: Subjective data from interview 1**

Key: 1 = yes, 2 = no

NB: due to a misprint on the questionnaire no useful data could be extracted from question 6.

| Partic | Q1 | Q2 | Q3 | Q4 | Q7 | Q8 | Q9 | Q10 | Q11 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 2 |
| 2 | 1 | 1 | 3 | 2 | 2 | 2 | 2 | 2 | 1 |
| 3 | 2 | 2 | 3 | 1 | 2 | 2 | 2 | 2 | 1 |
| 4 | 1 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 1 |
| 5 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 6 | 2 | 2 | 3 | 1 | 1 | 2 | 1 | 2 | 1 |
| 7 | 2 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 1 |
| 8 | 1 | 2 | 5 | 2 | 1 | 2 | 2 | 1 | 1 |
| 9 | 2 | 2 | 4 | 1 | 2 | 2 | 2 | 2 | 1 |
| 10 | 2 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 11 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 1 | 1 |
| 12 | 1 | 2 | 4.5 | 2 | 1 | 1 | 2 | 1 | 1 |
| 13 | 1 | 1 | 3 | 2 | 1 | 2 | 1 | 2 | 2 |
| 14 | 2 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 2 |
| 15 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 1 |
| 16 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 1 |
| 17 | 2 | 2 | 3 | 1 | 1 | 2 | 1 | 2 | 2 |
| 18 | 1 | 2 | 3 | 2 | 1 | 1 | 2 | 1 | 1 |
| 19 | 1 | 1 | 3.5 | 1 | 1 | 2 | 1 | 1 | 2 |
| 20 | 2 | 1 | 3 | 1 | 1 | 1 | 2 | 1 | 2 |
| 21 | 1 | 1 | 2.5 | 1 | 1 | 2 | 2 | 2 | 2 |
| 22 | 2 | 2 | 3 | 1 | 1 | 2 | 2 | 1 | 1 |
| 23 | 2 | 1 | 3 | 1 | 1 | 2 | 1 | 1 | 2 |
| 24 | 1 | 2 | 5 | 1 | 1 | 2 | 2 | 2 | 1 |

**Table 23: Subjective data from interview 2**

Key: 1 = yes, 2 = no

For question 11, 1 = interviewee 1, 2 = interviewee 2

NB: due to a misprint on the questionnaire no useful data could be extracted from question 6.

| Participant | Interview 1 | Interview 2 |
|---|---|---|
| 1 | 1 | 0 |
| 2 | 1 | 5 |
| 3 | 2 | 0 |
| 4 | 0 | 0 |
| 5 | 3 | 4 |
| 6 | 0 | 0 |
| 7 | 0 | 0 |
| 8 | 2 | 6 |
| 9 | 3 | 0 |
| 10 | 0 | 0 |
| 11 | 1 | 0 |
| 12 | 0 | 0 |
| 13 | 3 | 4 |
| 14 | 0 | 0 |
| 15 | 0 | 0 |
| 16 | 0 | 0 |
| 17 | 7 | 6 |
| 18 | 0 | 0 |
| 19 | 1 | 2 |
| 20 | 5 | 5 |
| 21 | 4 | 4 |
| 22 | 0 | 0 |
| 23 | 1 | 5 |
| 24 | 7 | 7 |

**Table 24: Responses to question 5**

Key:

0 No change
1 Frame rate changed when interviewee moved
2 Yes
3 Frame rate changes frequently
4 Frame rate changed at specific point
5 Comment on audio
6 Comment on synchrony
7 Don't know

***Physiological data***

| Partic | BI | 16fps | Time1 | Time2 | Time3 | Time4 | Time5 | Time6 |
|---|---|---|---|---|---|---|---|---|
| 1 | 8.49 | 9.33 | 9.75 | 10.53 | 10.35 | 9.81 | 10.92 | 12.53 |
| 2 | 2.38 | 2.98 | 3.12 | 3.37 | 3.71 | 1.57 | 2.33 | 3.03 |
| 3 | 7.84 | 10.49 | 10.34 | 11.65 | 12.63 | 13.08 | 14.85 | 15.29 |
| 4 | 3.29 | 3.82 | 4.27 | 4.47 | 4.84 | 5.20 | 5.79 | 5.82 |
| 5 | 8.70 | 8.78 | 7.79 | 8.05 | 8.97 | 8.12 | 8.77 | 9.06 |
| 6 | 4.09 | 6.54 | 5.71 | 6.00 | 5.46 | 4.37 | 3.81 | 4.23 |
| 7 | 1.81 | 1.85 | 0.76 | 0.72 | 0.78 | 3.98 | 1.83 | 2.03 |
| 8 | 12.76 | 14.46 | 12.88 | 12.98 | 11.74 | 10.65 | 9.90 | 10.69 |
| 9 | 11.48 | 12.86 | 10.94 | 12.83 | 11.95 | 11.50 | 12.14 | 12.28 |
| 10 | 3.47 | 5.21 | 5.29 | 6.40 | 7.22 | 8.58 | 9.22 | 9.95 |
| 11 | 5.73 | 5.78 | 6.53 | 6.87 | 6.70 | 6.00 | 5.73 | 6.31 |
| 12 | 3.53 | 4.38 | 2.94 | 3.21 | 3.60 | 3.35 | 3.58 | 4.18 |
| 13 | 6.50 | 9.62 | 10.32 | 11.42 | 11.07 | 10.94 | 11.38 | 11.47 |
| 14 | 3.85 | 5.26 | 5.08 | 5.49 | 5.43 | 5.92 | 4.92 | 7.06 |
| 15 | 9.25 | 17.71 | 18.78 | 18.74 | 18.38 | 18.51 | 18.65 | 18.08 |
| 16 | 17.44 | 19.32 | 19.91 | 19.43 | 19.00 | 17.99 | 16.87 | 17.46 |
| 17 | 2.74 | 2.51 | 2.05 | 2.68 | 2.08 | 2.92 | 2.94 | 4.35 |
| 18 | 1.36 | 3.14 | 1.28 | 1.44 | 1.49 | 3.28 | 1.66 | 1.36 |
| 19 | 8.75 | 11.13 | 11.60 | 11.97 | 12.60 | 12.49 | 13.61 | 12.95 |
| 20 | 6.26 | 6.63 | 7.47 | 8.72 | 8.67 | 7.92 | 8.02 | 9.34 |
| 21 | 2.93 | 2.73 | 2.95 | 3.20 | 3.46 | 3.32 | 3.22 | 3.47 |
| 22 | 2.42 | 3.29 | 3.18 | 2.85 | 2.92 | 3.31 | 2.70 | 3.18 |
| 23 | 9.11 | 10.91 | 11.76 | 12.08 | 12.07 | 12.61 | 11.62 | 12.71 |
| 24 | 9.58 | 11.13 | 12.96 | 12.70 | 12.84 | 12.27 | 11.59 | 11.77 |

**Table 25: SC means**

| Partic | BI | 16fps | Time1 | Time2 | Time3 | Time4 | Time5 | Time6 |
|---|---|---|---|---|---|---|---|---|
| 1 | missing | 0.39 | 0.30 | 0.39 | 0.32 | 0.26 | 1.22 | 1.00 |
| 2 | 0.35 | 0.19 | 0.18 | 0.20 | 0.35 | 0.36 | 0.38 | 0.62 |
| 3 | missing | 0.45 | 1.35 | 0.83 | 0.69 | 0.69 | 1.14 | 1.88 |
| 4 | missing | 0.12 | 0.12 | 0.22 | 0.15 | 0.26 | 0.23 | 0.28 |
| 5 | missing | 0.49 | 0.67 | 1.18 | 1.06 | 0.57 | 1.15 | 1.46 |
| 6 | 0.98 | 0.77 | 0.39 | 0.83 | 1.07 | 0.59 | 0.10 | 0.72 |
| 7 | 0.41 | 0.61 | 0.10 | 0.06 | 0.08 | 0.72 | 0.40 | 0.63 |
| 8 | missing | 1.52 | 1.21 | 1.64 | 2.10 | 2.45 | 0.94 | 1.83 |
| 9 | missing | 0.70 | 0.69 | 1.89 | 0.61 | 0.95 | 0.92 | 1.16 |
| 10 | missing | 0.97 | 0.70 | 1.07 | 1.11 | 1.13 | 0.99 | 1.96 |
| 11 | missing | 0.81 | 1.09 | 1.30 | 1.51 | 1.07 | 1.29 | 1.69 |
| 12 | 0.93 | 0.52 | 0.38 | 0.49 | 0.44 | 0.83 | 0.30 | 0.65 |
| 13 | 1.33 | 0.44 | 0.98 | 1.01 | 0.51 | 0.46 | 0.47 | 0.66 |
| 14 | 1.59 | 0.62 | 0.77 | 1.18 | 0.87 | 0.86 | 0.54 | 1.15 |
| 15 | 3.24 | 1.20 | 1.07 | 1.51 | 2.04 | 1.83 | 1.91 | 1.65 |
| 16 | 1.72 | 1.07 | 1.62 | 1.33 | 1.62 | 1.24 | 1.27 | 1.96 |
| 17 | missing | 0.46 | 0.34 | 0.90 | 0.21 | 0.45 | 0.32 | 0.92 |
| 18 | missing | 0.92 | 0.24 | 0.27 | 0.24 | 1.17 | 0.32 | 0.13 |
| 19 | missing | 0.73 | 1.09 | 1.38 | 1.13 | 1.28 | 1.37 | 1.30 |
| 20 | missing | 0.34 | 0.71 | 1.14 | 1.09 | 1.09 | 1.18 | 1.42 |
| 21 | missing | 0.24 | 0.21 | 0.17 | 0.22 | 0.31 | 0.24 | 0.28 |
| 22 | missing | 0.42 | 1.07 | 0.77 | 1.13 | 0.78 | 0.88 | 1.38 |
| 23 | 1.46 | 0.47 | 0.57 | 0.86 | 0.88 | 0.75 | 0.51 | 0.82 |
| 24 | 0.87 | 1.85 | 1.07 | 0.91 | 0.66 | 0.57 | 0.43 | 0.91 |

**Table 26: SC standard deviations**

| Partic | BI | 16fps | Time 1 | Time 2 | Time 3 | Time 4 | Time 5 | Time 6 |
|---|---|---|---|---|---|---|---|---|
| 1 | 71.69 | 70.32 | 72.09 | 73.28 | 72.63 | 68.21 | 70.54 | 68.73 |
| 2 | 96.18 | 96.03 | 96.92 | 98.34 | 97.78 | 90.77 | 94.46 | 93.71 |
| 3 | 69.04 | 65.62 | 65.42 | 67.23 | 68.93 | 66.66 | 66.32 | 65.02 |
| 4 | 82.19 | 81.50 | 85.30 | 84.80 | 82.65 | 80.28 | 78.37 | 78.08 |
| 5 | 77.99 | 73.49 | 75.83 | 75.69 | 76.61 | 72.31 | 73.58 | 71.89 |
| 6 | 87.99 | 97.85 | 99.76 | 103.15 | 106.19 | 93.86 | 96.3 | 100.36 |
| 7 | 70.43 | 74.39 | 78.68 | 76.45 | 76.13 | 72.07 | 73.88 | 72.55 |
| 8 | 93.09 | 94.50 | 93.41 | 94.70 | 94.53 | 90.79 | 91.71 | 93.68 |
| 9 | 58.73 | 72.88 | 73.44 | 74.79 | 72.33 | 72.56 | 74.99 | 75.71 |
| 10 | 101.56 | 103.15 | 104.91 | 102.89 | 105.31 | 104.89 | 108.48 | 107 |
| 11 | 79.87 | 83.58 | 77.83 | 70.90 | 77.02 | 77.33 | 77.61 | 76.71 |
| 12 | 80.77 | 77.08 | 73.08 | 74.82 | 73.66 | 72.92 | 68.67 | 70.89 |
| 13 | 74.45 | 75.51 | 74.75 | 75.48 | 73.36 | 75.75 | 73.04 | 72.25 |
| 14 | 75.53 | 72.63 | 72.26 | 73.59 | 73.61 | 70.71 | 68.12 | 69.69 |
| 15 | 71.76 | 72.31 | 78.14 | 82.24 | 85.3 | 88.33 | 89.67 | 85.31 |
| 16 | 94.19 | 84.02 | 82.6 | 82.97 | 82.19 | 74.9 | 75.73 | 77.53 |
| 17 | 57.84 | 54.05 | 55.25 | 56.95 | 56.49 | 56.35 | 57.26 | 58.23 |
| 18 | 77.20 | 87.81 | 86.31 | 86.50 | 85.16 | 85.82 | 85.36 | 85.98 |
| 19 | 76.87 | 71.66 | 74.73 | 74.43 | 77.62 | 76.03 | 77.70 | 76.66 |
| 20 | 80.28 | 74.46 | 74.45 | 73.70 | 76.49 | 72.19 | 70.53 | 74.81 |
| 21 | 72.26 | 67.30 | 67.73 | 67.45 | 66.05 | 65.59 | 65.36 | 64.70 |
| 22 | 75.65 | 77.98 | 79.85 | 78.19 | 79.47 | 72.52 | 74.32 | 75.64 |
| 23 | 69.98 | 65.29 | 64.34 | 65.20 | 66.03 | 65.91 | 66.29 | 72.80 |
| 24 | 59.96 | 71.14 | 71.09 | 70.45 | 61.71 | 67.14 | 67.88 | 65.63 |

**Table 27: HR means**

| Partic | BI | 16fps | Time 1 | Time 2 | Time 3 | Time 4 | Time 5 | Time 6 |
|---|---|---|---|---|---|---|---|---|
| 1 | Missing | 4.57 | 5.64 | 5.92 | 4.57 | 4.53 | 4.57 | 5.69 |
| 2 | 11 | 3.54 | 3.97 | 3.59 | 4.65 | 4.91 | 4.41 | 3.73 |
| 3 | Missing | 7.27 | 4.90 | 6.15 | 7.36 | 8.66 | 9.57 | 7.48 |
| 4 | Missing | 5.09 | 5.11 | 7.16 | 5.30 | 5.12 | 5.11 | 4.70 |
| 5 | Missing | 4.02 | 4.10 | 7.62 | 6.61 | 4.39 | 6.76 | 5.69 |
| 6 | 21.80 | 3.85 | 7.94 | 8.05 | 9.78 | 7.61 | 7.60 | 5.70 |
| 7 | 9.60 | 4.19 | 5.90 | 8.99 | 5.24 | 13.54 | 8.29 | 11.43 |
| 8 | Missing | 7.57 | 7.73 | 5.79 | 3.45 | 10.05 | 3.33 | 3.10 |
| 9 | 15.79 | 6.69 | 4.88 | 7.07 | 14.91 | 12.78 | 11.88 | 6.82 |
| 10 | 9.44 | 6.23 | 4.56 | 8.08 | 6.59 | 4.06 | 4.23 | 4.57 |
| 11 | 12.22 | 6.41 | 11.94 | 14.27 | 12.03 | 6.25 | 2.94 | 3.94 |
| 12 | 8.63 | 5.77 | 4.16 | 5.79 | 6.49 | 5.20 | 7.22 | 8.21 |
| 13 | 16.31 | 11.56 | 14.60 | 16.72 | 11.27 | 15.40 | 12.81 | 9.26 |
| 14 | 6.76 | 4.52 | 4.12 | 4.96 | 4.81 | 6.94 | 6.25 | 10.71 |
| 15 | 11.55 | 8.40 | 15.13 | 21.64 | 18.00 | 6.22 | 7.74 | 9.14 |
| 16 | 15.41 | 4.59 | 4.78 | 4.50 | 6.06 | 7.03 | 9.13 | 8.50 |
| 17 | Missing | 1.24 | 3.32 | 4.48 | 2.05 | 2.05 | 2.39 | 2.21 |
| 18 | Missing | 3.19 | 3.64 | 6.70 | 4.02 | 6.25 | 3.83 | 6.27 |
| 19 | Missing | 10.82 | 11.92 | 8.61 | 7.52 | 6.58 | 10.12 | 9.43 |
| 20 | Missing | 6.11 | 9.31 | 7.86 | 6.91 | 8.08 | 7.79 | 9.65 |
| 21 | Missing | 6.21 | 10.13 | 7.76 | 5.80 | 10.43 | 8.22 | 6.03 |
| 22 | Missing | 4.66 | 6.61 | 7.39 | 5.65 | 11.29 | 12.15 | 14.97 |
| 23 | 8.25 | 7.62 | 5.57 | 4.70 | 8.04 | 11.91 | 4.86 | 8.99 |
| 24 | 18.33 | 6.36 | 4.07 | 11.67 | 13.56 | 6.32 | 10.61 | 10.37 |

**Table 28: HR standard deviations**

# Experiment 2

| Participant | Gender |
|---|---|
| 1 | Male |
| 2 | Female |
| 3 | Male |
| 4 | Male |
| 5 | Male |
| 6 | Female |
| 7 | Female |
| 8 | Female |
| 9 | Male |
| 10 | Female |
| 11 | Female |
| 12 | Female |
| 13 | Male |
| 14 | Female |
| 15 | Female |
| 16 | Female |
| 17 | Male |
| 18 | Male |
| 19 | Male |
| 20 | Male |
| 21 | Male |
| 22 | Male |
| 23 | Female |

**Table 29: Gender of participants**

| Partic | Bl | Ref | 5%pl | Quiet | Bad mike | Echo | Loud | 20%pl |
|---|---|---|---|---|---|---|---|---|
| 1 | 4.69 | 5.95 | 6.59 | 7.01 | 7.04 | 6.45 | 6.45 | 6.25 |
| 2 | 3.68 | 4.86 | 4.83 | 4.89 | 4.91 | 4.95 | 4.67 | 4.65 |
| 3 | 2.6 | 6.11 | 5.37 | 6.54 | 5.32 | 6.03 | 5.91 | 6.31 |
| 4 | 5.63 | 13.13 | 16.92 | 15.13 | 13.78 | 14.67 | 16.47 | 15.75 |
| 5 | 4.3 | 9.55 | 10.78 | 9.95 | 9.31 | 11.88 | 12.41 | 10.09 |
| 6 | 11.07 | 17.64 | 18.35 | 18.16 | 18.66 | 17.75 | 20.23 | 17.85 |
| 7 | 8.42 | 13.97 | 11.67 | 12.5 | 11.73 | 12.86 | 12.92 | 14.19 |
| 8 | 3.33 | 5.01 | 4.17 | 5.14 | 4.49 | 5.63 | 5.06 | 4.98 |
| 9 | 4.56 | 7.34 | 7.24 | 6.74 | 8.03 | 8.09 | 7.31 | 8.69 |
| 10 | 3.84 | 8.18 | 7.66 | 7.85 | 6.69 | 6.1 | 6.62 | 7.84 |
| 11 | 6.43 | 8.8 | 9.59 | 9.65 | 10.25 | 9.7 | 10.05 | 9.37 |
| 12 | 1.27 | 2.43 | 2.82 | 3.33 | 3.22 | 3.21 | 3.66 | 3.22 |
| 13 | 2.77 | 4.41 | 3.65 | 3.92 | 3.85 | 3.64 | 4.6 | 4.78 |
| 14 | 7.33 | 10.1 | 11.7 | 11.76 | 10.58 | 11.27 | 11.77 | 11.5 |
| 15 | 4.14 | 4.99 | 5.1 | 4.24 | 5.38 | 5.1 | 5.95 | 5.08 |
| 16 | 11.16 | 13.05 | 15.1 | 14.8 | 15.08 | 14.58 | 15.44 | 14.35 |
| 17 | 8.48 | 11.97 | 14.83 | 13.57 | 14.87 | 15.51 | 12.93 | 14.7 |
| 18 | 6.95 | 9.71 | 13.1 | 9.32 | 11.03 | 11.71 | 11.3 | 11.4 |
| 19 | 5.05 | 7.78 | 7.32 | 7.58 | 6.99 | 8.76 | 8.26 | 7.84 |
| 20 | 11.28 | 15.77 | 17.92 | 18.01 | 19.28 | 19.01 | 19.05 | 19.72 |
| 21 | 3.66 | 6.38 | 8.39 | 7.82 | 7.8 | 7.63 | 7.59 | 7.23 |
| 22 | 4.91 | 7.07 | 7.39 | 7.8 | 7.46 | 7.81 | 7.56 | 7.26 |
| 23 | 6.29 | 9.98 | 10.52 | 9.89 | 11.25 | 11.01 | 10.96 | 11.42 |

**Table 30: SC means presentation 1**

| Partic | Bl | Ref | 5%pl | Quiet | Bad mike | Echo | Loud | 20%pl |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.76 | 0.24 | 0.71 | 0.41 | 0.55 | 0.5 | 0.38 | 0.35 |
| 2 | 0.96 | 0.19 | 0.54 | 0.43 | 0.31 | 0.43 | 0.4 | 0.54 |
| 3 | 0.67 | 0.25 | 0.43 | 0.55 | 0.24 | 0.36 | 0.52 | 0.47 |
| 4 | 1.84 | 0.36 | 0.63 | 0.4 | 0.28 | 0.29 | 0.53 | 0.42 |
| 5 | 1.42 | 0.62 | 0.67 | 0.84 | 0.85 | 0.75 | 1.51 | 0.63 |
| 6 | 2.95 | 1.06 | 1.41 | 1.25 | 1.76 | 1.27 | 2.47 | 1.77 |
| 7 | 2.65 | 1.66 | 1.52 | 2.26 | 1.79 | 2.68 | 1.62 | 2.05 |
| 8 | 0.94 | 0.32 | 0.44 | 0.61 | 0.66 | 0.79 | 0.55 | 0.71 |
| 9 | 1.06 | 0.35 | 1.15 | 0.76 | 0.8 | 0.93 | 0.68 | 0.65 |
| 10 | 2.37 | 0.61 | 1.39 | 1.29 | 1.37 | 1.51 | 1.54 | 1.77 |
| 11 | 1.56 | 0.33 | 0.74 | 0.49 | 0.62 | 0.35 | 0.96 | 0.7 |
| 12 | 0.09 | 0.37 | 0.58 | 0.33 | 0.66 | 0.37 | 0.67 | 0.31 |
| 13 | 0.85 | 0.52 | 1.13 | 1.4 | 0.95 | 1.05 | 1.05 | 1.02 |
| 14 | 2.16 | 0.38 | 0.19 | 0.46 | 0.47 | 0.29 | 0.48 | 0.29 |
| 15 | 1.1 | 0.35 | 0.42 | 0.22 | 0.34 | 0.31 | 0.58 | 0.38 |
| 16 | 1.32 | 0.33 | 0.52 | 0.43 | 0.17 | 0.33 | 0.33 | 0.33 |
| 17 | 1.87 | 0.36 | 0.47 | 0.3 | 0.86 | 0.77 | 0.57 | 0.43 |
| 18 | 2.66 | 1.17 | 1.07 | 1.12 | 0.85 | 1.67 | 1.49 | 1.17 |
| 19 | 0.86 | 0.65 | 1.63 | 1.39 | 1.89 | 1.06 | 1.3 | 1.29 |
| 20 | 3.32 | 0.32 | 0.84 | 0.92 | 1.16 | 0.95 | 0.95 | 0.81 |
| 21 | 0.53 | 0.22 | 0.39 | 0.26 | 0.35 | 0.22 | 0.21 | 0.29 |
| 22 | 1.5 | 0.3 | 0.66 | 0.58 | 0.57 | 0.5 | 0.46 | 0.45 |
| 23 | 1.69 | 0.99 | 1.3 | 0.51 | 1.06 | 0.7 | 0.77 | 1.58 |

**Table 31: SC presentation 1 standards deviations**

| Partic | Ref | 5%pl | Quiet | Bad mike | Echo | Loud | 20%pl |
|---|---|---|---|---|---|---|---|
| 1 | 6.53 | 7.53 | 7.22 | 7.18 | 7.27 | 7.39 | 6.99 |
| 2 | 5.02 | 5.12 | 6.39 | 4.85 | 5.85 | 5.91 | 5.75 |
| 3 | 7.10 | 8.56 | 7.34 | 6.26 | 6.48 | 7.25 | 7.33 |
| 4 | 16.66 | 16.89 | 16.11 | 16.17 | 16.09 | 15.44 | 16.70 |
| 5 | 11.54 | 11.13 | 11.50 | 12.44 | 12.46 | 11.86 | 11.58 |
| 6 | 20.74 | 17.35 | 18.48 | 16.05 | 17.76 | 18.90 | 18.45 |
| 7 | 12.05 | 12.38 | 16.95 | 12.56 | 13.31 | 13.91 | 10.60 |
| 8 | 6.03 | 3.99 | 3.26 | 4.05 | 4.44 | 4.62 | 4.85 |
| 9 | 7.99 | 8.41 | 8.27 | 8.69 | 8.94 | 8.42 | 8.42 |
| 10 | 5.35 | 8.29 | 6.89 | 7.72 | 6.50 | 6.59 | 5.52 |
| 11 | 10.39 | 10.71 | 10.92 | 10.30 | 10.68 | 10.43 | 9.77 |
| 12 | 3.36 | 3.83 | 3.78 | 3.53 | 3.73 | 4.02 | 3.55 |
| 13 | 5.18 | 5.89 | 4.57 | 4.12 | 5.54 | 5.12 | 6.62 |
| 14 | 11.99 | 12.05 | 12.60 | 12.68 | 11.79 | 11.77 | 12.35 |
| 15 | 4.24 | 4.67 | 4.15 | 4.75 | 4.79 | 4.94 | 4.28 |
| 16 | 15.90 | 15.26 | 15.42 | 15.16 | 15.47 | 15.63 | 15.58 |
| 17 | 15.33 | 16.19 | 16.66 | 16.24 | 15.89 | 15.78 | 15.93 |
| 18 | 13.09 | 14.13 | 12.81 | 10.14 | 14.86 | 12.30 | 11.50 |
| 19 | 8.20 | 9.70 | 10.25 | 9.43 | 9.80 | 9.62 | 10.44 |
| 20 | 17.77 | 20.22 | 18.91 | 18.65 | 18.89 | 15.85 | 20.72 |
| 21 | 8.12 | 8.57 | 9.01 | 8.40 | 8.62 | 8.70 | 8.23 |
| 22 | 8.22 | 9.37 | 7.96 | 8.94 | 8.97 | 8.50 | 7.97 |
| 23 | 10.34 | 12.14 | 10.82 | 9.07 | 11.30 | 10.52 | 9.63 |

**Table 32: SC presentation 2 means**

| Partic | Ref | 5%pl | Quiet | Bad mike | Echo | Loud | 20%pl |
|---|---|---|---|---|---|---|---|
| 1 | 0.3 | 0.7 | 0.63 | 0.43 | 0.41 | 0.26 | 0.28 |
| 2 | 0.43 | 0.42 | 0.31 | 0.24 | 0.39 | 0.35 | 0.38 |
| 3 | 0.39 | 0.37 | 0.45 | 0.47 | 0.43 | 0.45 | 0.26 |
| 4 | 0.44 | 1.13 | 0.58 | 0.7 | 0.77 | 0.72 | 0.48 |
| 5 | 0.54 | 0.76 | 0.45 | 0.54 | 0.8 | 0.96 | 0.8 |
| 6 | 1.26 | 1.91 | 0.95 | 1.61 | 2.02 | 2.19 | 1.26 |
| 7 | 2.66 | 2.4 | 4 | 1.58 | 2.22 | 2.27 | 1.56 |
| 8 | 0.86 | 0.73 | 0.27 | 0.21 | 0.73 | 0.59 | 0.43 |
| 9 | 0.68 | 0.76 | .87 | 0.52 | 0.57 | 0.68 | 0.48 |
| 10 | 1.1 | 1.04 | 1.08 | 1.17 | 1.46 | 1.39 | 1.46 |
| 11 | 0.8 | 0.8 | 0.74 | 0.78 | 0.87 | 0.7 | 0.51 |
| 12 | 0.29 | 0.46 | 0.42 | 0.39 | 0.44 | 0.68 | 0.33 |
| 13 | 0.65 | 1.27 | 1.32 | 1.23 | 1.01 | 1.77 | 1.25 |
| 14 | 0.43 | 0.49 | 0.52 | 0.66 | 0.32 | 0.37 | 0.53 |
| 15 | 0.4 | 0.45 | 0.43 | 0.57 | 0.49 | 0.49 | 0.52 |
| 16 | 0.22 | 0.17 | 0.19 | 0.18 | 0.35 | 0.44 | 0.19 |
| 17 | 0.65 | 0.65 | 0.83 | 0.91 | 0.73 | 0.59 | 0.58 |
| 18 | 1.49 | 0.98 | 1.67 | 1.04 | 1.4 | 1.83 | 1.31 |
| 19 | 2.11 | 1.26 | 1.64 | 1.47 | 1.29 | 1.68 | 1.56 |
| 20 | 1.12 | 1.27 | 0.84 | 1.14 | 1.25 | 0.9 | 1.44 |
| 21 | 0.45 | 0.48 | 0.32 | 0.33 | 0.25 | 0.34 | 0.27 |
| 22 | 0.7 | 0.74 | 0.42 | 0.56 | 0.6 | 0.94 | 0.69 |
| 23 | 0.84 | 1.47 | 1.25 | 0.7 | 1.33 | 1.8 | 1.08 |

**Table 33: SC presentation 2 standard deviations**

| Partic | BI | Ref | 5%pl | Quiet | Bad mike | Echo | Loud | 20%pl |
|---|---|---|---|---|---|---|---|---|
| 1 | 76.87 | 73.06 | 70.7 | 71.27 | 71.14 | 70.63 | 72.53 | 71.74 |
| 2 | 87.4 | 85.27 | 77.01 | 81.56 | 83.22 | 81.2 | 84.1 | 82.13 |
| 3 | 79.58 | 75.58 | 76.59 | 80.07 | 79.27 | 82.01 | 78.59 | 79.34 |
| 4 | 67.58 | 71.52 | 70.73 | 72.73 | 65.18 | 71.68 | 73.24 | 66.41 |
| 5 | 83.77 | 85.86 | 79.05 | 78.53 | 75.81 | 82.09 | 77.24 | 80.97 |
| 6 | 81.38 | 86.47 | 83.81 | 80.22 | 83.75 | 83.53 | 79.87 | 80.91 |
| 7 | 93.46 | 87.47 | 86.12 | 84.1 | 88.19 | 84.93 | 86.5 | 88.06 |
| 8 | 75.12 | 66.01 | 62.38 | 63.77 | 63.21 | 61.76 | 61.75 | 63.13 |
| 9 | 86.62 | 84.79 | 79.6 | 79.94 | 80.63 | 80.3 | 79.08 | 81.54 |
| 10 | 68.97 | 80.99 | 71 | 69.93 | 68.14 | 69.65 | 70.84 | 70.28 |
| 11 | 74.77 | 78.97 | 82.11 | 77.99 | 80.06 | 76.75 | 76.82 | 78.71 |
| 12 | 82.04 | 82.55 | 83.22 | 84.67 | 82.78 | 82.8 | 83.18 | 83.83 |
| 13 | 62.21 | 60.64 | 63.31 | 64.37 | 60.25 | 61.83 | 59.89 | 62.52 |
| 14 | 92.6 | 88.79 | 84.23 | 84.16 | 88.37 | 84.6 | 86.37 | 82.67 |
| 15 | 86.38 | 84.65 | 74.04 | 74.45 | 86.56 | 83.98 | 77.96 | 78.11 |
| 16 | 79.47 | 78.88 | 76.71 | 76.34 | 73.32 | 78.1 | 74.65 | 79.68 |
| 17 | 81.18 | 80.98 | 80.68 | 78.97 | 80.76 | 79.47 | 80.54 | 78.27 |
| 18 | 75.27 | 71.89 | 73.58 | 71.85 | 74.27 | 73.22 | 72.24 | 71.41 |
| 19 | 92.35 | 88.32 | 84.63 | 82.45 | 86.88 | 88.32 | 84.29 | 87.11 |
| 20 | 64.66 | 69.82 | 71.7 | 69.73 | 78.35 | 74.71 | 77.48 | 75.75 |
| 21 | 64.95 | 65.96 | 65.1 | 63.05 | 64.16 | 63.98 | 65.23 | 64.65 |
| 22 | 79.09 | 77.98 | 77.5 | 75.32 | 77.69 | 75.04 | 78.51 | 75.13 |
| 23 | 74.2 | 75.09 | 72.83 | 72.7 | 71.8 | 72.78 | 70.86 | 72.84 |

**Table 34: HR presentation 1 means**

| Partic | BI | Ref | 5%pl | Quiet | Bad mike | Echo | Loud | 20%pl |
|---|---|---|---|---|---|---|---|---|
| 1 | 5.25 | 3.99 | 5.26 | 4.16 | 4.04 | 3.45 | 5.73 | 3.11 |
| 2 | 7.93 | 4.98 | 5.85 | 5.06 | 5.83 | 5.89 | 5.61 | 5.51 |
| 3 | 6.78 | 4.21 | 3.52 | 5.94 | 4.89 | 3.79 | 5.76 | 4.94 |
| 4 | 18.95 | 12.27 | 11.87 | 9.71 | 16.91 | 9.85 | 12.73 | 16.15 |
| 5 | 6.01 | 7.63 | 7.29 | 6.09 | 6.35 | 6.89 | 6.55 | 6.15 |
| 6 | 6.94 | 4.23 | 4.31 | 4.96 | 6.44 | 4.39 | 6.13 | 4.23 |
| 7 | 4.84 | 5.18 | 5.92 | 3.4 | 4.35 | 4.41 | 3.76 | 3.4 |
| 8 | 7.75 | 3.22 | 2.18 | 2.94 | 1.48 | 3.85 | 4.57 | 8.37 |
| 9 | 8.56 | 5.16 | 6.96 | 5.92 | 6.27 | 7.25 | 5.35 | 6.04 |
| 10 | 3.61 | 3.97 | 4.3 | 2.27 | 1.62 | 2.49 | 3.13 | 2.38 |
| 11 | 10.65 | 4.7 | 6 | 5.18 | 6.15 | 9.04 | 6.74 | 4.09 |
| 12 | 3.62 | 3.88 | 3.78 | 2.47 | 3.44 | 3.37 | 3.4 | 4.02 |
| 13 | 5.09 | 3.01 | 2.52 | 5.6 | 3.69 | 3.69 | 3.48 | 2.53 |
| 14 | 8.57 | 8.78 | 19.44 | 7.04 | 8.54 | 10.28 | 9.39 | 19.3 |
| 15 | 6.83 | 5.31 | 4.44 | 7.16 | 3.16 | 8.18 | 5.24 | 3.29 |
| 16 | 8.55 | 9.17 | 11.45 | 9.06 | 10.97 | 8.11 | 11.56 | 7.59 |
| 17 | 4.92 | 3.29 | 4.1 | 3.41 | 4.4 | 3.88 | 3.56 | 2.61 |
| 18 | 5.55 | 3.68 | 4.25 | 3.62 | 4.79 | 4.89 | 5.31 | 3.9 |
| 19 | 6.01 | 5.27 | 5.78 | 5.33 | 7 | 8.14 | 7.05 | 4.14 |
| 20 | 6.89 | 5.92 | 6.3 | 5.72 | 7 | 8.78 | 7.78 | 6.36 |
| 21 | 8.24 | 8.21 | 13.88 | 8.7 | 7.16 | 6.26 | 13.27 | 10.21 |
| 22 | 4.48 | 3.67 | 3.37 | 3.1 | 4.39 | 4.07 | 3.98 | 2.82 |
| 23 | 5.12 | 4 | 3.53 | 3.25 | 3.32 | 3.39 | 3.47 | 5.12 |

**Table 35: HR presentation 1 standard deviations**

| Partic | Ref | 5%pl | Quiet | Bad mike | Echo | Loud | 20%pl |
|---|---|---|---|---|---|---|---|
| 1 | 70.23 | 68.48 | 70.58 | 69.98 | 71.24 | 68.33 | 69.93 |
| 2 | 80.67 | 85.22 | 83.23 | 81.01 | 79.34 | 79.98 | 81.56 |
| 3 | 81.63 | 87.8 | 84.03 | 81.24 | 83.39 | 83.05 | 84.34 |
| 4 | 67.94 | 67.34 | 67.63 | 75.94 | 71.05 | 66.88 | 76.06 |
| 5 | 83.55 | 81.54 | 81.38 | 89.24 | 83.11 | 82.38 | 84.28 |
| 6 | 84.21 | 81.71 | 83.55 | 84.74 | 84.93 | 86.65 | 84.5 |
| 7 | 84.1 | 83.82 | 79.84 | 83.25 | 82.57 | 84.27 | 84.57 |
| 8 | 61.22 | 59.45 | 58.93 | 61.06 | 60.63 | 62.09 | 60.47 |
| 9 | 83.7 | 80.02 | 80.39 | 82.05 | 83.76 | 82.81 | 85.45 |
| 10 | 71.49 | 70.11 | 69.41 | 71.03 | 72.03 | 70.16 | 71.82 |
| 11 | 83.04 | 84.83 | 87.62 | 86.64 | 85.23 | 85.2 | 83.87 |
| 12 | 84.37 | 81.52 | 83.09 | 82.78 | 84.6 | 84.28 | 83.2 |
| 13 | 64.23 | 65.64 | 63.13 | 64.68 | 63.05 | 62.55 | 63.04 |
| 14 | 90.27 | 87.38 | 94.87 | 86.88 | 86.27 | 93.99 | 91.07 |
| 15 | 77.09 | 75.38 | 75.98 | 84.95 | 77.24 | 76.71 | 77.83 |
| 16 | 74.23 | 74.09 | 75.46 | 70.13 | 72.38 | 69.69 | 74.43 |
| 17 | 80.17 | 78.51 | 80.02 | 77.75 | 79.22 | 78.91 | 80.33 |
| 18 | 74.23 | 72.23 | 73.86 | 74.45 | 74.21 | 74.63 | 75.41 |
| 19 | 88.74 | 86.11 | 87.36 | 89.71 | 90.23 | 87.12 | 88.59 |
| 20 | 76.28 | 68.35 | 76.09 | 74.14 | 72.87 | 74.68 | 74.88 |
| 21 | 65.77 | 64.99 | 66.37 | 66.78 | 66.11 | 68.33 | 65.17 |
| 22 | 78.29 | 76.94 | 75.61 | 77.07 | 76.58 | 77.54 | 76.68 |
| 23 | 71.9 | 68.22 | 71.71 | 72.21 | 68.52 | 69.45 | 67.99 |

**Table 36: HR presentation 2 means**

| Partic | Ref | 5%pl | Quiet | Bad mike | Echo | Loud | 20%pl |
|---|---|---|---|---|---|---|---|
| 1 | 4.57 | 5.87 | 5.34 | 4.6 | 5.7 | 5.64 | 6.23 |
| 2 | 10.51 | 4.04 | 8.66 | 8.15 | 12.14 | 21.97 | 12.98 |
| 3 | 5.8 | 9.95 | 8.59 | 4.79 | 8.68 | 9.56 | 7 |
| 4 | 13.49 | 19.23 | 10.96 | 7.89 | 11.46 | 16.29 | 8.16 |
| 5 | 5.07 | 4.11 | 6.29 | 3.65 | 5.23 | 6.29 | 6.67 |
| 6 | 5.84 | 5.25 | 5.27 | 3.67 | 5.43 | 7.44 | 5.19 |
| 7 | 4.54 | 6.27 | 12.55 | 6.06 | 7.18 | 7.12 | 5.47 |
| 8 | 3.34 | 3.02 | 2.71 | 8.07 | 6.17 | 12.97 | 2.71 |
| 9 | 6.58 | 5.55 | 7.48 | 6.2 | 8.78 | 6.82 | 7.53 |
| 10 | 2.82 | 2.54 | 2.71 | 3.37 | 2.86 | 2.73 | 2.8 |
| 11 | 7.63 | 8.44 | 9.02 | 6.74 | 9.96 | 9.05 | 7.88 |
| 12 | 3.7 | 4.65 | 3.81 | 3.65 | 4.6 | 4.36 | 3.89 |
| 13 | 4.09 | 4.39 | 2.89 | 2.95 | 4.33 | 3.77 | 3.56 |
| 14 | 9.31 | 15.57 | 17.45 | 16.48 | 17.98 | 12.61 | 13.11 |
| 15 | 5.42 | 3.92 | 4.76 | 9.39 | 5.04 | 3.25 | 4.4 |
| 16 | 10.37 | 9.12 | 10.68 | 16.41 | 11.08 | 9.54 | 8.59 |
| 17 | 3.36 | 4.73 | 11.01 | 5.46 | 6.29 | 5.6 | 7.56 |
| 18 | 4.95 | 7.01 | 5.16 | 4.29 | 5.07 | 4.4 | 3.72 |
| 19 | 4.54 | 5.93 | 5.48 | 3.9 | 7.97 | 7.43 | 4.6 |
| 20 | 8.29 | 6.9 | 6.44 | 8.09 | 7.85 | 6.63 | 5.34 |
| 21 | 6.8 | 5.79 | 9.31 | 6.25 | 8.11 | 7.91 | 7.37 |
| 22 | 4.09 | 4.86 | 3.95 | 3.52 | 4.84 | 3.64 | 4.23 |
| 23 | 2.92 | 6.11 | 7.11 | 8.9 | 6.21 | 7.27 | 6.51 |

**Table 37: HR presentation 2 standard deviations**

| Partic | Bl | Ref | 5%pl | Quiet | Bad mike | Echo | Loud | 20%pl |
|---|---|---|---|---|---|---|---|---|
| 1 | 48.61 | 45.38 | 48.81 | 48.72 | 48.75 | 52.02 | 49.6 | 45.87 |
| 2 | 47.42 | 41.52 | 40.67 | 45.44 | 40.45 | 42.85 | 42.36 | 41.71 |
| 3 | 49.8 | 40.27 | 46.28 | 46.1 | 44.52 | 44.87 | 40.37 | 46.62 |
| 4 | 66.53 | 68.14 | 66.75 | 65.81 | 69.95 | 68.01 | 66.37 | 69.68 |
| 5 | 54.55 | 43.88 | 48.6 | 52.96 | 50.12 | 45.34 | 50.44 | 48.24 |
| 6 | 49.05 | 45.26 | 51.26 | 45.7 | 45.9 | 50.6 | 48.27 | 45.18 |
| 7 | 51.54 | 39.3 | 39.08 | 37.89 | 38.24 | 38.74 | 42.44 | 36.98 |
| 8 | 49.11 | 39.78 | 46.09 | 48.19 | 47.53 | 43.26 | 42.66 | 47.39 |
| 9 | 62.09 | 50.74 | 58.91 | 53.96 | 55.95 | 49.01 | 54.4 | 47.51 |
| 10 | 48.9 | 33.22 | 37.19 | 40.54 | 38.5 | 39.25 | 39.05 | 39.93 |
| 11 | 44.73 | 43.59 | 45.65 | 41.7 | 42.17 | 39.49 | 40.46 | 43.09 |
| 12 | 33.91 | 26.31 | 30.4 | 27.38 | 28.34 | 29.73 | 27.96 | 27.04 |
| 13 | 59.47 | 51.09 | 54.6 | 51.43 | 53.02 | 52.04 | 49 | 49.04 |
| 14 | 29.58 | 62.24 | 44.98 | 45.51 | 39.52 | 56.27 | 38.72 | 43.89 |
| 15 | 38.39 | 25.03 | 24.73 | 24.9 | 24.97 | 24.82 | 24.75 | 24.85 |
| 16 | 43.84 | 32.9 | 31.14 | 32.15 | 32.64 | 35.18 | 30.88 | 31.26 |
| 17 | 58.38 | 47.35 | 54.39 | 55.25 | 51.4 | 49.39 | 48.16 | 51.86 |
| 18 | 52.7 | 43.81 | 42.67 | 42.91 | 40.91 | 40.5 | 41.44 | 42.49 |
| 19 | 51.68 | 40.67 | 43.44 | 41.64 | 46.23 | 40.82 | 40.03 | 45.38 |
| 20 | 49.41 | 52.61 | 55.88 | 57.67 | 50.39 | 52.97 | 48.61 | 53.84 |
| 21 | 30.1 | 29.22 | 29.87 | 28.61 | 29.49 | 28.92 | 28.9 | 29.18 |
| 22 | 41.42 | 31.97 | 33.01 | 37.53 | 35.09 | 35.09 | 36.66 | 34.65 |
| 23 | 56.06 | 40.75 | 43.84 | 43.95 | 41.62 | 41.74 | 41.49 | 40.89 |

**Table 38: BVP presentation 1 means**

| Partic | Bl | Ref | 5%pl | Quiet | Bad mike | Echo | Loud | 20%pl |
|---|---|---|---|---|---|---|---|---|
| 1 | 6.29 | 4.08 | 7.01 | 6.08 | 5.23 | 5.86 | 6.19 | 3.2 |
| 2 | 6.57 | 2.49 | 2.77 | 2.57 | 2.81 | 2.71 | 3.37 | 2.57 |
| 3 | 8.65 | 3.33 | 1.92 | 4.68 | 4.23 | 3.23 | 4.51 | 4.21 |
| 4 | 10.75 | 5.28 | 6.61 | 6.31 | 5.34 | 4.68 | 6.99 | 3.62 |
| 5 | 3.9 | 3.75 | 5.78 | 2.46 | 4.04 | 3.84 | 5.18 | 3.38 |
| 6 | 6.26 | 6.58 | 3.93 | 4.34 | 5.98 | 4.69 | 5.42 | 4.47 |
| 7 | 5.23 | 2.16 | 2.11 | 2.9 | 2.98 | 2.62 | 1.88 | 2.81 |
| 8 | 6.47 | 2.32 | 3.73 | 2.98 | 3.42 | 7.33 | 3.72 | 5.26 |
| 9 | 7.81 | 6.78 | 6.4 | 6.74 | 6.04 | 7.65 | 5.53 | 5.63 |
| 10 | 4.31 | 1.2 | 2.14 | 2.55 | 1.78 | 1.79 | 2.6 | 1.85 |
| 11 | 8.72 | 2.47 | 3.01 | 3.76 | 3.5 | 4.8 | 5.46 | 2.18 |
| 12 | 3.44 | 0.63 | 1.64 | 1.42 | 2.24 | 1.26 | 0.86 | 0.73 |
| 13 | 6.97 | 4.1 | 1.81 | 3.66 | 4.15 | 4.26 | 6.15 | 4.31 |
| 14 | 1.62 | 7.8 | 9.29 | 11.76 | 10.68 | 12.74 | 10.21 | 9.91 |
| 15 | 6.72 | 0.14 | 0.08 | 0.09 | 0.12 | 0.09 | 0.09 | 0.07 |
| 16 | 5.63 | 1.59 | 2.05 | 2.17 | 2.46 | 3 | 1.34 | 1.1 |
| 17 | 7.87 | 4.7 | 2.72 | 4.02 | 5.76 | 6.81 | 5.92 | 5.91 |
| 18 | 5.19 | 2.81 | 1.01 | 3.29 | 3.28 | 2.03 | 2.57 | 1.94 |
| 19 | 5.68 | 3.56 | 3.36 | 4.17 | 4.32 | 3.62 | 4.03 | 3.38 |
| 20 | 7.43 | 3.37 | 4.17 | 4.38 | 5.13 | 6.22 | 4.57 | 4.38 |
| 21 | 1.13 | 0.51 | 0.73 | 0.52 | 0.49 | 0.48 | 0.75 | 0.48 |
| 22 | 5.05 | 2.17 | 1.86 | 3.23 | 2.62 | 2.41 | 3.32 | 1.86 |
| 23 | 5.78 | 2.72 | 2.69 | 5.4 | 3.61 | 4.51 | 4.09 | 4.7 |

**Table 39: BVP presentation 1 standard deviations**

| Partic | Ref | 5%pl | Quiet | Bad mike | Echo | Loud | 20%pl |
|---|---|---|---|---|---|---|---|
| 1 | 41.33 | 40.13 | 38.41 | 40.83 | 43.73 | 36.03 | 42 |
| 2 | 38.33 | 41.82 | 38.23 | 37.79 | 34.43 | 30.82 | 35.16 |
| 3 | 43.95 | 40.17 | 39.18 | 43.67 | 42.86 | 38.7 | 42.13 |
| 4 | 69.78 | 67.66 | 68.61 | 63.68 | 68.12 | 69.26 | 63.49 |
| 5 | 51.05 | 51.61 | 52.66 | 48.64 | 47.4 | 47.33 | 49.8 |
| 6 | 51.3 | 49.2 | 55.76 | 48.11 | 55.42 | 40.17 | 51.88 |
| 7 | 37.44 | 32.14 | 31.86 | 35.07 | 33.35 | 35.42 | 36.24 |
| 8 | 46.17 | 43.36 | 39.04 | 37.06 | 42.77 | 36.87 | 44.38 |
| 9 | 48.13 | 51.88 | 49.86 | 49.89 | 47.67 | 46.91 | 47 |
| 10 | 37.24 | 35.78 | 39.79 | 39.7 | 35.19 | 41.1 | 37.43 |
| 11 | 40.77 | 38.03 | 40.61 | 38.5 | 42.22 | 37.26 | 39.57 |
| 12 | 26.69 | 26.28 | 27.08 | 26.8 | 27.96 | 26.64 | 26.98 |
| 13 | 46.09 | 48.08 | 49.63 | 51.57 | 46.82 | 51.67 | 49.09 |
| 14 | 48.17 | 37.87 | 38.45 | 38.49 | 41.43 | 29.82 | 38.45 |
| 15 | 24.64 | 24.82 | 24.81 | 24.72 | 24.7 | 24.76 | 25.05 |
| 16 | 29.92 | 29.87 | 30.13 | 30.12 | 29.46 | 30.58 | 29.41 |
| 17 | 50.54 | 50.14 | 46.79 | 44.52 | 47.6 | 46.2 | 47.44 |
| 18 | 39.27 | 40.01 | 37.83 | 41.31 | 38.07 | 39.95 | 40.19 |
| 19 | 45.88 | 48.42 | 47.09 | 47.89 | 45 | 47.77 | 46.95 |
| 20 | 57.82 | 51.85 | 53.33 | 55.99 | 52.77 | 51.35 | 47.56 |
| 21 | 29.57 | 31.46 | 31.88 | 31.94 | 30.07 | 31.49 | 30.53 |
| 22 | 37.04 | 32.93 | 34.53 | 35.39 | 32.66 | 34.34 | 37.07 |
| 23 | 44.2 | 38.87 | 39.85 | 42.9 | 40.38 | 39.63 | 40.35 |

**Table 40: BVP presentation 2 means**

| Partic | Ref | 5%pl | Quiet | Bad mike | Echo | Loud | 20%pl |
|---|---|---|---|---|---|---|---|
| 1 | 4.73 | 4.65 | 4.54 | 3.71 | 4.21 | 3.96 | 5.01 |
| 2 | 4.1 | 3.38 | 2.61 | 5.11 | 2.87 | 2.36 | 3.04 |
| 3 | 3.81 | 4.81 | 4.57 | 4.18 | 5.43 | 4.03 | 3.98 |
| 4 | 4.79 | 6.94 | 5.11 | 6.75 | 6.35 | 5.4 | 7.57 |
| 5 | 3.01 | 2.42 | 4.3 | 3.33 | 5.08 | 5.11 | 4.98 |
| 6 | 7.72 | 4.27 | 5.26 | 3.48 | 4.99 | 4.63 | 4.29 |
| 7 | 1.91 | 2.24 | 1.84 | 2.51 | 2.19 | 2.72 | 3.13 |
| 8 | 4.16 | 4.78 | 4.02 | 2.46 | 6.25 | 3.42 | 4.26 |
| 9 | 7.07 | 7.15 | 6.52 | 6.58 | 6.79 | 6.25 | 5.56 |
| 10 | 2.4 | 1.73 | 1.7 | 2.05 | 1.78 | 1.95 | 1.32 |
| 11 | 3.72 | 3.94 | 3.9 | 2.6 | 5.48 | 4.79 | 5.23 |
| 12 | 0.64 | 0.71 | 0.92 | 1.11 | 1.97 | 0.89 | 0.78 |
| 13 | 4.37 | 6.24 | 4.05 | 3.37 | 4.2 | 3.86 | 5.2 |
| 14 | 6.01 | 6.61 | 7.6 | 10.18 | 6.84 | 2.14 | 4.74 |
| 15 | 0.07 | 0.09 | 0.11 | 0.13 | 0.11 | 0.07 | 0.12 |
| 16 | 1.27 | 1.46 | 1.45 | 1.28 | 0.82 | 1.61 | 0.99 |
| 17 | 6.03 | 5.58 | 6.07 | 8.26 | 5.23 | missing | 7.82 |
| 18 | 3.68 | 2.01 | 2.87 | 1.34 | 3.14 | 2.55 | 2.2 |
| 19 | 3.12 | 5.61 | 5.06 | 4.09 | 6 | 6.12 | 6.32 |
| 20 | 4.45 | 5.84 | 6.53 | 4.55 | 5.57 | 6.22 | 2.56 |
| 21 | 0.35 | 1.15 | 1 | 0.83 | 0.75 | 0.82 | 0.76 |
| 22 | 2.36 | 2.07 | 1.71 | 2.6 | 1.38 | 2.3 | 3.98 |
| 23 | 3.46 | 3.69 | 3.86 | 4.12 | 3.54 | 4.11 | 3.77 |

**Table 41: BVP presentation 2 standard deviations**

# Experiment 3

## *Subjective data*

| Participant | Loud | 20% pl | Bad mike | 5% pl |
|---|---|---|---|---|
| 1 | 64 | 45 | 44 | 67 |
| 2 | 45 | 31 | 32 | 41 |
| 3 | 90 | 55 | 44 | 34 |
| 4 | 73 | 75 | 83 | 46.73* |
| 5 | 20 | 20 | 40 | 25 |
| 6 | 65 | 63 | 70 | 70 |
| 7 | 36 | 25 | 52 | 35 |
| 8 | 60 | 40 | 45 | 50 |
| 9 | 80 | 70 | 64 | 54 |
| 10 | 12 | 43 | 5 | 44 |
| 11 | 80 | 20 | 40 | 60 |
| 12 | 86 | 65 | 55 | 34 |
| 13 | 30 | 5 | 20 | 61 |
| 14 | 65 | 16 | 34 | 65 |
| 15 | 56 | 24 | 69 | 41 |
| 16 | 80 | 45 | 96 | 34 |
| 17 | 70 | 30 | 40 | 50 |
| 18 | 70 | 20 | 30 | 30 |
| 19 | 60 | 30 | 50 | 50 |
| 20 | 65 | 17 | 35 | 49 |
| 21 | 25 | 15 | 40 | 47 |
| 22 | 80 | 50 | 70 | 60 |
| 23 | 75 | 25 | 65 | 30 |

**Table 42: Question 1 responses**

*= Missing data so replaced with mean of condition.

| Participant | Loud | 20% pl | Bad mike | 5% pl |
|---|---|---|---|---|
| 1 | 75 | 35 | 35 | 75 |
| 2 | 65 | 21 | 11 | 71 |
| 3 | 90 | 50 | 50 | 40 |
| 4 | 63 | 74 | 83 | 42 |
| 5 | 35 | 1 | 20 | 35 |
| 6 | 95 | 70 | 70 | 80 |
| 7 | 65 | 24 | 34 | 67 |
| 8 | 65 | 30 | 45 | 54 |
| 9 | 90 | 80 | 40 | 64 |
| 10 | 22 | 25 | 3 | 54 |
| 11 | 90 | 50 | 40 | 80 |
| 12 | 86 | 35 | 26 | 35 |
| 13 | 30 | 10 | 30 | 71 |
| 14 | 74 | 20 | 53 | 72 |
| 15 | 67 | 18 | 79 | 31 |
| 16 | 95 | 70 | 100 | 40 |
| 17 | 95 | 70 | 70 | 70 |
| 18 | 70 | 30 | 30 | 30 |
| 19 | 70 | 30 | 40 | 50 |
| 20 | 72 | 12 | 35 | 50 |
| 21 | 65 | 20 | 80 | 50.4* |
| 22 | 90 | 60 | 70 | 70 |
| 23 | 72.8* | 10 | 65 | 20 |

**Table 43: Question 2 responses**

*= Missing data so replaced with mean of condition.

| Participant | Loud | Bad mike | 20% packet loss | 5% packet loss |
|---|---|---|---|---|
| 1 | Y | Y | Y | N |
| 2 | Y | Y | Y | Didn't answer |
| 3 | Y | N | Y | Y |
| 4 | Y | Y | Y | N |
| 5 | Y | Y | N | Y |
| 6 | N | Y | Y | Y |
| 7 | Y | Y | Y | Y |
| 8 | Y | Y | Y | Y |
| 9 | Y | Y | Y | Y |
| 10 | Y | Y | Y | Y |
| 11 | Y | Y | Y | Y |
| 12 | Y | Y | Y | Y |
| 13 | Y | Y | Y | Y |
| 14 | Y | Y | Y | N |
| 15 | Y | N | Y | Y |
| 16 | Y | Y | Y | Y |
| 17 | N | Y | Y | Y |
| 18 | Y | Y | Y | Y |
| 19 | Y | N | N | Y |
| 20 | Y | Y | N | Y |
| 21 | Y | Y | Didn't answer | N |
| 22 | Y | Y | Y | Y |
| 23 | N | Y | N | Y |

**Table 44: Question 3 responses**

| Participant | Loud | 20% pl | Bad mike | 5% pl |
|---|---|---|---|---|
| 1 | 75 | 55 | 54 | 75 |
| 2 | 50 | 62 | 51 | 64.09* |
| 3 | 84 | 90 | 90 | 93 |
| 4 | 82 | 83 | 83 | 74 |
| 5 | 35 | 30 | 30 | 35 |
| 6 | 54 | 74 | 74 | 80 |
| 7 | 44 | 74 | 65 | 69 |
| 8 | 70 | 40 | 30 | 40 |
| 9 | 77 | 85 | 60 | 80 |
| 10 | 11 | 23 | 23 | 33 |
| 11 | 50 | 80 | 70 | 50 |
| 12 | 36 | 55 | 55 | 76 |
| 13 | 65 | 80 | 70 | 100 |
| 14 | 34 | 75 | 71 | 61 |
| 15 | 43 | 52 | 51 | 69 |
| 16 | 75 | 40 | 90 | 70 |
| 17 | 90 | 90 | 90 | 85 |
| 18 | 80 | 80 | 20 | 80 |
| 19 | 60 | 70 | 50 | 60 |
| 20 | 62 | 57 | 65 | 80 |
| 21 | 35 | 20 | 60 | 74.5* |
| 22 | 90 | 80 | 80 | 100 |
| 23 | 80 | 50 | 70 | 40 |

**Table 45: Question 4 responses**

*= Missing data so replaced with mean of condition.

| Participant | Loud | 20% pl | Bad mike | 5% pl |
|---|---|---|---|---|
| 1 | 54 | 87 | 25 | 77 |
| 2 | 64 | 72 | 61 | 72.09* |
| 3 | 100 | 100 | 100 | 100 |
| 4 | 84 | 83 | 84 | 73 |
| 5 | 40 | 15 | 40 | 30 |
| 6 | 75 | 85 | 80 | 85 |
| 7 | 53 | 74 | 84 | 75 |
| 8 | 80 | 50 | 50 | 64 |
| 9 | 90 | 85 | 80 | 90 |
| 10 | 5 | 77.27* | 34 | 34 |
| 11 | 90 | 90 | 90 | 90 |
| 12 | 25 | 76 | 1 | 75 |
| 13 | 90 | 100 | 87 | 100 |
| 14 | 33 | 73 | 73 | 69 |
| 15 | 48 | 61 | 47 | 69 |
| 16 | 95 | 90 | 90 | 70 |
| 17 | 100 | 100 | 100 | 100 |
| 18 | 70 | 80 | 20 | 80 |
| 19 | 70 | 70 | 50 | 70 |
| 20 | 77 | 66 | 85 | 90 |
| 21 | 25 | 55 | 70 | 80.8* |
| 22 | 100 | 90 | 80 | 100 |
| 23 | 80 | 50 | 70 | 60 |

**Table 46: Question 5 responses**

*= Missing data so replaced with mean of condition.

| Participant | Loud | Bad mike | 20%pl | 5% pl |
|---|---|---|---|---|
| 1 | N | Y | N | N |
| 2 | N | N | N | Y |
| 3 | N | N | Y | N |
| 4 | N | N | N | N |
| 5 | N | N | N | N |
| 6 | N | N | N | N |
| 7 | N | N | N | N |
| 8 | N | Y | N | Y |
| 9 | N | N | N | N |
| 10 | N | N | N | N |
| 11 | N | Y | N | N |
| 12 | N | N | N | N |
| 13 | Y | N | N | Y |
| 14 | Y | N | Y | N |
| 15 | Y | N | Y | N |
| 16 | N | N | N | Y |
| 17 | N | N | N | N |
| 18 | N | Y | Y | Y |
| 19 | Y | N | N | N |
| 20 | N | N | N | N |
| 21 | Y | N | Didn't answer | N |
| 22 | N | N | N | N |
| 23 | N | Y | N | N |

**Table 47: Question 6 responses**

| Participant | Loud | Bad mike | 20% pl | 5% pl |
|---|---|---|---|---|
| 1 | Y | Y | Y | N |
| 2 | Y | Y | N | N |
| 3 | Y | Y | N | Y |
| 4 | Don't know | Y | Y | N |
| 5 | Didn't answer | Didn't answer | Didn't answer | Didn't answer |
| 6 | Y | Y | Y | N |
| 7 | Y | Y | Y | N |
| 8 | Y | N | Y | N |
| 9 | Y | Y | Y | N |
| 10 | Y | Y | Y | Y |
| 11 | Y | Y | Don't know | N |
| 12 | Y | Y | Y | N |
| 13 | Y | Y | Y | N |
| 14 | No answer | No answer | No answer | No answer |
| 15 | Y | N | Y | N |
| 16 | Y | Don't know | Don't know | N |
| 17 | Y | Y | N | Don't know |
| 18 | Y | Y | Don't know | N |
| 19 | Y | N | Y | Don't know |
| 20 | Y | N | Y | N |
| 21 | Y | N | Y | N |
| 22 | Y | Y | Y | N |
| 23 | N | Y | Y | N |

**Table 48: Question 7responses**

***Physiological data***

| Partic | Bl | Normal | Loud | Normal | 20%pl | Normal | Bad mike | Normal | 5%pl |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2.52 | 2.96 | 3.44 | 3.82 | 3.30 | 3.98 | 3.47 | 4.66 | 4.88 |
| 2 | 9.11 | 11.21 | 12.11 | 12.90 | 13.33 | 11.90 | 13.22 | 13.31 | 13.42 |
| 3 | 1.77 | 1.81 | 1.82 | 2.07 | 2.00 | 2.55 | 2.61 | Missing* | Missing* |
| 4 | 1.79 | 4.09 | 4.61 | 3.12 | 3.45 | 5.90 | 6.03 | 5.19 | 5.35 |
| 5 | 1.95 | 2.54 | 2.74 | 2.40 | 2.93 | 3.05 | 3.81 | 2.72 | 1.82 |
| 6 | 5.10 | 7.40 | 8.07 | 7.11 | 6.75 | 7.43 | 6.70 | 7.99 | 8.35 |
| 7 | 7.74 | 9.35 | 9.60 | 9.10 | 9.26 | 9.88 | 9.96 | 9.72 | 9.94 |
| 8 | 1.98 | 2.77 | 2.69 | 2.87 | 3.17 | 3.62 | 3.32 | 3.31 | 2.89 |
| 9 | 5.42 | 7.77 | 7.79 | 7.91 | 7.24 | 6.17 | 6.37 | 6.85 | 6.90 |
| 10 | 7.89 | 7.03 | 5.03 | 4.65 | 3.72 | 6.61 | 5.16 | 9.51 | 7.39 |
| 11 | 3.74 | 6.96 | 7.97 | 5.30 | 5.56 | 5.03 | 4.50 | 5.18 | 4.09 |
| 12 | 2.95 | 4.43 | 5.45 | 4.17 | 5.69 | 5.48 | 5.55 | 5.23 | 7.39 |
| 13 | 4.94 | 6.10 | 6.42 | 7.32 | 7.68 | 7.22 | 7.61 | 6.98 | 7.90 |
| 14 | 6.13 | 7.19 | 7.76 | 8.32 | 8.45 | 10.04 | 9.85 | 10.61 | 10.46 |
| 15 | 2.44 | Missing* | Missing* | 4.63 | 4.00 | 5.05 | 4.27 | 5.57 | 5.34 |
| 16 | 2.67 | 5.93 | 5.34 | 4.44 | 3.66 | 6.68 | 6.42 | 6.31 | 5.98 |
| 17 | 2.48 | 4.98 | 4.28 | 4.48 | 3.66 | 5.18 | 4.68 | 5.49 | 6.24 |
| 18 | 0.49 | 0.47 | 0.46 | 0.44 | 0.45 | 0.56 | 0.57 | 0.48 | 0.56 |
| 19 | 6.03 | 8.32 | 7.85 | 8.94 | 8.55 | 7.36 | 7.27 | 8.09 | 8.36 |
| 20 | 3.84 | 7.80 | 7.67 | 7.45 | 7.50 | 5.82 | 5.56 | 6.98 | 7.08 |
| 21 | 9.97 | 13.94 | 14.99 | 13.48 | 13.83 | 13.33 | 13.35 | 15.80 | 16.01 |
| 22 | 3.02 | 9.29 | 6.83 | 5.55 | 6.98 | 7.41 | 7.01 | 5.49 | 5.21 |
| 23 | 17.31 | 24.38 | 23.57 | 23.27 | 23.43 | 22.36 | 22.50 | 21.43 | 21.62 |

**Table 49: SC experiment 3 means**

*Missing so replaced with overall mean for condition

| Partic | Bl | Normal | Loud | Normal | 20%pl | Normal | Bad mike | Normal | 5%pl |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.17 | 0.63 | 0.45 | 0.77 | 0.61 | 1.15 | 0.67 | 0.54 | 0.71 |
| 2 | 3.71 | 1.34 | 1.65 | 1.81 | 1.60 | 2.10 | 1.58 | 1.20 | 1.10 |
| 3 | 0.36 | 0.27 | 0.22 | 0.17 | 0.14 | 0.35 | 0.19 | Missing | Missing |
| 4 | 0.39 | 0.22 | 0.32 | 0.14 | 0.29 | 0.38 | 0.49 | 0.23 | 0.42 |
| 5 | 0.45 | 0.53 | 0.38 | 0.30 | 0.25 | 0.59 | 0.90 | 0.49 | 0.08 |
| 6 | 1.73 | 0.75 | 1.07 | 0.52 | 0.64 | 1.91 | 1.34 | 0.89 | 1.28 |
| 7 | 1.60 | 0.50 | 0.71 | 0.51 | 0.72 | 0.52 | 0.64 | 0.56 | 0.83 |
| 8 | 1.07 | 0.27 | 0.33 | 0.24 | 0.59 | 0.40 | 0.36 | 0.64 | 0.88 |
| 9 | 1.54 | 0.55 | 0.59 | 0.53 | 0.12 | 0.54 | 0.57 | 0.57 | 0.30 |
| 10 | 2.08 | 1.72 | 0.45 | 0.35 | 0.24 | 0.45 | 0.54 | 1.02 | 0.34 |
| 11 | 1.30 | 0.71 | 0.86 | 0.24 | 0.39 | 0.40 | 0.10 | 0.36 | 0.61 |
| 12 | 2.84 | 1.46 | 1.59 | 1.12 | 2.32 | 2.17 | 1.27 | 1.54 | 1.71 |
| 13 | 1.15 | 0.34 | 0.67 | 0.18 | 0.30 | 0.25 | 0.46 | 0.41 | 0.86 |
| 14 | 1.56 | 0.45 | 0.63 | 0.25 | 0.50 | 1.16 | 0.70 | 1.16 | 0.76 |
| 15 | 1.80 | Missing | Missing | 1.28 | 1.03 | 1.70 | 1.14 | 1.57 | 1.91 |
| 16 | 0.80 | 0.59 | 0.37 | 0.41 | 0.41 | 0.57 | 0.43 | 0.52 | 0.37 |
| 17 | 0.97 | 0.92 | 1.03 | 0.77 | 0.93 | 2.04 | 1.09 | 1.38 | 1.15 |
| 18 | 0.10 | 0.01 | 0.01 | 0.01 | 0.01 | 0.03 | 0.05 | 0.01 | 0.03 |
| 19 | 1.24 | 0.76 | 0.83 | 0.60 | 0.51 | 0.26 | 0.51 | 0.65 | 0.55 |
| 20 | 0.84 | 0.77 | 0.62 | 0.42 | 0.59 | 0.47 | 0.74 | 0.69 | 0.53 |
| 21 | 2.69 | 1.05 | 1.26 | 1.32 | 0.93 | 1.05 | 0.59 | 1.38 | 0.71 |
| 22 | 1.46 | 1.52 | 1.42 | 0.68 | 0.76 | 0.57 | 0.45 | 0.83 | 0.39 |
| 23 | 1.77 | 1.52 | 1.49 | 1.79 | 1.50 | 1.53 | 1.13 | 1.02 | 1.00 |

**Table 50: SC standard deviations**

| Partic | Bl | Normal | Loud | Normal | 20%pl | Normal | Bad mike | Normal | 5%pl |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 76.92 | 80.22 | 80.69 | 75.67 | 76.06 | 72.58 | 76.03 | 72.99 | 76.64 |
| 2 | 74.78 | 74.4 | 77.57 | 76.56 | 80.3 | 76.42 | 76.84 | 74.94 | 74.24 |
| 3 | 68.45 | 68.51 | 68.75 | 70.63 | 70.61 | 72.73 | 71.2 | Missing* | Missing* |
| 4 | 54.54 | 49.23 | 48.5 | 48.7 | 49.71 | 52.93 | 49.94 | 48.59 | 52.7 |
| 5 | 78.66 | 81.86 | 84.68 | 71.79 | 75.03 | 83.77 | 83.69 | 86.58 | 89.14 |
| 6 | 96.3 | 90.01 | 92.08 | 91.23 | 92.53 | 80.84 | 87.29 | 86.19 | 86.88 |
| 7 | 101.58 | 99.51 | 99.63 | 96.61 | 99.4 | 100.64 | 102.6 | 95.52 | 103.34 |
| 8 | 75.75 | 67.66 | 75.62 | 71.04 | 69.75 | 72.3 | 74.17 | 72.31 | 73.83 |
| 9 | 72.15 | 69.94 | 69.75 | 67.68 | 67.1 | 67.27 | 70.45 | 69.9 | 68.99 |
| 10 | 61.13 | 59.9 | 58.59 | 59.9 | 54.87 | 58.69 | 59.23 | 55.74 | 55.15 |
| 11 | 82.05 | 91.86 | 89.58 | 74.18 | 74 | 78.38 | 72.17 | 79.31 | 78.84 |
| 12 | 77.83 | 74.62 | 77.95 | 75.41 | 77.89 | 76.58 | 76.84 | 75.52 | 78.87 |
| 13 | 60.41 | 60.66 | 58.38 | 59.96 | 60.85 | 58.66 | 60.01 | 61.74 | 60.77 |
| 14 | 94.57 | 90.5 | 91.57 | 92.7 | 92.98 | 92.06 | 90.57 | 86.26 | 86.51 |
| 15 | 72.95 | Missing* | Missing* | 66.75 | 68.22 | 64 | 60.17 | 62.04 | 64.39 |
| 16 | 71.3 | 68.74 | 67.34 | 62.56 | 67.22 | 67.54 | 66.96 | 66.83 | 66.98 |
| 17 | 66.16 | 62.67 | 63.41 | 61.19 | 64.05 | 61.76 | 59.09 | 60.99 | 59.81 |
| 18 | 60.54 | 149.62 | 150 | 41.38 | 41.38 | 71.7 | 74.05 | 71.33 | 71.06 |
| 19 | 88.5 | 83.48 | 85.74 | 80.36 | 81.85 | 85.63 | 88.66 | 84.67 | 86.27 |
| 20 | 77.25 | 73.66 | 73.67 | 70.51 | 74.24 | 75.09 | 72.85 | 72.94 | 73.91 |
| 21 | 59.62 | 72.27 | 70.09 | 71.39 | 71.19 | 66.46 | 62.84 | 67.42 | 67.32 |
| 22 | 77.9 | 78.56 | 73.1 | 74.89 | 74.59 | 71.09 | 70.19 | 75.08 | 75.52 |
| 23 | 90.545 | 87.35 | 84.53 | 84.29 | 85.4 | 85.95 | 88.63 | 90.05 | 90.41 |

**Table 51: HR means**

*Missing so replaced with overall mean for condition

| Partic | Bl | Normal | Loud | Normal | 20%pl | Normal | Bad mike | Normal | 5%pl |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 10.71 | 6.51 | 5.21 | 8.9 | 5.19 | 7.19 | 6.74 | 12.3 | 13.6 |
| 2 | 7.02 | 5.1 | 6.6 | 5.3 | 11.3 | 8.66 | 10.75 | 10.39 | 9.61 |
| 3 | 9.27 | 6.23 | 4.64 | 4.72 | 7.83 | 13.57 | 8.13 | Missing | Missing |
| 4 | 17.54 | 4.92 | 4.45 | 4.55 | 4.47 | 15.61 | 10.78 | 7.53 | 14.26 |
| 5 | 10.95 | 3.47 | 4.62 | 4.29 | 5.19 | 6.55 | 5.85 | 3.51 | 4.2 |
| 6 | 9.12 | 4.58 | 3.46 | 4.35 | 5.7 | 9.09 | 8.15 | 6.12 | 6.46 |
| 7 | 10.84 | 5.17 | 5.22 | 4.82 | 5.38 | 6.7 | 13.1 | 19.4 | 5.9 |
| 8 | 8.14 | 7.67 | 6.97 | 11.95 | 7.62 | 12.44 | 7.11 | 10.5 | 8.29 |
| 9 | 7.64 | 12.04 | 12.88 | 12.54 | 6.57 | 8.23 | 6.84 | 13.66 | 5.34 |
| 10 | 7.68 | 15.18 | 12.71 | 15.06 | 7.53 | 14.91 | 14.78 | 10.56 | 10.44 |
| 11 | 11.55 | 6.26 | 13.45 | 11.03 | 15.42 | 8.89 | 13.87 | 5.33 | 7.25 |
| 12 | 10.54 | 6.12 | 13.72 | 6.28 | 7.88 | 8.19 | 10.34 | 8.42 | 8.37 |
| 13 | 10.41 | 5.72 | 2.79 | 5.91 | 14.51 | 7.42 | 10.98 | 9.5 | 13.01 |
| 14 | 7.49 | 8.78 | 6.15 | 3.76 | 6.01 | 5.77 | 4.72 | 9.46 | 13.4 |
| 15 | 11.95 | Missing | Missing | 10.3 | 7.21 | 10.43 | 6.56 | 8.57 | 12.53 |
| 16 | 7.49 | 6.39 | 9.36 | 10.44 | 5.13 | 7.67 | 5.21 | 4.18 | 7.6 |
| 17 | 10.09 | 10.42 | 8.08 | 6.34 | 10.74 | 14.9 | 19.29 | 6.95 | 10.94 |
| 18 | 3.96 | 3.72 | 0 | $5.1^{-06}$ | $5.07^{-06}$ | 3.6 | 2.74 | 2.3 | 3.38 |
| 19 | 9.69 | 3.88 | 6.78 | 5.48 | 6.3 | 5.02 | 7.96 | 4.19 | 6.84 |
| 20 | 11.28 | 13.97 | 9.44 | 11.78 | 14.34 | 11.55 | 10.42 | 5.41 | 8.52 |
| 21 | 16.15 | 6.19 | 8.45 | 6.74 | 6.54 | 5.46 | 5.58 | 6.86 | 6.48 |
| 22 | 7.86 | 8.07 | 9.72 | 6.76 | 5.67 | 9.94 | 6.32 | 8.51 | 5.95 |
| 23 | 14.17 | 17.35 | 11.89 | 7.38 | 11.06 | 8.72 | 8.91 | 8.16 | 7.42 |

**Table 52: HR standard deviations**

| Partic | Bl | Normal | Loud | Normal | 20%pl | Normal | Bad mike | Normal | 5%pl |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 42.33 | 41.46 | 33.1 | 35.19 | 34.3 | 44.73 | 37.34 | 33.99 | 32.78 |
| 2 | 50.92 | 44.6 | 46.23 | 46.5 | 49.94 | 54.55 | 45.91 | 39 | 38.63 |
| 3 | 28.86 | 31.36 | 29.83 | 30.17 | 28.91 | 29.69 | 29.33 | Missing* | Missing* |
| 4 | 29.66 | 25.3 | 25.19 | 25.82 | 25.02 | 24.34 | 24.32 | 24.92 | 24.54 |
| 5 | 43.1 | 53.22 | 53.49 | 33.66 | 38.08 | 39.33 | 43.87 | 56.17 | 57.56 |
| 6 | 42.36 | 38.21 | 37.92 | 32.29 | 36.15 | 34.22 | 36.84 | 33.95 | 33.4 |
| 7 | 53.18 | 43.25 | 42.77 | 44.8 | 39.7 | 58.1 | 55.43 | 63.34 | 44.24 |
| 8 | 26.82 | 24.14 | 24.21 | 24.08 | 25.31 | 24.19 | 24.04 | 24.22 | 24.14 |
| 9 | 48.73 | 26.91 | 27.16 | 24.89 | 24.38 | 28.79 | 26.01 | 26.43 | 25.49 |
| 10 | 44.23 | 27.77 | 27.47 | 28.79 | 31.91 | 27.61 | 26.99 | 25.65 | 24.88 |
| 11 | 34.62 | 39.92 | 36.17 | 26.88 | 24.77 | 24.9 | 29.38 | 25.51 | 32.66 |
| 12 | 45.17 | 40.23 | 37.3 | 41.62 | 39.1 | 41.64 | 39.31 | 41.06 | 39.64 |
| 13 | 38.34 | 32.67 | 33.05 | 32.25 | 34.31 | 41.23 | 32.72 | 32.51 | 31.37 |
| 14 | 40.22 | 30.7 | 30.88 | 50.8 | 38.49 | 51.55 | 47.12 | 42.92 | 44.87 |
| 15 | 56.74 | Missing* | Missing* | 51.62 | 55.09 | 45.76 | 53.59 | 48.67 | 51.73 |
| 16 | 58.08 | 57.78 | 56.56 | 58.96 | 56.77 | 35.51 | 30.72 | 49.39 | 53.28 |
| 17 | 44.33 | 60.64 | 42.32 | 60.32 | 47.06 | 46.57 | 52.43 | 30.9 | 35.57 |
| 18 | 24.32 | 26.08 | 25.98 | 27.49 | 27.49 | 24.46 | 24.74 | 23.89 | 24.13 |
| 19 | 37.51 | 30.85 | 34.16 | 32.9 | 35.58 | 28.32 | 30.54 | 28.65 | 28.32 |
| 20 | 44.39 | 37.53 | 41.19 | 37.13 | 37.32 | 38.2 | 39.5 | 42.3 | 42.08 |
| 21 | 60.5 | 59.21 | 65.11 | 50.94 | 54.02 | 60.72 | 64.9 | 59.5 | 59.93 |
| 22 | 25.45 | 25.8 | 26.79 | 26.64 | 26.95 | 24.47 | 24.29 | 25.14 | 24.8 |
| 23 | 39.19 | 26.25 | 29.45 | 44.24 | 44.67 | 41.65 | 40.69 | 38.74 | 39.45 |

**Table 53: BVP means**

*Missing so replaced with overall mean for condition

| Partic | Bl | Normal | Loud | Normal | 20%pl | Normal | Bad mike | Normal | 5%pl |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 8.89 | 5.76 | 3.55 | 4.43 | 3.43 | 6.84 | 3.49 | 3.44 | 3.49 |
| 2 | 12.8 | 5.69 | 5.18 | 5.31 | 6.29 | 7.3 | 7.52 | 4.24 | 4.87 |
| 3 | 0.67 | 1.23 | 0.73 | 0.99 | 0.54 | 1.3 | 1.05 | Missing | Missing |
| 4 | 9.42 | 0.28 | 0.46 | 0.47 | 0.29 | 0.29 | 0.2 | 0.43 | 0.38 |
| 5 | 10.98 | 7.53 | 8.59 | 2.28 | 6 | 4.73 | 4.74 | 6.55 | 5.48 |
| 6 | 8.51 | 3.73 | 4.08 | 3.01 | 3 | 4.05 | 5.09 | 3.34 | 2.28 |
| 7 | 7.66 | 5.61 | 7.19 | 5.67 | 5.05 | 7.56 | 9.53 | 7.75 | 6.85 |
| 8 | 1.77 | 0.06 | 0.26 | 0.47 | 0.68 | 0.23 | 0.16 | 0.18 | 0.14 |
| 9 | 12.14 | 3.14 | 4.5 | 1.03 | 0.17 | 3.11 | 0.53 | 1.31 | 0.48 |
| 10 | 9.95 | 2.55 | 1.86 | 1.75 | 4.87 | 3.27 | 2.66 | 1.35 | 0.57 |
| 11 | 8.12 | 3.37 | 4.06 | 6.3 | 0.66 | 0.79 | 6.59 | 1.59 | 8.33 |
| 12 | 6.63 | 3.47 | 4.23 | 5.58 | 4.56 | 4.13 | 4.53 | 4.1 | 4.75 |
| 13 | 6.18 | 1.98 | 1.91 | 1.96 | 3.22 | 4.09 | 2.19 | 2.29 | 2.1 |
| 14 | 7.22 | 0.9 | 0.83 | 4.15 | 5.24 | 6.01 | 6.21 | 6.52 | 7.01 |
| 15 | 10.45 | Missing | Missing | 5.59 | 3.09 | 6.44 | 3.8 | 7.17 | 8.26 |
| 16 | 8.07 | 11.82 | 11.64 | 10.92 | 9.03 | 11.48 | 6.72 | 7.08 | 9.05 |
| 17 | 11.49 | 10.44 | 11.22 | 5.9 | 15.46 | 11.25 | 17.25 | 2.76 | 4.67 |
| 18 | 0.36 | 1.03 | $3.98^{-06}$ | $1.98^{-06}$ | $2.03^{-06}$ | 0.36 | 0.16 | 0 | 0.2 |
| 19 | 11.06 | 2.8 | 3.04 | 2.98 | 3.93 | 1.05 | 2.1 | 1.48 | 1.1 |
| 20 | 8.53 | 6.18 | 6.46 | 6.3 | 6.24 | 5.67 | 6.45 | 6.93 | 6.59 |
| 21 | 14.46 | 6.24 | 5.59 | 4.42 | 3.34 | 4.28 | 3.96 | 5.18 | 6.51 |
| 22 | 2.18 | 1.1 | 1.38 | 0.95 | 1.54 | 0.63 | 0.19 | 0.68 | 0.38 |
| 23 | 5.43 | 0.99 | 39 | 7.34 | 8.07 | 6.9 | 4.91 | 3.57 | 3.52 |

**Table 54: BVP standard deviations**

## Experiment 4

| Partic | Frame rate order | Actor order | Gender |
|---|---|---|---|
| 1 | 25fps first | Female actor first | Female |
| 2 | 5fps first | Male actor first | Male |
| 3 | 25fps first | Male actor first | Female |
| 4 | 5fps first | Female actor first | Female |
| 5 | 25fps first | Male actor first | Female |
| 6 | 5fps first | Female actor first | Male |
| 7 | 25fps first | Male actor first | Female |
| 8 | 5fps first | Male actor first | Female |
| 9 | 25fps first | Female actor first | Female |
| 10 | 5fps first | Male actor first | Female |
| 11 | 25fps first | Male actor first | Female |
| 12 | 5fps first | Female actor first | Female |
| 13 | 25fps first | Male actor first | Female |
| 14 | 5fps first | Female actor first | Male |
| 15 | 25fps first | Female actor first | Female |
| 16 | 5fps first | Male actor first | Female |
| 7 | 25fps first | Female actor first | Female |
| 18 | 5fps first | Female actor first | Male |
| 19 | 25fps first | Female actor first | Female |
| 20 | 5fps first | Male actor first | Female |
| 21 | 5fps first | Male actor first | Female |
| 22 | 5fps first | Male actor first | Female |
| 23 | 25fps first | Female actor first | Female |
| 24 | 5fps first | Male actor first | Male |

**Table 55: Gender of participants and orders received**

| Partic | Bl | Stdev | 5fps | Stdev | 25fps | Stdev |
|---|---|---|---|---|---|---|
| 1 | 8.16 | 1.08 | 10.61 | .78 | 9.78 | .50 |
| 2 | 1.69 | .45 | 3.67 | .81 | 4.17 | .57 |
| 3 | 2.16 | .41 | 3.84 | .57 | 3.55 | .42 |
| 4 | 4.53 | .93 | 6.55 | .40 | 6.86 | .63 |
| 5 | 2.33 | .59 | 4.96 | .66 | 5.07 | .94 |
| 6 | 3.93 | 1.19 | 8.54 | 2.52 | 8.24 | 1.29 |
| 7 | 1.51 | .32 | 2.75 | .41 | 3.33 | .23 |
| 8 | 11.65 | 1.02 | 12.73 | .45 | 12.58 | .50 |
| 9 | 3.28 | 1.23 | 12.98 | 2.28 | 12.79 | 2.02 |
| 10 | 4.30 | .82 | 7.80 | 1.12 | 7.47 | 1.30 |
| 11 | 4.46 | 1.15 | 9.90 | 1.52 | 8.36 | 1.15 |
| 12 | 2.99 | .95 | 5.78 | .79 | 6.91 | 1.30 |
| 13 | 2.00 | .68 | 10.27 | .77 | 9.16 | .61 |
| 14 | 1.66 | .11 | 2.70 | .14 | 4.58 | .22 |
| 15 | 6.76 | 2.05 | 12.09 | 1.69 | 10.16 | .69 |
| 16 | .75 | .30 | 3.37 | .47 | 3.82 | .39 |
| 17 | 1.11 | .41 | 4.56 | .46 | 4.55 | .37 |
| 18 | 2.00 | .43 | 3.49 | .39 | 3.77 | .43 |
| 19 | .45 | .12 | 1.35 | .16 | 1.35 | .29 |
| 20 | 3.41 | .37 | 5.49 | .74 | 4.51 | .35 |
| 12 | .38 | .03 | 2.01 | .37 | 2.08 | .21 |
| 22 | 2.94 | .31 | 3.26 | .15 | 3.73 | .40 |
| 23 | 6.31 | 1.20 | 11.29 | .73 | 10.61 | .40 |
| 24 | 2.62 | .34 | 4.01 | .30 | 5.02 | .38 |

**Table 56: SC means and standard deviations**

| Partic | Bl | Stdev | 5fps | Stdev | 25fps | Stdev |
|---|---|---|---|---|---|---|
| 1 | 77.47 | 20.93 | 86.58 | 9.32 | 91.55 | 8.44 |
| 2 | 80.61 | 15.88 | 64.28 | 18.54 | 78.62 | 15.23 |
| 3 | 95.58 | 4.56 | 111.56 | 5.86 | 127.38 | 6.10 |
| 4 | 72.93 | 5.06 | 83.27 | 8.23 | 83.06 | 5.93 |
| 5 | 72.39 | 10.69 | 83.13 | 9.49 | 95.12 | 14.80 |
| 6 | 76.02 | 5.75 | 73.10 | 9.68 | 73.15 | 7.96 |
| 7 | 72.58 | 6.17 | 72.49 | 10.38 | 71.89 | 13.46 |
| 8 | 84.36 | 8.36 | 88.53 | 14.79 | 53.17 | 14.65 |
| 9 | 60.06 | 9.01 | 47.24 | 20.06 | 55.10 | 16.51 |
| 10 | 69.07 | 12.67 | 83.75 | 15.23 | 68.16 | 7.98 |
| 11 | 68.47 | 11.18 | 75.67 | 7.03 | 78.21 | 7.96 |
| 12 | 78.76 | 4.11 | 85.02 | 11.24 | 88.25 | 19.33 |
| 13 | 73.24 | 7.19 | 83.11 | 8.85 | 105.62 | 12.03 |
| 14 | 73.49 | 5.00 | 80.73 | 6.17 | 77.23 | 4.61 |
| 15 | 86.07 | 8.55 | 94.85 | 10.91 | 95.52 | 9.28 |
| 16 | 52.65 | 4.13 | 56.62 | 5.14 | 52.68 | 5.80 |
| 17 | 74.16 | 4.51 | 86.73 | 7.59 | 87.80 | 6.25 |
| 18 | 83.72 | 7.97 | 80.93 | 5.07 | Missing * | Missing |
| 19 | 77.97 | 7.11 | 97.09 | 9.84 | 100.06 | 7.05 |
| 20 | 83.79 | 6.44 | 100.06 | 7.05 | 97.09 | 9.84 |
| 12 | 86.02 | 5.93 | 91.55 | 5.77 | 84.03 | 4.51 |
| 22 | 97.63 | 18.09 | 115.77 | 11.76 | 99.85 | 14.92 |
| 23 | 91.67 | 7.77 | 86.91 | 12.78 | 95.96 | 16.23 |
| 24 | 46.03 | 18.49 | 40.93 | 6.71 | 37.56 | 4.26 |

**Table 57: HR means and standard deviations**

* Missing value so replaced with overall mean of condition. The standard deviation was not analysed in an ANOVA and therefore their mean was omitted in the standard deviation graph

| Partic | BI | Stdev | 5fps | Stdev | 25fps | Stdev |
|---|---|---|---|---|---|---|
| 1 | 64.58 | 9.20 | 42.65 | 4.70 | 43.77 | 3.76 |
| 2 | 53.79 | 15.65 | 64.48 | 10.52 | 65.59 | 7.00 |
| 3 | 46.95 | 12.64 | 39.10 | 2.98 | 30.19 | 2.88 |
| 4 | 29.74 | 4.22 | 25.24 | 0.33 | 24.95 | 0.24 |
| 5 | 40.56 | 13.95 | 30.70 | 1.90 | 32.42 | 3.39 |
| 6 | 38.26 | 10.24 | 26.41 | 0.90 | 25.09 | 0.28 |
| 7 | 25.02 | 0.47 | 25.82 | 0.70 | 25.09 | 0.33 |
| 8 | 25.02 | 1.12 | 24.19 | 0.14 | 24.71 | 0.39 |
| 9 | 48.09 | 4.21 | 53.97 | 17.35 | 63.63 | 14.05 |
| 10 | 58.39 | 12.50 | 28.22 | 2.35 | 41.73 | 10.62 |
| 11 | 51.23 | 13.48 | 48.83 | 6.83 | 44.93 | 9.81 |
| 12 | 30.74 | 1.21 | 27.36 | 1.15 | 26.33 | 0.95 |
| 13 | 48.98 | 6.24 | 36.63 | 3.54 | 31.25 | 2.19 |
| 14 | 37.40 | 11.76 | 29.64 | 3.32 | 28.87 | 1.97 |
| 15 | 27.98 | 3.69 | 24.61 | 0.19 | 24.62 | 0.20 |
| 16 | 25.83 | 1.20 | 27.08 | 0.69 | 28.72 | 0.61 |
| 17 | 27.89 | 2.77 | 25.27 | 0.48 | 25.48 | 0.35 |
| 18 | 58.14 | 7.45 | 48.13 | 7.89 | Missing | Missing |
| 19 | 54.44 | 7.54 | 24.18 | 0.24 | 24.32 | 0.21 |
| 20 | 23.86 | 0.45 | 24.32 | 0.21 | 24.18 | 0.24 |
| 12 | 25.40 | 0.69 | 25.14 | 0.49 | 27.11 | 1.48 |
| 22 | 24.42 | 0.27 | 24.50 | 0.28 | 26.57 | 4.19 |
| 23 | 39.84 | 9.23 | 27.26 | 1.37 | 27.33 | 1.87 |
| 24 | 24.98 | 1.68 | 24.03 | 0.12 | 24.53 | 0.16 |

**Table 58: BVP means and standard deviations**

## Experiment 5

| Partic | BI | BVBA | GVBA | BVGA | GVGA |
|---|---|---|---|---|---|
| 1 | Missing | 9.87 | 6.19 | 8.54 | Missing |
| 2 | 6.8 | 12.75 | 13.25 | 11.22 | 9.07 |
| 3 | 12.24 | 20.24 | 15.32 | 20.60 | 18.27 |
| 4 | 10.67 | 19.13 | 19.90 | 20.58 | 19.60 |
| 5 | 3.46 | 7.10 | 6.44 | 7.30 | 5.43 |
| 6 | 5.15 | 6.59 | 7.34 | 7.69 | 6.89 |
| 7 | 10.63 | 14.56 | 13.29 | 14.84 | 14.53 |
| 8 | 4.21 | 8.04 | 7.82 | 7.78 | 6.93 |
| 9 | 3.81 | 7.61 | 7.41 | 6.10 | 7.01 |
| 10 | 3.43 | 5.12 | 4.54 | 5.00 | 7.78 |
| 11 | 3.43 | 6.38 | 6.63 | 5.75 | 7.14 |

**Table 59: SC means**

| Partic | BI | BVBA | GVBA | BVGA | GVGA |
|---|---|---|---|---|---|
| 1 | Missing | 0.66 | 0.77 | 0.74 | Missing |
| 2 | 0.57 | 0.58 | 0.64 | 0.84 | 0.24 |
| 3 | 2.77 | 1.93 | 1.07 | 1.57 | 2.49 |
| 4 | 2.69 | 0.58 | 0.74 | 0.71 | 0.97 |
| 5 | 0.76 | 0.56 | 0.53 | 0.76 | 0.46 |
| 6 | 0.46 | 0.49 | 0.37 | 0.53 | 0.20 |
| 7 | 2.01 | 0.87 | 1.05 | 0.72 | 0.83 |
| 8 | 0.83 | 0.65 | 0.63 | 0.54 | 0.45 |
| 9 | 1.01 | 0.32 | 0.49 | 0.28 | 0.32 |
| 10 | 0.88 | 0.43 | 0.38 | 0.53 | 1.04 |
| 11 | 0.4 | 0.69 | 0.47 | 0.51 | 0.66 |

**Table 60: SC standard deviations**

| Partic | BI | BVBA | GVBA | BVGA | GVGA |
|---|---|---|---|---|---|
| 1 | 65.06 | 60.58 | 59.92 | 60.71 | 64.77 |
| 2 | 76.80 | 78.88 | 81.61 | 82.29 | 82.59 |
| 3 | 86.90 | 74.59 | 71.19 | 68.98 | 75.15 |
| 4 | 68.17 | 65.84 | 68.93 | 65.23 | 63.49 |
| 5 | 83.42 | 84.50 | 82.45 | 86.45 | 82.89 |
| 6 | 66.09 | 70.27 | 66.10 | 66.71 | 68.84 |
| 7 | 83.01 | 80.65 | 73.23 | 77.82 | 74.48 |
| 8 | 72.77 | 68.37 | 67.12 | 68.64 | 71.19 |
| 9 | 79.56 | 69.43 | 67.86 | 84.37 | 77.41 |
| 10 | 71.44 | 70.19 | 70.07 | 73.03 | 77.11 |
| 11 | 106.80 | 94.65 | 85.38 | 104.99 | 86.15 |

**Table 61: HR means**

| Partic | BI | BVBA | GVBA | BVGA | GVGA |
|---|---|---|---|---|---|
| 1 | 12.9 | 8.02 | 4.46 | 6.55 | 12.72 |
| 2 | 21.22 | 24.01 | 11.54 | 17.94 | 23.71 |
| 3 | 11.48 | 25.87 | 27.63 | 27.07 | 21.97 |
| 4 | 10.18 | 11.93 | 26.84 | 14.87 | 8.53 |
| 5 | 5.09 | 6.47 | 7.35 | 5.9 | 6.58 |
| 6 | 16.79 | 14.42 | 9.06 | 15.49 | 12.1 |
| 7 | 6.51 | 9.02 | 10.54 | 8.98 | 18.14 |
| 8 | 9.39 | 8.63 | 7.91 | 7.9 | 6.68 |
| 9 | 7.61 | 9.75 | 8 | 9.23 | 9.18 |
| 10 | 23.94 | 25.87 | 28.98 | 23.84 | 15.26 |
| 11 | 6.39 | 15.26 | 19.34 | 8.39 | 13.43 |

**Table 62: HR standard deviations**

| Partic | BI | BVBA | GVBA | BVGA | GVGA |
|---|---|---|---|---|---|
| 1 | 25.98 | 47.74 | 29.09 | 33.42 | 25.23 |
| 2 | 29.94 | 26.23 | 26.78 | 24.95 | 24.27 |
| 3 | 37.39 | 37.01 | 29.60 | 33.41 | 34.50 |
| 4 | 36.48 | 24.97 | 26.56 | 24.72 | 25.55 |
| 5 | 31.85 | 37.65 | 38.36 | 30.13 | 30.08 |
| 6 | 24.59 | 25.63 | 24.13 | 24.37 | 24.36 |
| 7 | 47.88 | 45.64 | 39.97 | 42.71 | 37.87 |
| 8 | 33.23 | 39.70 | 37.29 | 33.89 | 27.55 |
| 9 | 41.31 | 50.31 | 54.59 | 47.06 | 49.86 |
| 10 | 25.73 | 25.68 | 31.88 | 26.41 | 26.87 |
| 11 | 31.52 | 43.04 | 34.44 | 28.88 | 31.41 |

**Table 63: BVP means**

| Partic | BI | BVBA | GVBA | BVGA | GVGA |
|---|---|---|---|---|---|
| 1 | 1.28 | 10.74 | 2.46 | 8.5 | 0.56 |
| 2 | 1.72 | 1.65 | 1.88 | 1.07 | 0.56 |
| 3 | 6.4 | 9.65 | 6.74 | 7.75 | 9.65 |
| 4 | 4.9 | 0.78 | 1.28 | 0.86 | 1.45 |
| 5 | 3.25 | 9.4 | 9.35 | 2.58 | 2.45 |
| 6 | 0.87 | 6.26 | 0.28 | 0.32 | 0.36 |
| 7 | 6.13 | 5.22 | 5.42 | 5.92 | 7.75 |
| 8 | 9.33 | 9.28 | 6.5 | 6.1 | 1.56 |
| 9 | 13.07 | 8.11 | 9.83 | 12.65 | 7.92 |
| 10 | 4.21 | 2.78 | 10.66 | 2 | 1.95 |
| 11 | 3.58 | 14.33 | 7.78 | 2.34 | 4.9 |

**Table 64: BVP standard deviations**