



FP6- 506811

**AMI
AUGMENTED MULTI-PARTY INTERACTION**

<http://www.amiproject.org/>

Integrated Project
Information Society Technologies

**D6.4 MEETING BROWSER EVALUATION
REPORT**

Due date: 31/12/2006

Submission date: 31/12/2006

Project start date: 1/1/2004

Duration: 36 months

Lead Contractor: PHI

Revision: 1.0

Project co-funded by the European Commission in the 6th Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	<input checked="" type="checkbox"/>
PP	Restricted to other programme participants (including the Commission Services)	<input type="checkbox"/>
RE	Restricted to a group specified by the consortium (including the Commission Services)	<input type="checkbox"/>
CO	Confidential, only for members of the consortium (including the Commission Services)	<input type="checkbox"/>



D6.4 MEETING BROWSER EVAL REPORT

Anita Cremers (TNO)
Wilfried Post (TNO)
Erwin Elling (TNO)
Betsy van Dijk (TWENTE)
Bram van der Wal (PHI)
Jean Carletta (UoE)
Mike Flynn (IDIAP)
Pierre Wellner (IDIAP)
Simon Tucker (SHEF)

Content

Introduction	2
1. Browser Evaluation Test (BET) method.....	3
1.1 Overview of the BET Method.....	3
1.1.1 Collecting Observations	4
1.1.2 Ordering Observations.....	5
2. Task-based Browser Evaluation Test (TBET) method.....	8
2.1 Objective.....	8
2.2 Method	8
2.2.1 Subjects	9
2.2.2 Conditions.....	9
2.2.3 Measures	9
2.2.4 Procedure	10
2.3 Description of Browsers tested.....	12
2.3.1 Condition 0: Baseline condition	12
2.3.2 Condition 1: Browser 1.....	15
2.3.3 Condition 2: Browser 2.....	17
2.3.4 Overview of functionalities	18
3. Results of BET.....	19
3.1 Conditions tested with the BET.....	19
3.2 Raw uncalibrated results.....	23
3.3 The Speed/Accuracy Trade-off Model	24
3.3.1 Subjects' variable performance	24
3.3.2 A simple model	24
3.3.3 Comparing browsers	24
3.3.4 Method	25
3.3.5 Results.....	26
3.3.6 Validity of the model.....	26
3.4 Questionnaire responses	27
3.5 Conclusions.....	30
3.6 Acknowledgements	30
3.7 References	30
4. Results of TBET	31
4.1 Introduction	31
4.1.1 Pre-test.....	31
4.1.2 Tool assessment.....	32
4.1.3 Post test.....	33
4.2 Conclusions and further work	36
4.3 References	36
Conclusion	37
Appendix 1: Observer Instruction pages	38
Appendix 2: Observation Editing pages.....	43
Appendix 3: Subject Instruction pages.....	44
Appendix 4: BET Database Schema.....	45

List of figures

Figure 1. The BET method.	3
Figure 2 Observation Input form.....	4
Figure 3 Rating importance and creating false version of observations.....	5
Figure 4 Instrumented meeting room (Team Cockpit) at Soesterberg.	9
Figure 5 Desktop in the Baseline condition.	13
Figure 6 Shared project folder in Baseline condition..	14
Figure 7 Desktop in Condition 1 and Condition 2.....	15
Figure 8 Meeting browser in Condition 1.....	16
Figure 9 Meeting browser in Condition 2.....	17
Figure 10 Subject Calibration Condition.....	19
Figure 11: Base Condition.....	20
Figure 12 Speedup Condition.....	21
Figure 13 Overlap Condition.....	22
Figure 14: Interaural time difference.....	22
Figure 15: Graph showing speed and accuracy scores for all subjects and all conditions.....	23
Figure 16. The speed/accuracy trade-off model applied to the Calibrate condition.....	24
Figure 17: Questionnaire responses for Base Condition.....	27
Figure 18: Questionnaire responses for Speedup Condition.....	28
Figure 19: Questionnaire responses for Overlap Condition.....	29
Figure 20: Observer Introductory Instructions.....	38
Figure 21: Observer Task 1 Instructions.....	39
Figure 22: Observer Task 1 screenshot.....	39
Figure 23: Observer Task 2 instructions.....	40
Figure 24: Observer Task 2 instructions (continued).....	41
Figure 25: Observer Task 2 screenshot.....	42

Introduction

This deliverable acts a wrapper for two reports and describes the test methods used for the evaluation meeting browsers (BET) and task-based meeting browsers (TBET) conducted at IDIAP and TNO.

The evaluations take into account the relevant user requirements from D6.2 User Requirements updated version 2006-12-31.

Chapter 1 and 2 describe in detail the test methodologies for the BET and TBET evaluations, while Chapter 3 and 4 describe the outcomes of the different user tests and evaluates what improvements need to be done to make the browsers more users friendly and more efficient. Also a brief overview is given in both chapters towards future work that needs to be done in the AMIDA project for user requirements evaluation.

1. Browser Evaluation Test (BET) method

In many fields of research, an objective measure of system performance along with a standard data corpus and set of reference tasks has been of enormous benefit in helping researchers compare techniques and make progress. For example, in the field of speech recognition, this has made possible the construction of real time, large vocabulary systems that would not have been feasible ten years ago. The text retrieval conference (TREC) has also used standard corpora, tasks and metrics with great success: average precision doubled from 20% to 40% in the last seven years.

This work aims to develop similar objective metrics for meeting browsers, in order to complement studies that rely on subjective satisfaction ratings. It describes a *browser evaluation test* (or BET) for meeting browsers.

We define the task of **browsing** a meeting recording as an attempt to find a maximum number of **observations of interest** in a minimum amount of time.

A key problem in testing browsers, therefore, is identifying these *observations of interest*. The range of possibilities is enormous and depends upon meeting content and individual user interests. The BET aims to be:

- an objective measure of browser effectiveness based on user performance rather than satisfaction;
- independent of experimenter perception of the browsing task and meeting structure;
- produce directly comparable numeric scores, automatically; and
- replicable, through a publicly accessible web site allowing different researchers to evaluate their browsers and benchmark them.

An early version of the BET was described in [1]. For this report, however, we use a modified version of the test that aims to overcome some of the shortcomings of the initial version. The sections below present an overview of the modified method and describe each of its significant features in detail.

1.1 Overview of the BET Method

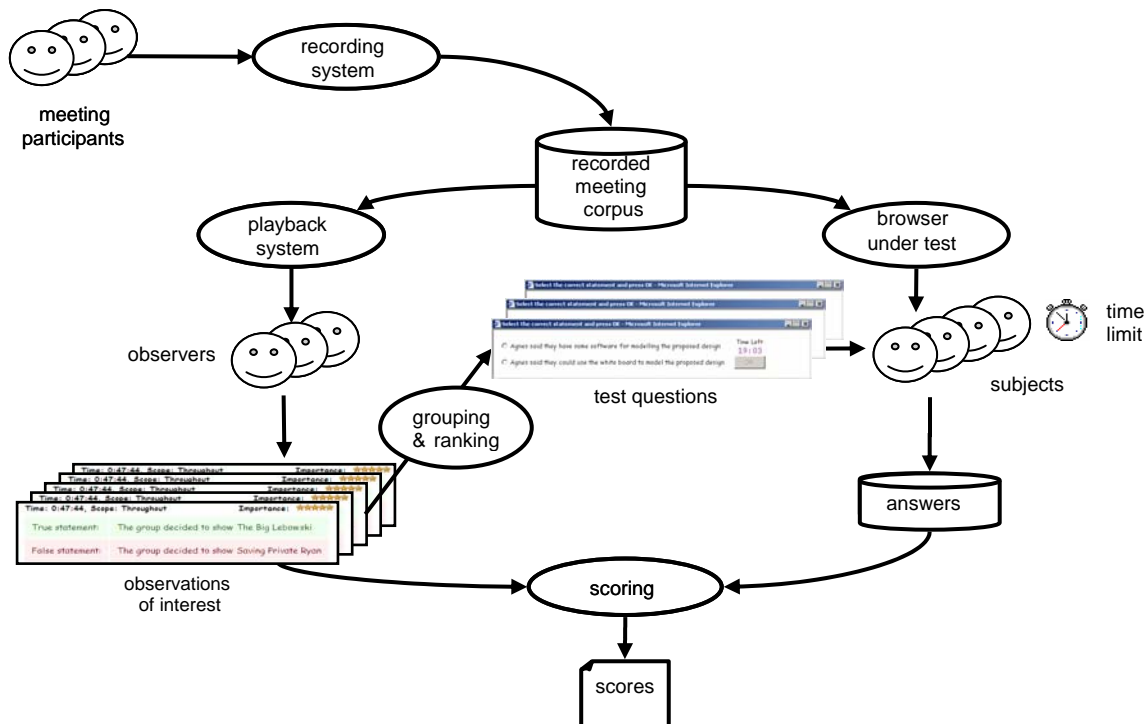


Figure 1. The BET method.

The BET method is illustrated in Figure 1. The significant features are described below, with further detail in subsequent sections:

- The *corpus* is a significant set of media recordings providing the data to be browsed.
- *Observers* watch selected meetings from the corpus, to produce a store of *observations*. Observers are not meeting participants.
- Later, during testing, the observations on some meeting are sampled to produce *tests*.
- *Subjects* use the *browser under test* to review the meeting, answering as many test questions as possible in a short time.
- *Answers* produced by the subjects are stored for scoring and analysis.
- *Scoring* compares the subjects' test answers to the original stored observations, to compute a *score* for the browser.

Using the BET requires one-time investment in creation of the corpus, collection of the observations and running of benchmark tests. Subsequent browser tests take advantage of this one-time effort to run tests and produce comparable scores. The BET differs from classic usability testing because task details are not predetermined by the experimenter, and the BET does not necessarily measure satisfaction.

The recorded meetings used for the BET were made in IDIAP's smart meeting room and are part of the AMI Corpus [2]. The specific recordings used were IB4010, IS1008c, and ISSCO-Meeting_024.

1.1.1 Collecting Observations

Questions to be used in browser tests are determined by a set of observers, who produce the *observations of interest*. Observers have available the full recordings from every media source, including slides. There is no time limit for the observers, but in the trial run, people spent about 4½ times the duration of the meeting to complete their observations. Each observer is instructed to produce observations that the meeting participants appear to consider interesting. This approach is meant to temper undue influence of each observer's own special interests, while avoiding the introduction of experimenter bias regarding the relative importance of particular meeting events.

Each observation is stated as a complementary pair of statements, one true and one false, both of which are later presented to subjects during testing. Observers are instructed to produce observations that should be difficult to guess without access to the recording (difficulty is verified later), and the observations should be simply and concisely stated.

BET observations by demo@amiproject.org on meeting IS1008c

Remember:

- Interesting to participants.
- Player positioned correctly.

Review the [full instructions](#).

True statement:

Scope:

Here
 Around
 Throughout

[See all my observations](#) | [Go on to Task 2](#)

Figure 2 Observation Input form

Observers create observations in two steps: first they create a list of true observations; second, they rate the importance of each observation and create a false version of each. The interface to the first step is show in Figure 2. Each observation is time-stamped with the media time into the recording, and submitted with an estimate of its locality: *nearby*, *around* or *throughout*. Locality can be used to determine the temporal correspondence between questions and their answers.

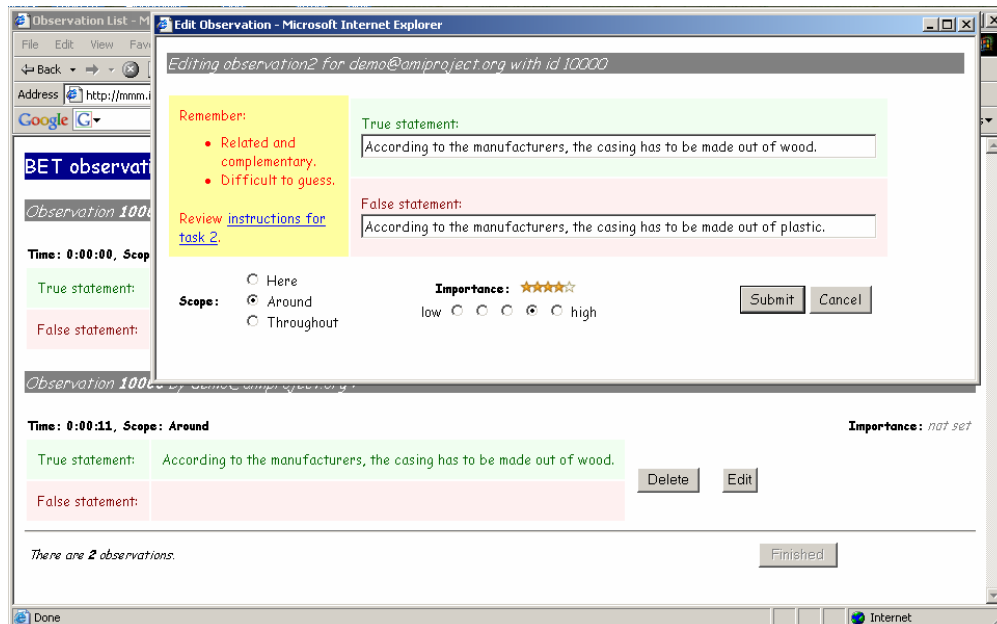


Figure 3 Rating importance and creating false version of observations

After observers have completed a list of true observations, they are asked to rate their importance and create a false version of each, as show in Figure 3. These tasks are done separately from the original input of observations for two reasons: first so that all observations can be taken into account when rating their relative importance; second, so that the need to create matched complementary pairs does not influence the choice of observations made. The first version of the BET had been criticized for producing a large number of “trivial” observations - perhaps for the ease of creating complimentary pairs.

1.1.2 Ordering Observations

1.1.2.1 Grouping

Because several different people observe the same meeting, in many cases they make the same or very similar observations. When the true statements of multiple observations are very similar or overlap, they are manually grouped by the experimenters. Subjects are tested using a single representative from the group, to prevent one observation pair from revealing the answer to another.

One way this can happen, for example, is when part of the true and false statements in a single observation are identical, revealing that this information must be true. When grouping observations, we find that the media times of observations in the same group are usually very close, but this is not a strict requirement.

Observers of observations in the same group are usually different, but there were some cases when a single observer made essentially the same observation twice. Because size of group affects the order of questions presented to subjects (as described below), we disqualify these cases. Each observer gets a maximum of one “vote” on the importance of the group, and we select the observation rated with the higher importance. If they have equal importance, an arbitrary selection is made.

1.1.2.2 Selecting group representatives

After observations have been placed into groups, a single observation from each group is manually selected by the experimenters to be the representative of the group. The criteria for selection are meant to not favor one type of browser over another, and they are as follows:

1. Must meet validity criteria (see below).
2. Concise & crisply expressed.
3. One factual point preferred rather than two or more.
4. Same keywords as group (in true statements)
5. Select least guessable.
6. Otherwise random (importance rating does not matter).

1.1.2.3 Validity criteria

Observation pairs were read by the experimenters, and some of them were judged as invalid for any of the reasons below. Care was taken to ensure that reasons for rejection would be browser-neutral, and not select observations that are better suited to a particular kind of browsing technique. Rejection of each observation required consensus between five different experimenters working on very different browser designs. Criteria for validity were as follows:

- The true statement must always be true and false statement must always be false. A common reason that this criterion is not met is when the observation refers to a fleeting moment in the recording that is not consistently true. (rejection code B)
- Statements are rejected if considered incomprehensible to a typical native English speaker because of serious grammatical problems, typographical errors, obscure words, or just too unclear. (rejection code C)
- Too easily guessable: the true statement is completely obvious without using any knowledge from the meeting, or the false statement is obviously wrong. (rejection code G)
- Not parallel enough, unrelated to each other, or not mutually exclusive. (rejection code P)
- Redundant observations not selected as group representative are marked with rejection code r.

Observations based on censored material, which participants asked to be left out of the recording were also removed. (rejection code x)

1.1.2.4 Editing

One of the main principles of the BET is that observations are experimenter-neutral so as not to bias them in favor of any particular browser design. Our initial policy had been that the original observations could not be edited. But after much discussion the experimenters chose to allow limited editing for any of five possible reasons. The edit should be browser-neutral and the original observation is kept in the database, along with one of the five following codes indicating the reason for the edit:

1. SP for spelling
2. GR ammar
3. EX plicitness (make obs more explicit)
4. BR evity (make obs shorter, to remove distractors)
5. CO mplementarity of observations (and parallelism)

1.1.2.5 Experimenter consensus

Judgment is required to group observations, validate observations, and select group representatives. Strict browser neutrality would require that these judgments should also be made by independent people without possible browser bias, but we performed the tasks ourselves. To ensure that the judgments are made consistently and fairly, the process was done transparently on collaborative web pages, where several “competing” browser development teams were able to check, comment on and discuss these decisions.

1.1.2.6 Adjusted importance

Some observers tend to rate their observations highly, while others rated them as less important. A four-star rating from someone who usually rates observations with four stars is not counted the same as a four-star rating from someone who usually rates observations with two stars.

The median importance is calculated for each observer, and the *adjusted importance* of each observation is determined by the difference between its importance rating relative to its observer's median importance. Median importance per observer is calculated based on all their observations, including rejected ones.

1.1.2.7 Ordering

By size of group, because this represents the number of "votes" the observation received.

For groups of equal size, sort by median adjusted importance.

For groups of equal size and median adjusted importance, sort by mean adjusted importance, then by mediaTime.

The proposed final order of questions is determined first by the size of the group, because this represents the number of "votes" each observation received, and is a kind of "inter-annotater agreement". Groups of equal size are sorted by median adjusted importance (see below). Finally, groups of equal size and median adjusted importance, are sorted by mean adjusted importance, and then by mediaTime.

2. Task-based Browser Evaluation Test (TBET) method

2.1 Objective

In (Post, Huis In't Veld & van den Boogaard, 2007) is pointed out that the success of a meeting is better determined from a series of meetings, such as in the context of a project with a clear goal. Further, the success of a meeting, or a project, depends not only on the means used (e.g. a meeting browser), but also on the (project or meeting) method, individual factors, team factors, type of task, organizational culture, environment, etc.

The objective of the presented study is to determine whether and how a multimodal meeting browser improves a meeting, and consequently might lead to a more efficient and satisfactory project process and higher quality results. The meeting browsers are thus evaluated in the context of the task in which they are being used, which explains the name Task Based Evaluation.

2.2 Method

An experiment was set up to compare meetings without meeting browser support with meetings supported with one of two variants of a multimodal meeting browser. The meeting factors mentioned in section 2.1 have been specified in the following experimental scenario¹.

Four subjects have to participate in a design project team, playing a specific role (project manager, industrial designer, user interface designer and marketing expert). They are told to take over a project carried out so far by a team that didn't do well enough. The subjects have to use all the materials used and produced by this previous team, including recordings of their three meetings. They have to prepare and carry out a final design meeting in which they have to come up with a television remote control prototype (in clay), according to specific requirements.

Both preparation and execution of the meeting is carried out in meeting rooms in Soesterberg (Figure 4) and Edinburgh, a well instrumented research environment for four subjects, with individual workplaces (including a private computer), a shared workplace (including electronic presentation boards), and, depending on the experimental condition, a set of tools. The participants and their computer interactions are observed and recorded by means of video cameras, microphones, and screen videos.

On preplanned points in time, subjects receive e-mails about the tasks to be carried out (sent by a virtual head of the department), some hints (sent by a virtual coach), and a series of questionnaires and rating scales.

¹ An exhaustive description of the scenario and the environment in which the scenario is played is provided in AMI Deliverable D1.3: Extended scenario definition.



Figure 4 Instrumented meeting room (Team Cockpit) at Soesterberg.

2.2.1 Subjects

Data from 22 project teams consisting of 4 participants were collected (87 subjects), half of them in Soesterberg and half in Edinburgh.

2.2.2 Conditions

The experimental conditions are as follows. In the basic condition, no browser is provided. Instead a folder structure in Microsoft Explorer is offered, organized by project phase (Project Kick-off; Functional Design Phase and Conceptual Design Phase). In these folders, participants can find documents, minutes, PowerPoint slides and audio/video recordings of three previous meetings. In the second and third conditions, all documents are still provided in a folder structure. For the meetings however, two variants of a meeting browser are provided, which both offer synchronization of the multimodal material. The only difference is that the browser in the third condition offers more functionalities than the one in the second condition. Section 2.3 provides elaborate descriptions of the different types of browsers that were used.

2.2.3 Measures

A specially developed evaluation instrument is used for measuring project process and outcome (Post et al., 2007). This instrument includes subjective workload rating scales and team questionnaires, and objective analysis of information transfer and project outcome. Table 1 provides an overview of the rating scales, questionnaires and observations. The current report does not describe the results of all measures, only the measures indicated with "*" are reported here, focusing on the usability of the meeting browsers. The rest of the measures will be reported in a separate paper.

Table 1 Overview of rating scales, questionnaires and observations.

When	What	How
Characteristics participants		
Beginning	* Background, experience (pretest)	questionnaire
	Occupational personality	questionnaire
	Leadership	questionnaire
	Memory	test
	Spatial orientation	test
Team measures		
After preparation	Mental effort	150 pt scale
During meeting	Behaviour	observation
End	Duration meeting	sec.
	Mental effort	150 pt scale
	Dominance	7 pt scale
	Info processing	4 items
	Leadership	4 items
	Process satisfaction	3 items
	Cohesiveness	5 items
	Work pace	4 items
	Communication	4 items
	Supporting behaviour	8 items
	Effectiveness	4 items
	Efficiency	7 items
	Outcome satisfaction	5 items
	Team satisfaction	2 items
Afterwards	Info transfer	# shared info
	Info outcome	# correctly applied info
	Quality product	Multiple Expert Assessment
Usability measures		
During preparation	Use of browser	logging
	Behaviour	observation
	* Usability browser 1	questionnaire
After preparation	* Usability browser 2	questionnaire
During meeting	Use of browser	logging
	Behaviour	observation
End	* Usability browser 3	questionnaire
	* Meeting room of the future (post test)	questionnaire

2.2.4 Procedure

The whole procedure took about 4 hours. First the test leader welcomed the participants in the Team Cockpit and explained the background and course of the experiment. Then the roles were divided between the participants. Subsequently the participants were asked to carry out two cognitive tests: an episodic memory test and a spatial orientation test. Episodic memory was tested, since participants in the experiment have to remember a lot of details concerning the design. If their episodic memory is good, they do not have to make use of the browsing software provided to them as often as people whose episodic memory is worse. Spatial orientation ability has been shown to have an impact on navigation behavior on the internet. People with a bad spatial orientation have a worse awareness of where they are and have been on the internet than people with a good spatial orientation. Following the tests participants were asked to fill in two questionnaires: a pretest questionnaire on their previous experiences with meetings, projects and ICT, and a questionnaire on the mental effort they experienced at that point in time.

Then the first phase of the scenario started. The participants were asked to take 15 minutes to familiarize themselves with the project, the previous team and their roles, and to make notes on relevant findings. They had to carry out this task individually, making use of the tools provided to them. After that, they

were given two questionnaires: again the mental effort questionnaire and a questionnaire on the usability of the tools they had used. After a 15 minute break, the second phase of the scenario started. The participants were asked to individually prepare for the upcoming meeting in the next 45 minutes. They had to follow instructions provided to them by email. They had to make use of tools to find specific information needed for preparation. Then, they were asked to fill in the mental effort and usability questionnaires again. The last phase of the scenario consisted of a meeting, which took 45 minutes. The participants were supposed to present their preparations, to discuss all available information and to finally come up with a clay prototype of the television remote control, which had to meet the requirements. Afterwards, they were presented with five final questionnaires: again mental effort and tool assessment, but also questionnaires on dominance (the participant's opinion on the dominance of team members) and team (the functioning of the team) and a general post-test questionnaire (on meetings, projects and the used tools).

2.3 Description of Browsers tested

The materials that were used for the experiment were all produced in the context of a previous similar scenario-based experiment that was carried out, and of which all meeting recordings are available in the AMI corpus (meetings ES2008a, b, c).

2.3.1 Condition 0: Baseline condition

Desktop

The desktop provides the following shortcuts (Figure 5):

1. Beamer
 - a. Allows to take control of the computer that is attached to the beamer in order to give presentations.
2. Shared Project Folder
 - a. A shared network folder with all the information from the previous team, including documents and **recordings of three previous meetings**.
 - b. Should be used to store and share all of the newly produced documents.
3. Internet Explorer
 - a. Can be used to access the Real Reaction corporate homepage.
 - b. Contains bookmarked websites that inspired the previous team.
4. Outlook
 - a. Contains all Email from the previous team.
 - b. Should be used to send mails to the personal coach, etc.
 - c. Mails with new information and instructions are received here.
5. Microsoft Office tools (Word, Powerpoint, Excel)
 - a. Word can be used to create and edit documents, notes, etc.
 - b. Powerpoint can be used to create and edit presentations.
 - c. Excel can be used to create and edit spreadsheets.
6. Recycle Bin
 - a. The Recycle Bin should not be used. All documents should be saved in the Shared Project Folder.

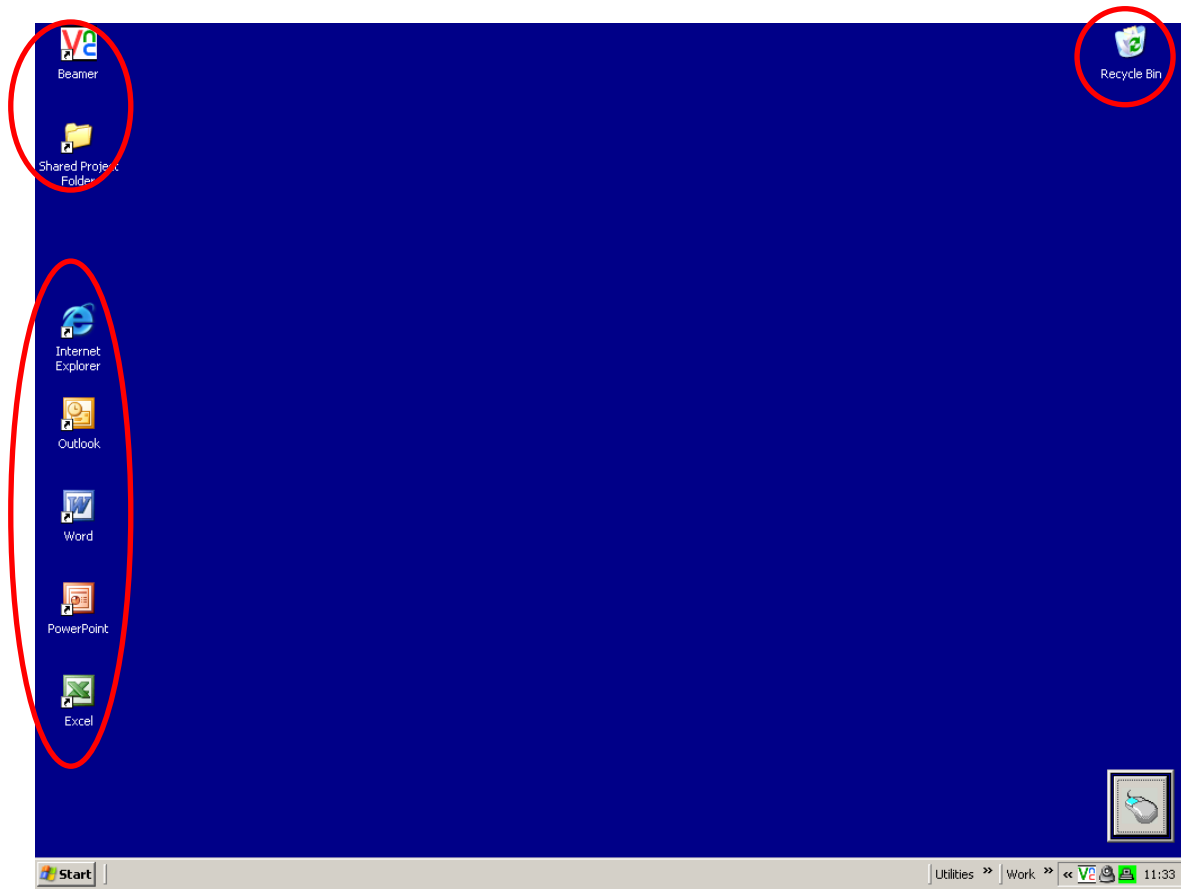


Figure 5 Desktop in the Baseline condition.

Shared Project folder

The Shared Project folder (Figure 6) contains folders for every previous meeting of the project, containing the documents and the meeting recordings that can be viewed via a regular media viewer. The folder should be used to store any new files.

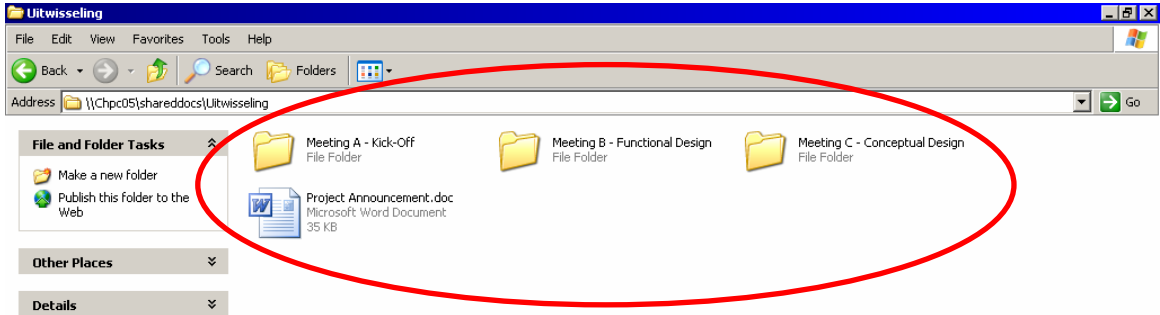


Figure 6 Shared project folder in Baseline condition..

2.3.2 Condition 1: Browser 1

Desktop

The desktop provides the following shortcuts (Figure 7):

1. Beamer
 - a. Allows to take control of the computer that is attached to the beamer in order to give presentations.
2. Shared Project Folder
 - a. A shared network folder with all the information from the previous team.
 - b. Should be used to store and share all of the newly produced documents.
3. **Meeting Browsers for 3 previous meetings (meetings A, B and C)**
4. Internet Explorer
 - a. Can be used to access the Real Reaction corporate homepage.
 - b. Contains bookmarked websites that inspired the previous team.
5. Outlook
 - a. Contains all Email from the previous team.
 - b. Should be used to send mails to your personal coach, etc.
 - c. Mails with new information and instructions are received here.
6. Microsoft Office tools (Word, Powerpoint, Excel)
 - a. Word can be used to create and edit documents, notes, etc.
 - b. Powerpoint can be used to create and edit presentations.
 - c. Excel can be used to create and edit spreadsheets.
7. Recycle Bin
 - a. The Recycle Bin should not be used. All documents should be saved in the Shared Project Folder.

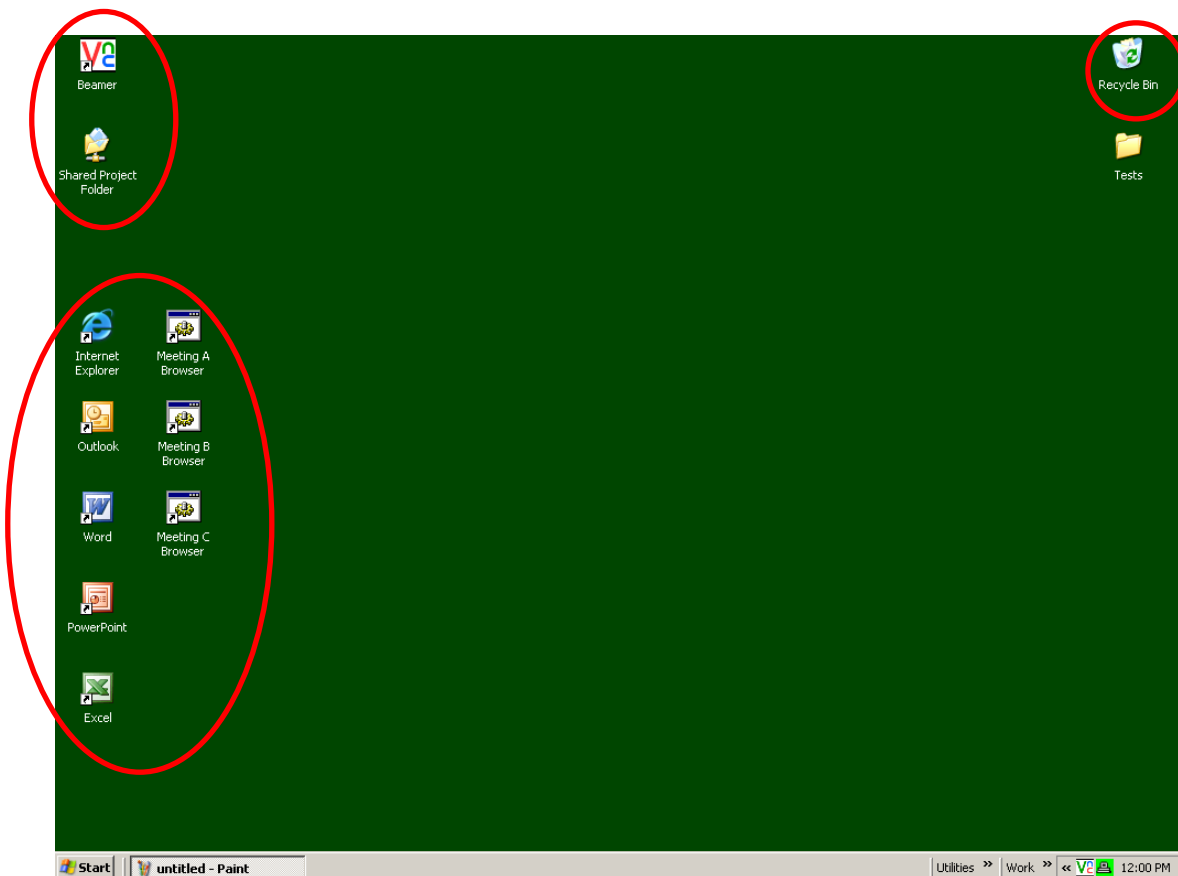


Figure 7 Desktop in Condition 1 and Condition 2.

Meeting Browser

The Meeting Browser provides the following functionalities (Figure 8):

1. Powerpoint presentations used during the meeting.
2. Video recordings of the meeting, you can switch between close-up shots of the four meeting participants (PM, UI, ME, ID).
3. Speaker activity log (who is speaking when), and indication of slide switches.
4. Transcript of all utterances.
5. 'Find': type in a word to search for this word in the transcript.

All information is synchronized, if you scroll through one of the windows, the other windows will change accordingly.

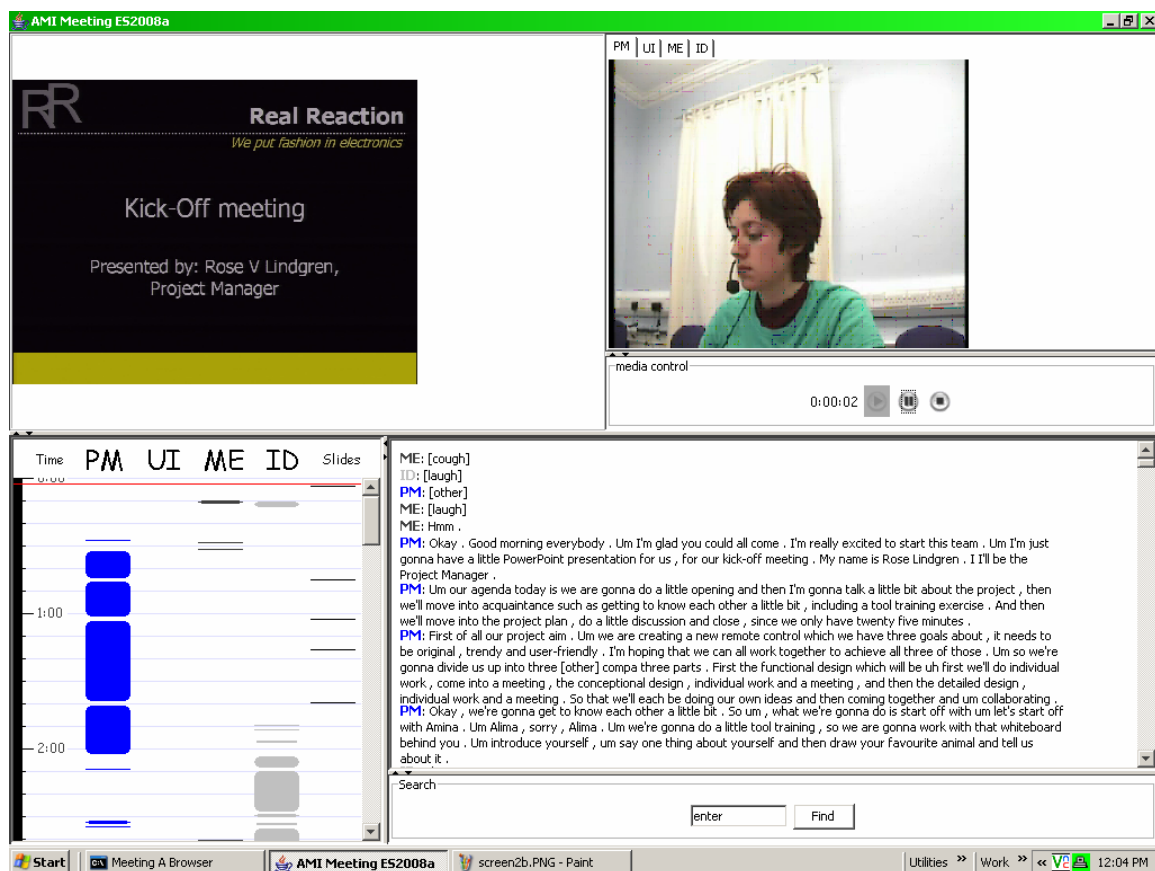


Figure 8 Meeting browser in Condition 1.

2.3.3 Condition 2: Browser 2

Desktop

The desktop in C2 is identical to the desktop in C1 (Figure 8). The meeting browser is different, however.

Meeting Browser

The Meeting Browser provides the following functionalities (Figure 9):

1. Powerpoint presentations used during the meeting.
2. Video recordings of the meeting, you can switch between close-up shots of the four meeting participants (PM, UI, ME, ID).
3. Minutes of the meeting, subdivided in Abstract, Actions, Decisions and Problems. If you click on a yellow area in the text, the transcript will show the corresponding yellow areas that were used as 'raw materials' for creating the text.
4. Speaker activity log (who is speaking when), and indication of slide switches.
5. Transcript of all utterances.
6. 'Find': type in a word to search for this word in the transcript.

All information is synchronized, if you scroll through one of the windows, the other windows will change accordingly.

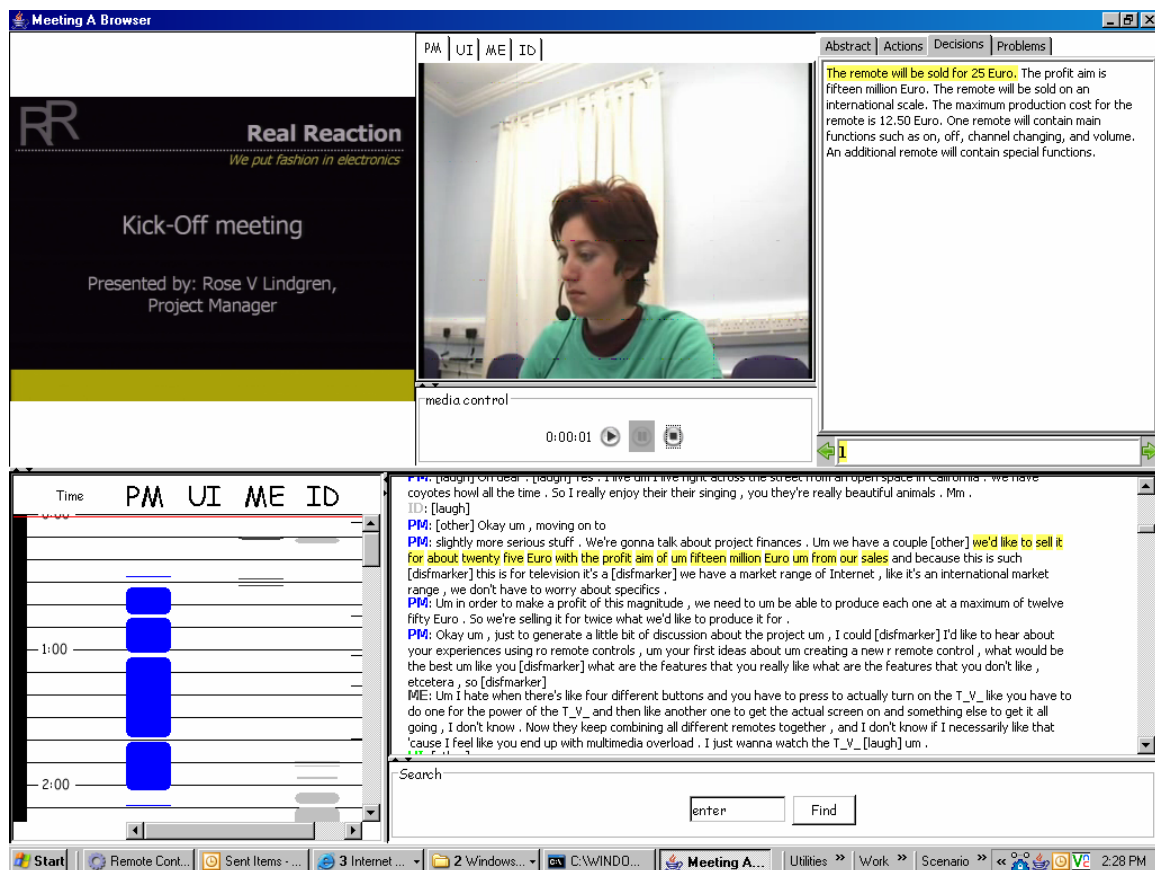


Figure 9 Meeting browser in Condition 2.

3. Results of BET

3.1 Conditions tested with the BET

We had usable data from 39 subjects on four conditions:

- Calibration (everyone)
- Base (15 subjects)
- Speedup (12 subjects)
- Overlap (12 subjects)

Screenshots from the three browser conditions tested are below. All conditions listened to audio through headphones:

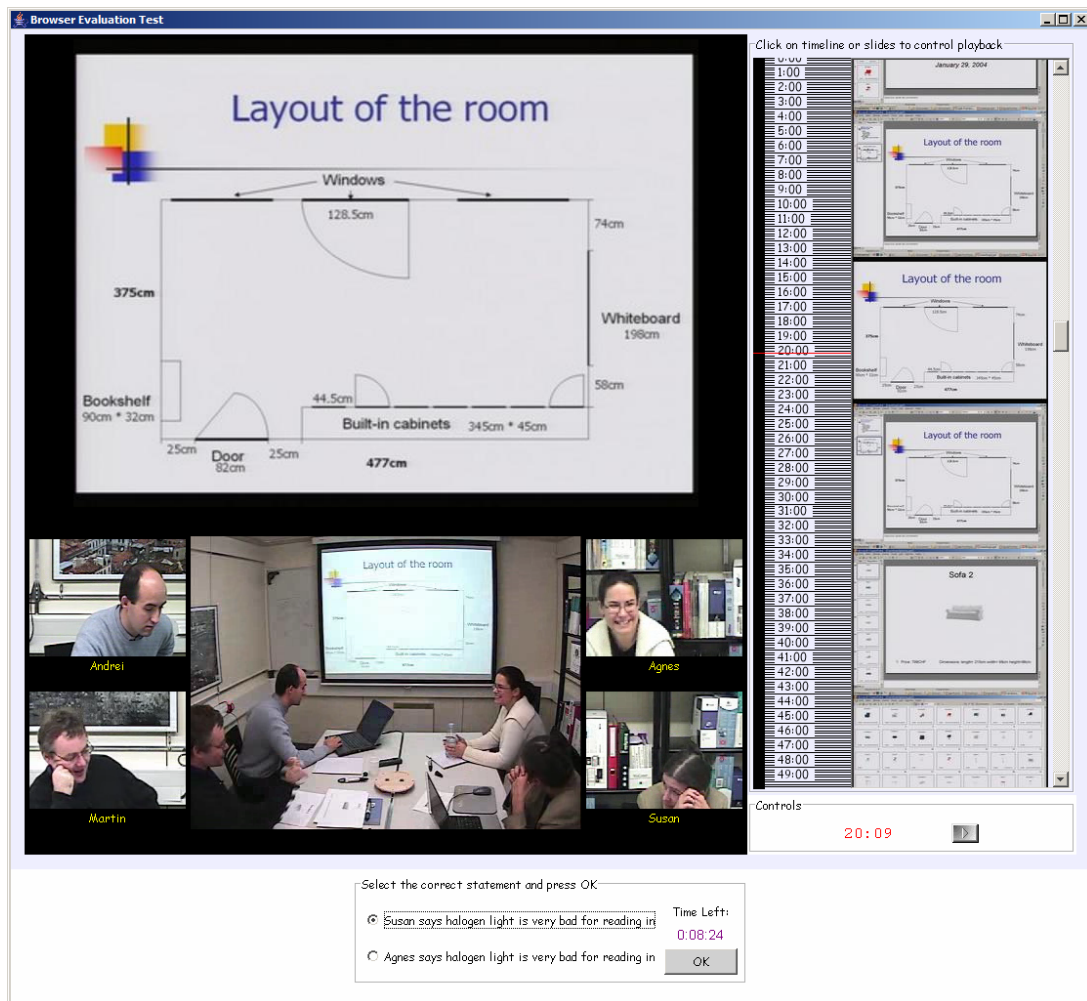


Figure 10 Subject Calibration Condition

The calibration condition resembles the type of browsers typically found in the industry today, and is likely to be somewhat familiar to most subjects. It presents a large slide view, 5 video views, a timeline, and slide thumbnails.



Figure 11: Base Condition

The base condition for these experiments played audio and included a timeline, scrollable speaker segmentations, a scrollable slide tray, and headshots with no live video.

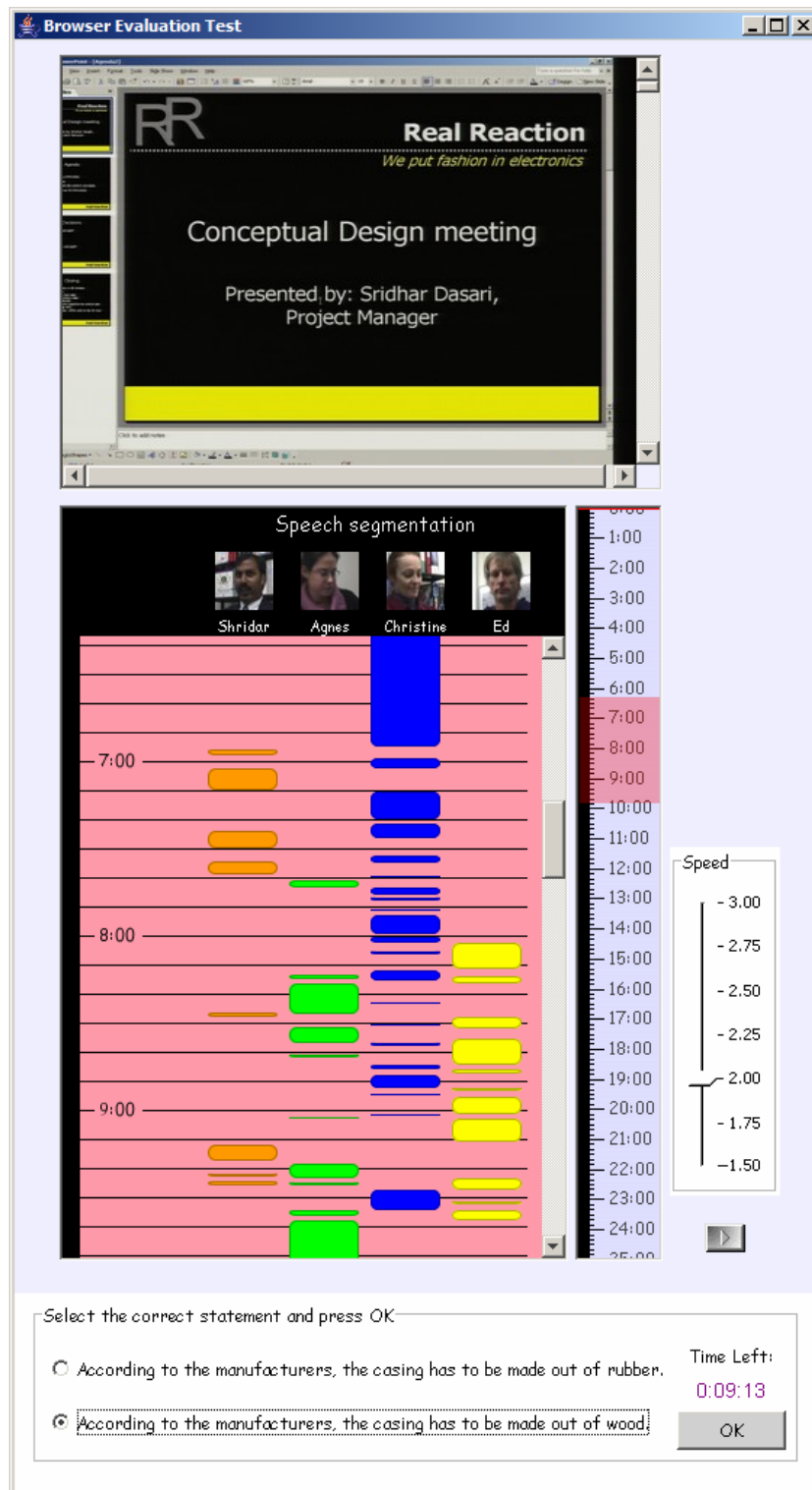


Figure 12 Speedup Condition

The Speedup condition was exactly like the base condition except that it required accelerated playback with a user-controlled speed between 1.5 and 3 times normal speed.

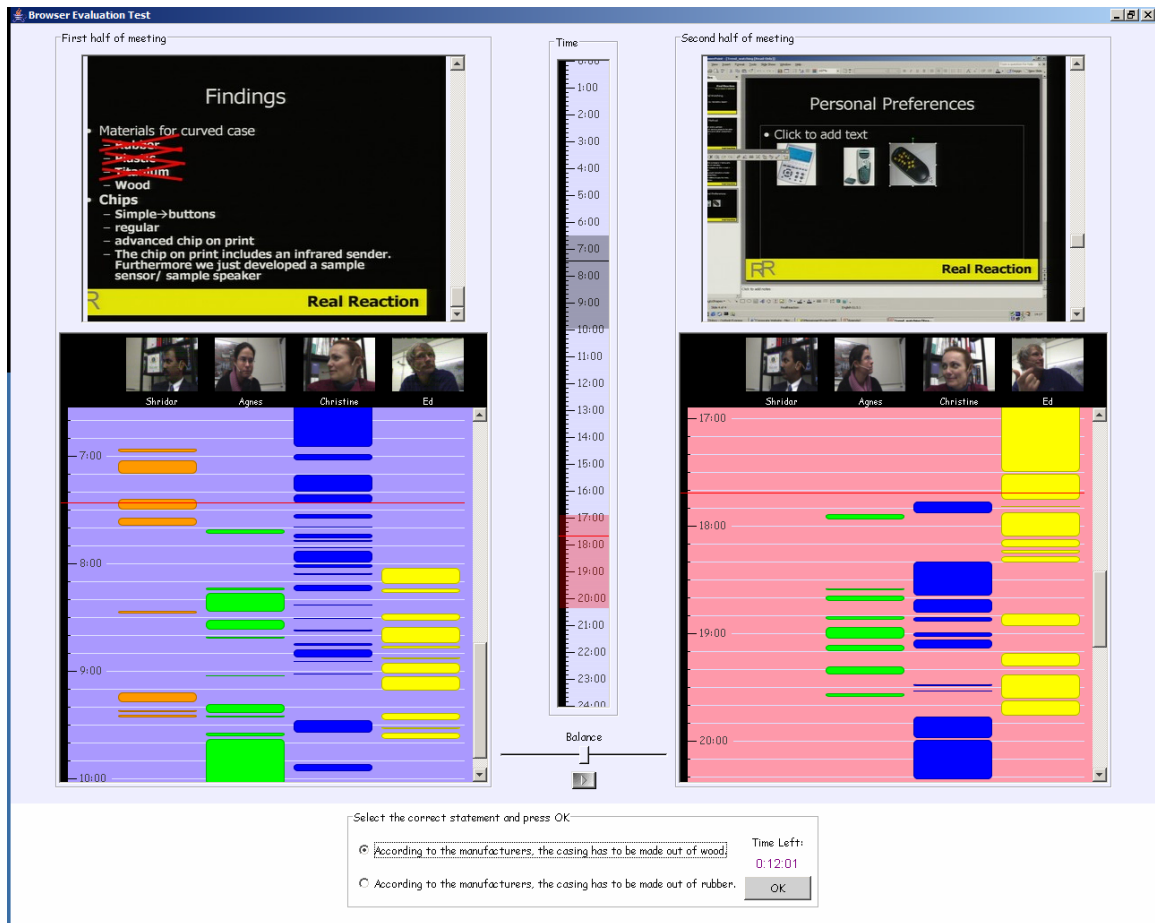


Figure 13 Overlap Condition

The Overlap condition duplicated the speedup condition with the first half of meeting on the left, with audio emerging from 45 degrees left (through headphones), and audio from the second half of the meeting emerging from 45 degrees right, simultaneously.

To achieve the 45 degree presentation of audio, the overlap condition used interaural time difference (see Figure 14 below).

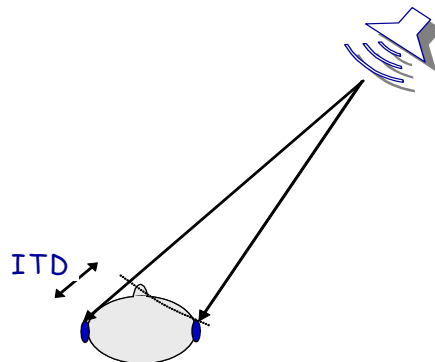


Figure 14: Interaural time difference

3.2 Raw uncalibrated results

Raw performance scores for both test meetings combined (but not accounting for the calibration task) were as follows for the three conditions.

Condition	Number of subjects	Mean Accuracy	Mean Speed (questions per min)
Base	15	77%	1.2
Speedup	12	83%	0.9
Overlap	12	74%	1.0

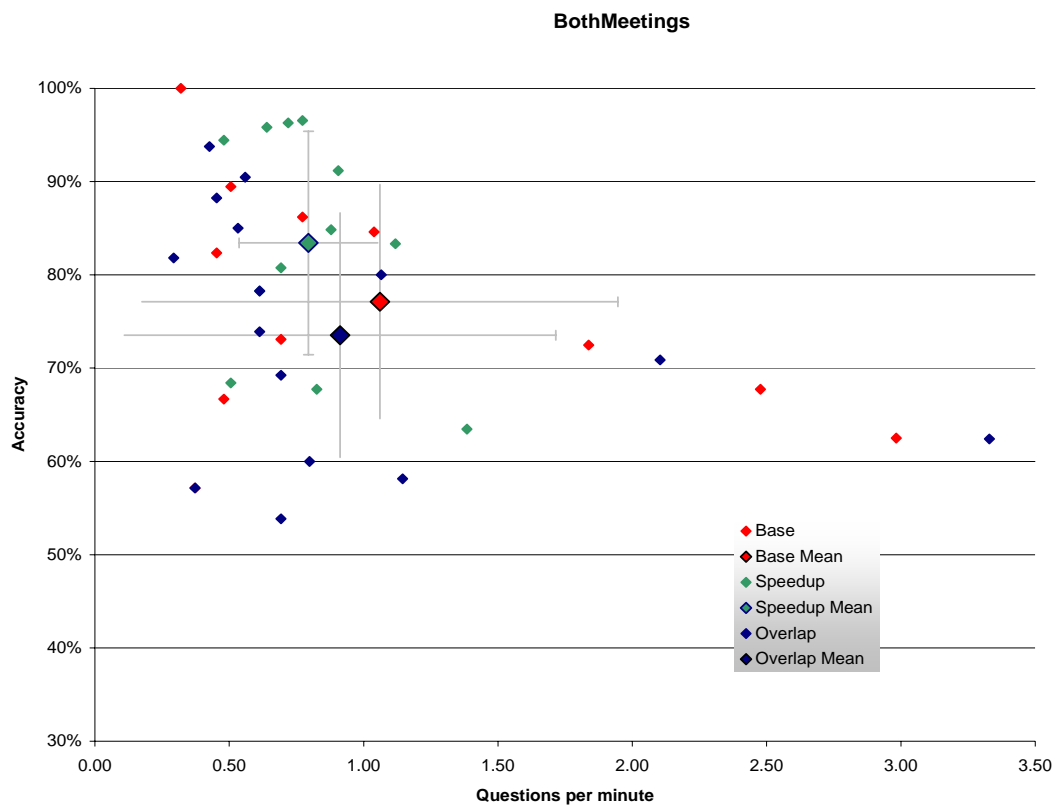


Figure 15: Graph showing speed and accuracy scores for all subjects and all conditions

3.3 The Speed/Accuracy Trade-off Model

3.3.1 Subjects' variable performance

Subjects have a fixed time to complete as many questions as possible, as correctly as possible – but where the balance lies between speed and accuracy is open to the subjects. Any individual subject might perform with huge variation.

When asked for high confidence, a subject might spend significant time finding and checking an answer. But, when pressed for time, a subject might make up their minds on the slimmest of evidence, or even guess. Thus, we might expect the same subject to exhibit a wide variety of behavior, ranging from slow-but-accurate through to fast-but-inaccurate.

While differences between subjects' aptitudes will account for some variability in the results, we suspect that much of the variation may be due to differences in where the subject lies on the speed/accuracy trade-off spectrum – even within the same test.

3.3.2 A simple model

We propose a model of the speed/accuracy trade-off, based on a simplified view of a subject's task. Each subject is looking for the answer to some questions, using a browser. The answer lies in the recording, let us assume, at one specific instant. If the subject decides to play a particular portion of the recording, the answer may or may not be found. If not, he or she might keep browsing for the answer, or give up and guess. The model fitted to the Calibrate condition is shown in Figure 16 below. With no help from a browser, the probability of a played segment containing the answer to a question is just random chance. However, a particular browser might be considered to boost this probability – an ideal browser might take the subject directly to the answer every time. Another browser might be better than random chance, but less than ideal.

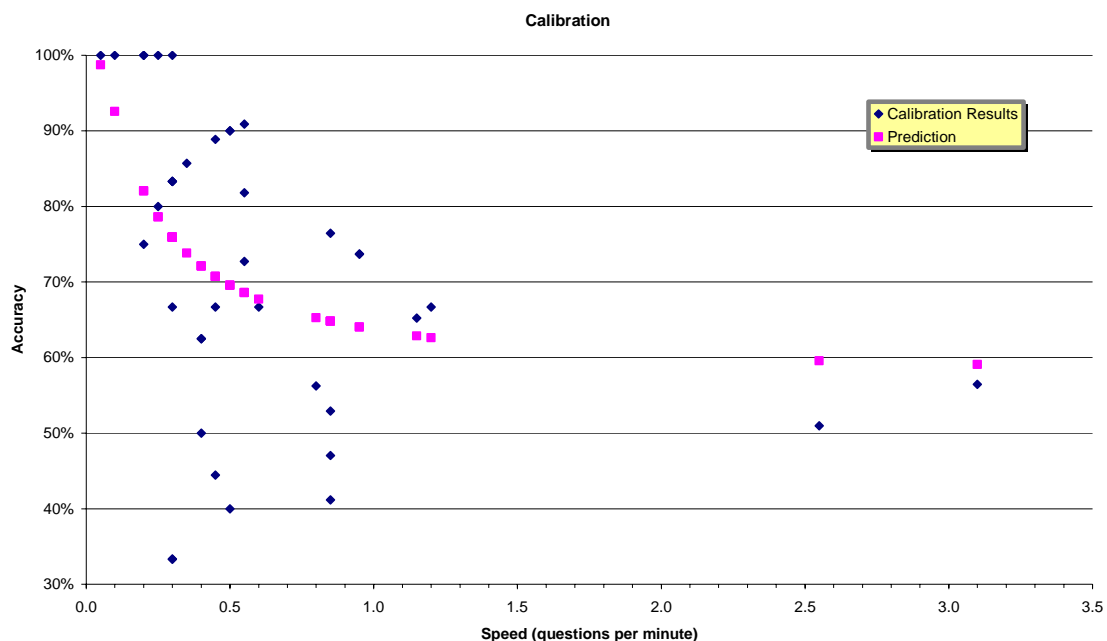


Figure 16. The speed/accuracy trade-off model applied to the Calibrate condition.

3.3.3 Comparing browsers

Thus, we can envisage a browser quality factor, by which a particular browser yields an advantage over purely random search. Such a factor, then, may provide a way of comparing browsers, without regard to the speed/accuracy trade-off. That is, any browser might be used by a diligent subject, prepared to spend much time answering questions, to obtain good accuracy. But the same browser in the hands of a

less patient subject might give poor results. The true merit of the browser lies in its ability to impart a greater speed for the diligent, but greater accuracy for the impatient.

3.3.4 Method

In more detail, the BET data may be analyzed as follows:

1. Each subject is given a browser, a recording, and a list of questions to answer.
2. We assume, naively, that the questions are independent, and that the subject does not accumulate knowledge of the media.
3. The answer to a particular question is assumed to lie at one particular point in the media.
4. For a given question, the subject uses the browser to decide on a segment of the recording to play. The probability of finding the right answer in one try is simply:

$$p_1(\text{answer}=\text{correct}) = \frac{QW}{L}$$

where W is the length of the segment played, and L is the length of the meeting. However, a browser may aid (or hinder) finding the answer, so the factor Q is used to model this. This factor is to be determined for each browser under test. For most browsers, we would hope $Q > 1$.

5. Whenever subjects find the answer to the question, they move on to the next question. However, if the answer is not found, they repeatedly use the browser to find and play segments of the recording. Thus, the probability of not finding the right answer, in a given number of tries i , is:

$$\begin{aligned} p_i(\text{answer}=\text{incorrect}) &= p_{i-1}(\text{answer}=\text{incorrect}) p_i(\text{answer}=\text{incorrect}) \\ &= p_i(\text{answer}=\text{incorrect})^i \end{aligned}$$

assuming the segments may or may not overlap, and the probability of finding the right answer after i tries is simply:

$$p_i(\text{answer}=\text{correct}) = 1 - p_i(\text{answer}=\text{incorrect})$$

6. After some time, the subject gives up and guesses the answer to the question. For simplicity, this time t is taken to be the mean amount of time spent on each question:

$$t = \frac{L}{2N}$$

where L is the length of the recording, N is the number of questions answered, and the divisor of 2 is due to the fact that BET tests are half the length of the recording. Assuming that the subject spends some time X between playing segments, and the segments are each of length W , then the subject has time for k tries:

$$k = \frac{t}{X + W}$$

7. When this time runs out, we assume the answer is guessed. The probability of being correct, p_{guess} , is given by the known likelihood from tests without media or browser of any kind. From previous experiments, this is known to be 56.7% [reference?].

8. Overall, an answer may be correct, either from finding the answer, or from guessing it. The accuracy is given by:

$$p(\text{answer}=\text{correct}) = p_k(\text{answer}=\text{correct}) + (p_{\text{guess}}) (p_k(\text{answer}=\text{incorrect}))$$

9. Thus we can predict the accuracy of each subject, knowing the number of questions answered during the BET test – given the quality factor Q for the browser. Conversely, we can determine the quality factor by fitting this model to the actual results of the tests, using least squares. Similarly, the other ‘constants’ of segment length (W) and time overhead between segments (X) can be estimated, either once and for all, or for each browser.

3.3.5 Results

Fitting the model to the Calibrate condition, we can obtain estimates for the segment length:

$$W \approx 25.11 \text{ seconds}$$

and the overhead between segments:

$$X \approx 7.49 \text{ seconds}$$

Using these two values for all other conditions, and fitting the model to each condition, we obtain estimates of the individual browser quality factors as shown in the table below.

Condition	Calibrate	Overlap	Speedup	Base
Quality Q	5.6	7.6	16.0	10.1

3.3.6 Validity of the model

Data collected for the BET has such high variability that it does not fit the trade-off model much better than simply using mean scores. It does fit *slightly* better, however, and this model might also make it possible to make use of the calibration test data, which was collected for such purposes. Evidence for this is described below.

3.3.6.1 Root mean square errors

The RMS error for actual accuracy compared to accuracy predicted by the tradeoff model is a little bit lower than the RMS error for actual accuracy compared to the mean accuracy. It was 15% for the model on meeting IS1008c compared to 14% for the mean. This is a very small difference in favor of the model.

3.3.6.2 Use of the calibration task

Every subject performed a calibration task in the hope that we could use their performance on the calibration task to normalize their results on the actual browser tests. Unfortunately, we found that normalizing test results according to performance on the calibration task led to even higher variability and a poorer fit of means and models to the data, so the value of the calibration task is questionable.

One possible explanation for this is that we had very low correlation between performances on the calibration task with performance on browser test. On IS1008c, for example, the correlation between the calibration task and the browser task was only 0.37 for the number of questions answered, and the correlation for accuracy was even lower, or 0.18

This result may be consistent with the model, however, because if each subject chose a different speed on the calibration test versus the browser test, then we would expect them to also have a different accuracy score. In order to explore this further, we measured the correlation of user accuracy on the calibration test with their accuracies on the actual test, but with actual accuracies adjusted to the same speed as the calibration test using the model. These adjusted scores correlate much better with a correlation of 0.51, which may indicate some advantage to using the model.

3.4 Questionnaire responses

In addition to taking performance measures, all subjects were given a short questionnaire at the end of their tests. The results of this questionnaire are reproduced below.

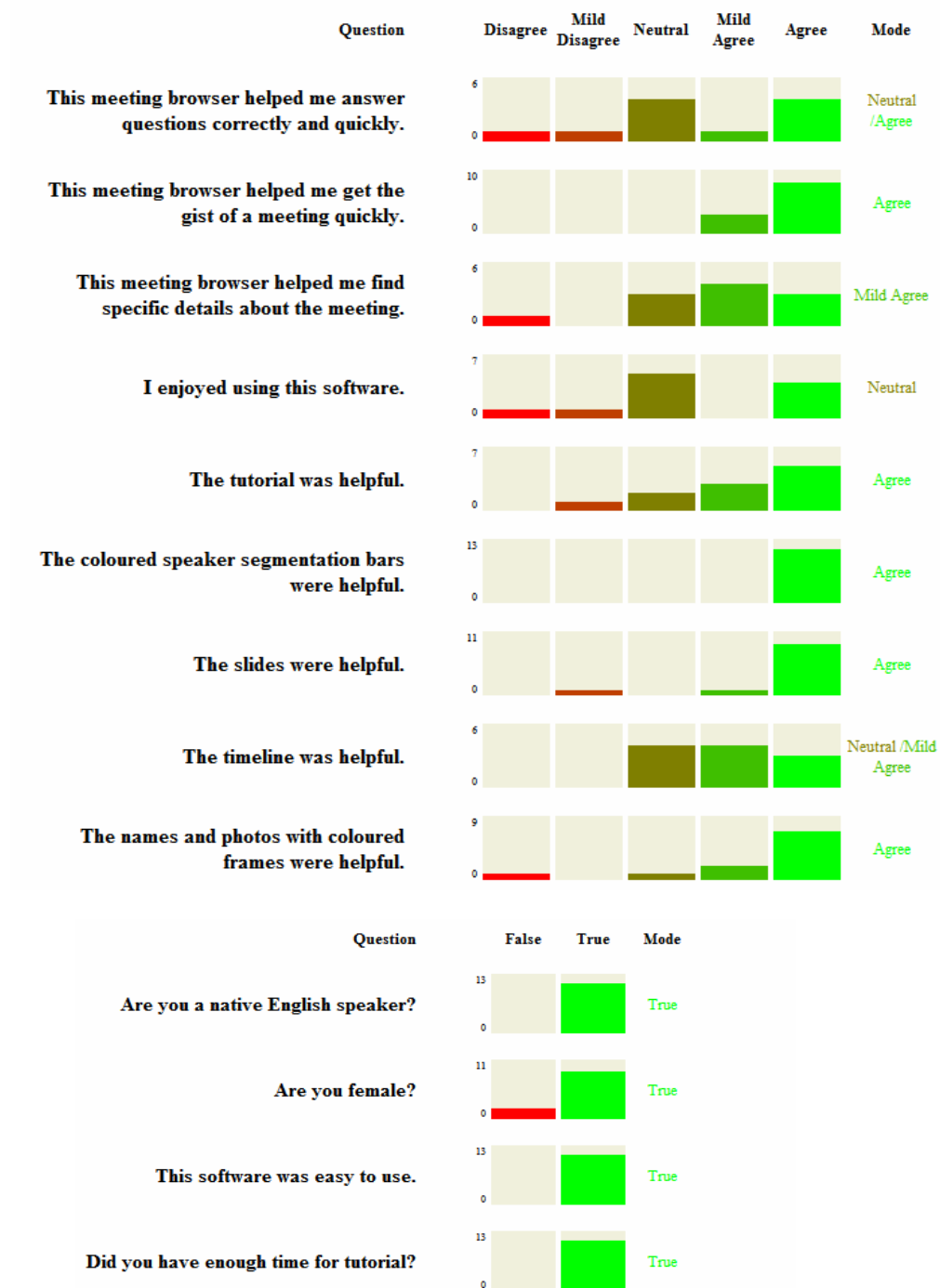


Figure 17: Questionnaire responses for Base Condition

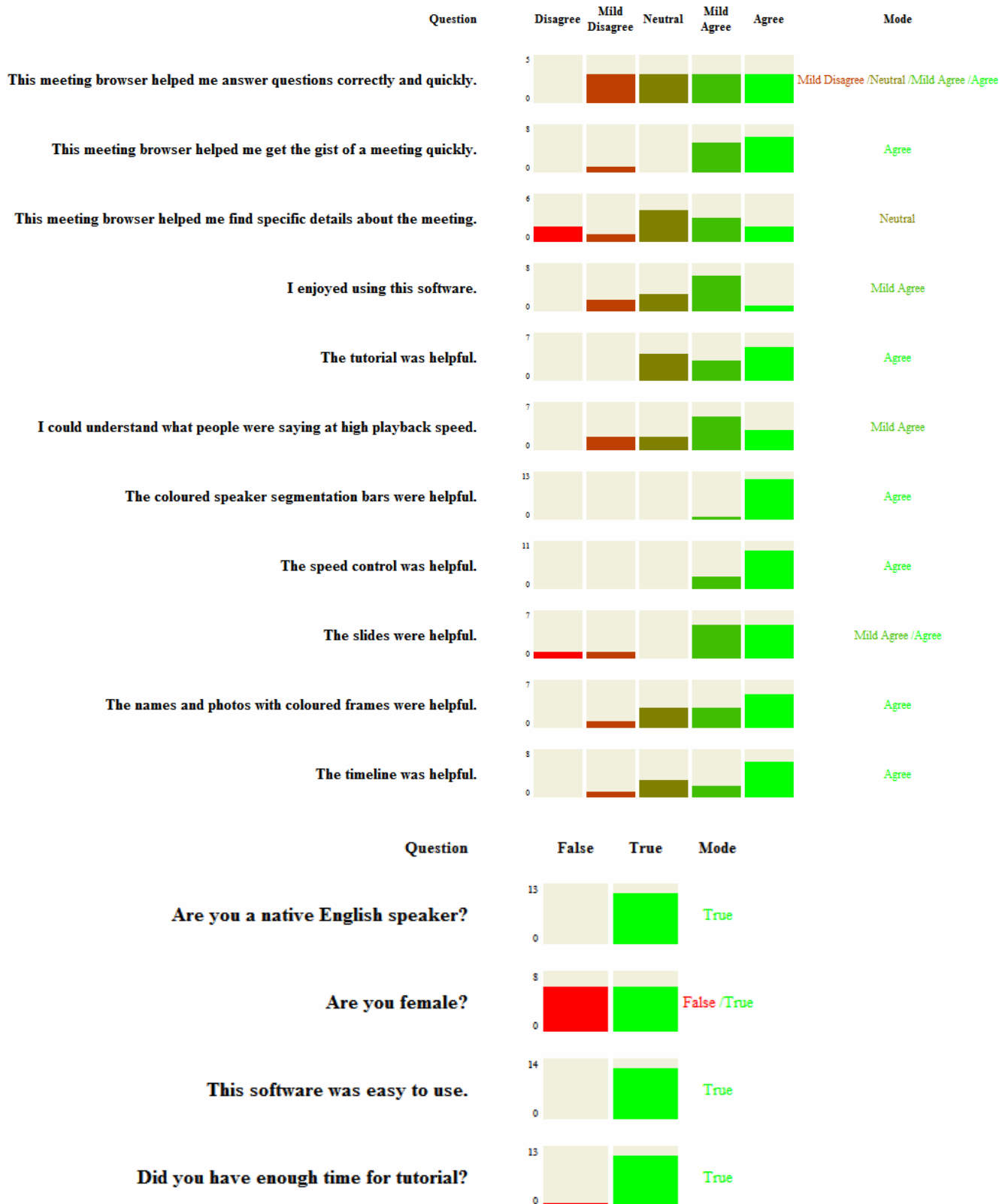


Figure 18: Questionnaire responses for Speedup Condition

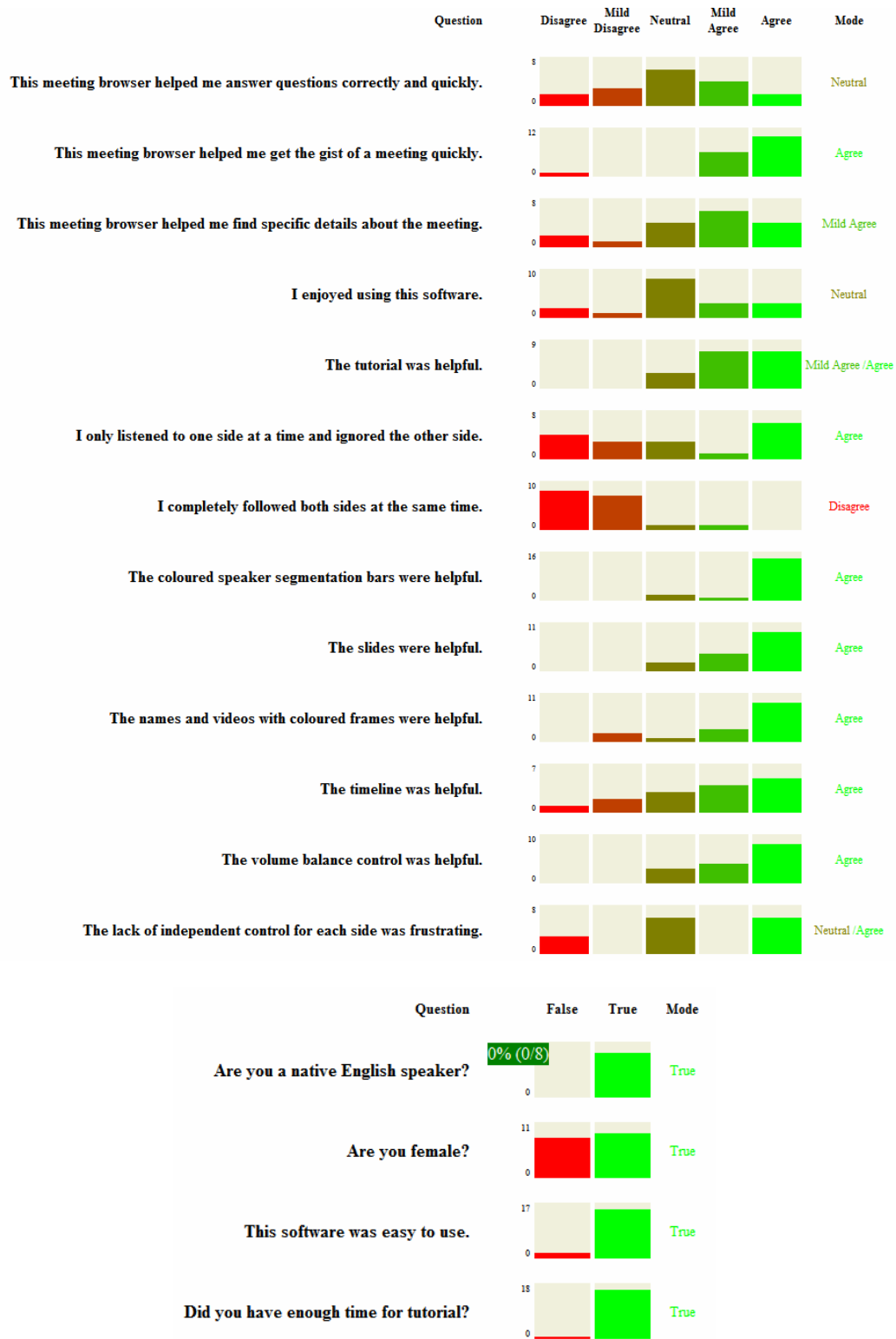


Figure 19: Questionnaire responses for Overlap Condition

3.5 Conclusions

The Speedup condition scores better than the base condition, and the Overlap condition scores worse. This ranking is preserved whether we compare simple means, or compare data for the three conditions fitted to the speed/accuracy trade-off model. In neither case, however, we do not have high confidence in the rankings because the data points have such high variability.

To improve the confidence of BET results, additional analysis or changes to the data collection methods may be required. One possibility to explore in the future is to gather more information from our subjects. The current BET collects approximately just one bit of key data per subject per minute, and some of those bits are guesses. Although detailed logs of browser interface activity are also collected, this data is more difficult to analyze in a structured manner.

This low number of data bits makes calibration difficult because small differences between subjects can cause such large adjustments. Approaches to collecting additional bits may include giving subjects a continuous range over which they can indicate confidence in their answers, or modifying the interface to encourage subjects to answer many more questions during the course of the test.

3.6 Acknowledgements

The design and implementation of the revised BET was heavily influenced by our colleagues Andrei Popescu-Belis, Agnes Lisowsky, and Denis Lalanne. Gerwin Van Doorn implemented the Speedup and Overlap browsers using the JFerret Browser Framework.

3.7 References

- [1] Wellner, P., Flynn, M., Tucker, S., and Whittaker, S. A meeting browser evaluation test. In *CHI '05 Human Factors in Computing Systems* (Portland, OR, USA, April 02 - 07, 2005).
- [2] Carletta, J.C., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, M., Post, W., Reidsma, D., and Wellner, P. (2006) The AMI Meeting Corpus: A Pre-Announcement. *Machine Learning for Multimodal Interaction: Second International Workshop, MIMI 2005*. Steve Renals and Samy Bengio, eds. Springer-Verlag Lecture Notes in Computer Science Volume 3869. ISBN 3-540-32549-2

4. Results of TBET

4.1 Introduction

In this report, results are presented of the pre-test questionnaire on background and experience of participants, of the three usability questionnaires (tool assessment) and of the post-test questionnaire, asking about general experiences, use of the tools and opinions on instrumented meeting rooms. All remaining measures that have been taken during the experiment will be reported on in a separate paper.

4.1.1 Pre-test

In total, 87 persons participated in the experiment. Their mean age was 23 years old (s.d. 7 years), 56% were male and 44% female. Almost all participants were students (97%). Most of them were students of computer science or information science (50%), 16% were students of psychology, 10% of philosophy and the rest of varying or unspecified subjects.

Participants were asked a number of questions on their patterns of computer use. Available answers were: never, monthly, weekly, daily (less than one hour), daily (between one and three hours) and daily (more than three hours). All participants use the computer on a daily basis. They also use the internet on a daily basis, equally varying between less than one hour, one to three hours and more than three hours. All also use email on a daily basis, half of them less than one hour, about one-third one to three hours, and the rest (one-sixth) more than three hours. About two-thirds of the participants chat daily, half of these less than one hour a day, the rest more than one hour a day. Most of the remaining one-third chats weekly, a small number chats monthly. Half of the participants search for multimedia content (music, video) on a daily basis, one third on a weekly basis, one-fifth on a monthly basis, the rest never. The following devices are used regularly by participants: laptop (76%), MP3-player (72%), GSM (40%) and PDA (16%).

Participants also received questions on their experience with meetings. Half of the participants participate in meetings on a weekly basis, one third on a monthly basis, one sixth never. Experiences with meetings could be expressed by five answers: never, hardly ever, sometimes, most of the times or always. More than half of the participants feel that most of the times objectives for their meetings are attained; one third feels that they are sometimes attained; one-sixth varies between the rest of the answers. About 80% of the participants feel that either sometimes or most of the time, time for their meetings is well-spent. The rest of the answers vary between the other options. About half of the participants feel that most of the times they like to participate in meetings. The others either like it sometimes or always. Hardly anybody never likes to participate.

The mean typical size of their meetings was six participants and they lasted for about one hour. Meetings are usually equally often attended in an educational and professional environment (both about 45%). The rest of the meetings is attended for leisure purposes (hobbies, charity). When asked to characterize their typical meetings, about 80% answer that the meetings have an informal atmosphere, about 10% formal. Other qualifications used are either positive (good: 6%) or negative (tedious: 4%). The large majority of the participants typically participate in meetings without having a specific role (94%). Smaller numbers of participants (15%) may also act as chairman or draw up the minutes (13%).

Participants were asked which means they typically use before (to prepare for the meeting), during and after meetings (to process the results) (percentages of participants). All percentages that differ substantially (about 20%) with respect to the other stages are bold.

	before	during	after
Personal recollection	48	43	60
Contact other participants	53	-	54
Related documents - use & annotate	56	45	53
Personal notes of the previous meeting(s) - Make personal notes	45	72	41
Consult external information sources (e.g. internet)	51	14	32
Minutes of the previous meeting(s)	33	25	31
Agenda	45	43	26
Contact external people (face-to-face, e-mail, telephone)	28	11	26
Other:	8	7	10
Pictures of previous meeting(s) - make pictures	5	9	6
Means to prepare a presentation - Give/discuss a presentation	15	24	2
Audio recording of previous meeting(s) - make recording	0	0	1
Video recording of previous meeting(s) - make recording	0	1	1
Make/discuss shared notes (e.g. on blackboard, whiteboard, flip-over)	-	44	-
Audio conferencing tools	-	0	-
Video conferencing tools	-	0	-

Participants typically include the following information in their personal notes (percentages of participants):

Things to do	89
Reminders	77
Decisions taken	58
Reference materials (names, phone number, webpages)	56
Things you want to tell others	47
Doodles (absent-minded scribbles)	33
Other	4

When participants have missed a meeting, they catch up in the following ways (percentages):

Ask other participants	89
Read meeting minutes	45
Consult notes of other participants	30
Consult video recording	2
Consult audio recording	1
Other:	8

Finally, participants were asked about their experience with working in project teams, which could be expressed by four answers: no, hardly any, average, or a lot. About 80% has either hardly any or average experience with working in project teams. The rest has either no experience or a lot of experience. About half of the participants have no experience at all in product or service development. The other half has hardly any or average experience.

4.1.2 Tool assessment

At three points during the scenario participants were asked their opinion on the ICT tools that were available to them: (1) after having familiarized themselves with the project, the previous team and their roles; (2) after having individually prepared for the upcoming meeting; and (3) after the meeting (at the end of the scenario). Questions were asked referring to the usability aspects effectiveness, efficiency and satisfaction. For every aspect four questions were asked, in a random order. Participants answered the questions on a seven-point scale, varying from not applicable at all (1) to very much applicable (7).

Mean total scores per assessment and per condition were calculated. All mean scores varied between 3.38 and 5.08, indicating general moderate usability. There were no differences between scores for effectiveness, efficiency and satisfaction.

Effects of point of measurement and condition occurred. Mean scores for all three aspects over conditions were all lowest for measurement point 1, and highest for point 3, indicating an increased perceived usability the longer the tools had been used. This can be considered a learning effect. Mean scores for all three aspects over the three measurement points were all lowest for condition 0 and highest for condition 2, indicating a better perceived usability if 'richer' tools are offered.

effectiveness			efficiency			satisfaction		
Tool 1	Tool 2	Tool 3	Tool 1	Tool 2	Tool 3	Tool 1	Tool 2	Tool 3
3.95	4.11	4.17	3.91	4.04	4.14	3.98	4.13	4.42
C0	C1	C2	C0	C1	C2	C0	C1	C2
3.45	4.26	4.62	3.67	4.07	4.40	3.74	4.18	4.67

4.1.3 Post test

Afterwards participants were asked about their general experiences, their use of the tools and their ideas about meeting in an instrumented meeting room.

Questions about general experiences dealt with the following issues. They were asked whether they found the objectives had been attained, whether the time had been well-spent and whether they had liked to participate in the project, on a five-point scale (never, hardly ever, sometimes, most of the times, always). Mean answers to these questions were about the same and rather positive, around 3.8. There were no differences between the three conditions.

Questions about the use of the tool dealt with the usefulness of the information on the computer, the usability of the tools that were offered to them, what they had missed in terms of information and search options, and the trust in the information they had experienced.

They were asked to indicate (on a seven-point scale) how *useful* they found the different types of information on the computer. The following mean answers were given for the three conditions:

	C0	C1	C2
Minutes	5.3	5.5	-
Presentations	4.6	5.0	5.0
E-mail / messages	3.7	4.4	5.4
Internet information	3.1	3.4	3.5
Close-up videos	3.1	3.3	3.8
Speaker activity log	-	4.2	4.0
Meeting transcripts	-	4.6	4.9
Abstracts	-	-	5.0
Actions	-	-	5.1
Decisions	-	-	5.8
Problems	-	-	5.5
Total mean	4.0	4.3	4.8

The general usefulness of the information is rated average. Interestingly, mean ratings for the three conditions show a small increase of usefulness in 'richer' conditions. For the types of information that had been available in all three conditions (presentations, email, internet, close up videos), the usefulness was also rated higher for 'richer' conditions. Interestingly, the usefulness of abstract, actions, decisions

and problems (which together form the minutes) in C2 was rated high, but comparable to the rating of minutes in C0 and C1.

Participants were also asked how *usable* they found the different tools (on a seven-point scale). The following mean answers were given for the three conditions:

	C0	C1	C2
File system (Explorer)	4.4	4.5	4.7
E-mail (Outlook)	4.0	4.4	5.0
Internet	3.4	4.0	3.9
Meeting browser	-	4.3	5.1
Total mean	3.9	4.3	4.7

Usability was rated at the same, average, level as usefulness. Mean ratings for the different conditions were also comparable to these for usefulness, showing a small increase of usability in 'richer' conditions.

In addition, participants were asked which information they had actually used, and which tool they had used to access the information. Results (in percentages of participants) for File system (F), Email (E), Internet (I) and Browser (B) are presented below. Hyphens indicate possibilities that were not offered in certain conditions. Light grey cells indicate inaccurate answers: answers that were given, but were not actually possible. Unfortunately, a lot of these answers were given, indicating that possibly participants had not fully understood the question. Expectations were that participants would make less use of the file system if they had the possibility of using the meeting browser (in C1 and C2). Instead, it is clear that the meeting browser is used extensively, but that the file system is still used as much as in C0. The minutes that were used extensively in C0 and C1, were replaced by abstract, actions, decisions and problems in C2. The minutes were still available through the file system, and were structured according to these four topics, explaining the high scores for these topics in File system in C2.

	Condition 0				Condition 1				Condition 2			
	F	E	I	B	F	E	I	B	F	E	I	B
Minutes	94	16	6	-	89	4	4	14	-	-	-	-
Presentation	94	34	6	-	79	21	7	43	86	32	29	46
E-mails / messages	22	88	3	-	11	89	11	0	39	89	18	21
Internet	19	9	47	-	4	7	64	0	39	0	39	7
Close up videos	59	0	3	-	7	0	0	54	14	0	4	43
Speaker activity log	-	-	-	-	7	0	0	54	21	0	0	54
Meeting transcripts	-	-	-	-	18	4	0	64	46	7	4	54
Abstracts	-	-	-	-	-	-	-	-	50	0	4	46
Actions	-	-	-	-	-	-	-	-	32	4	4	61
Decisions	-	-	-	-	-	-	-	-	50	7	7	57
Problems	-	-	-	-	-	-	-	-	39	25	4	46

Participants were asked what type of information they missed on the computer. About half (48%) of the participants indicated that they had not missed any information. The other half provided various suggestions for useful information they had missed, including: a better insight in the progress so far (including what the previous team had accomplished), an overview of decisions that had been taken, design sketches made so far, an overview of actions points for all team members, and easy access to documents of other team members. Finally, they wanted this information to be organized in a structured way, for instance chronologically.

Participants were also asked what type of search options they had missed on the computer. The majority (66%) did not miss any search options. Suggestions that were given by the remaining 34% included: global search through all available information (Google style), searching through emails, and better search options for the transcripts. There were no differences between the three conditions, except that searching through transcripts was only mentioned in C1 and C2.

Finally some issues related to using this type of information in 'real life' were addressed. Of the participants, 94% indicated that they had trusted the information on the computer. A large majority of the participants (83%) indicated they would like to participate in meetings that take place in an instrumented meeting room, although 73% indicated that this would affect their behavior, knowing that all communication is being logged.

4.2 Conclusions and further work

Results so far indicate that meeting browsers add to the perceived usefulness and usability of project-related information. When meeting browsers are available, users rate both the information offered and the usability of consulting this information higher, in terms of effectiveness, efficiency and satisfaction. Information that is available in the three versions of the browser is rated more useful and usable in browser conditions than in the baseline condition. When a meeting browser is available it is used as a complement to the other information, not as a replacement. The scores for both usefulness and usability were average, though, which means there is certainly room for improvement. Several suggestions for improvement of the tools were given, important ones being organizing the information in one well-structured tool, including action points for team members and decisions taken so far, and offering global search possibilities to access this information.

Participants in this experiment were rather young and relatively inexperienced with meetings and projects, which may have influenced the results in the sense that they are probably not fully aware of the requirements for meeting and functioning in a project team. They were experienced with using the computer, though, which suggests potential skills for using the tools and trust in the information offered. This could be a reason why they were quite apt at using the tools and were not really intimidated by the prospect of meeting in an instrumented meeting room.

The results reported so far are not complete, in the sense that they have not addressed yet the influence of using the different tools on team and project aspects, which were listed in Table 1. Once these results are analyzed it will be possible to answer the question whether and how a multimodal meeting browser improves a meeting, and consequently might lead to a more efficient and satisfactory project process and higher quality results.

The ideas for improvement mentioned above have already been implemented in a so-called 'project browser' concept, in which all project-related materials are integrated from the perspective of a user in a task setting (i.e. carrying out a role in a design project) (Cremers, Groenewegen, Kuijper and Post, 2007). This project browser will shortly be subjected to the identical task-based evaluation. Results will be compared to the results reported here.

4.3 References

- Rogelberg, S. G., Leach, D. J., Warr, P. B. & Burnfield, J. L. (in press). "Not another meeting!" Are meeting time demands related to employee well-being? *Journal of Applied Psychology*.
- Cremers, A.H.M., Groenewegen, P., Kuijper, I & Post, W.M. (2007). The Project Browser: supporting information access for a project team. Accepted for presentation at HCII 2007, Beijing, July 22-27.
- Post, W.M., Cremers, A.H.M. and Blanson Henkemans, O. (2004). A research environment for meeting behavior. In: A. Nijholt & T. Nishida (Eds.). *Proceedings of Social Intelligence Design 2004*, p. 159-165.
- Post, W.M., Elling, E., Cremers, A, & Kraaij, W. (2007). Experimental comparison of multimodal meeting browsers. Accepted for presentation at HCII 2007, Beijing, July 22-27.
- Post, W.M., Huis In't Veld, M.A.A. and van den Boogaard, S.A.A. (2007). Evaluation Meeting Support Tools. In: *Personal and Ubiquitous Computing*. Accepted for publication.

Conclusion

This deliverable described the different methods for evaluating Meeting Browsers and Task-based Meeting browsers using BET and TBET definitions as described in chapter 1 and chapter 2. Some very interesting user requirements, for our AMI meeting browsers, came up during the user tests, which will be input for further evaluations during the AMIDA project.

Appendix 1: Observer Instruction pages

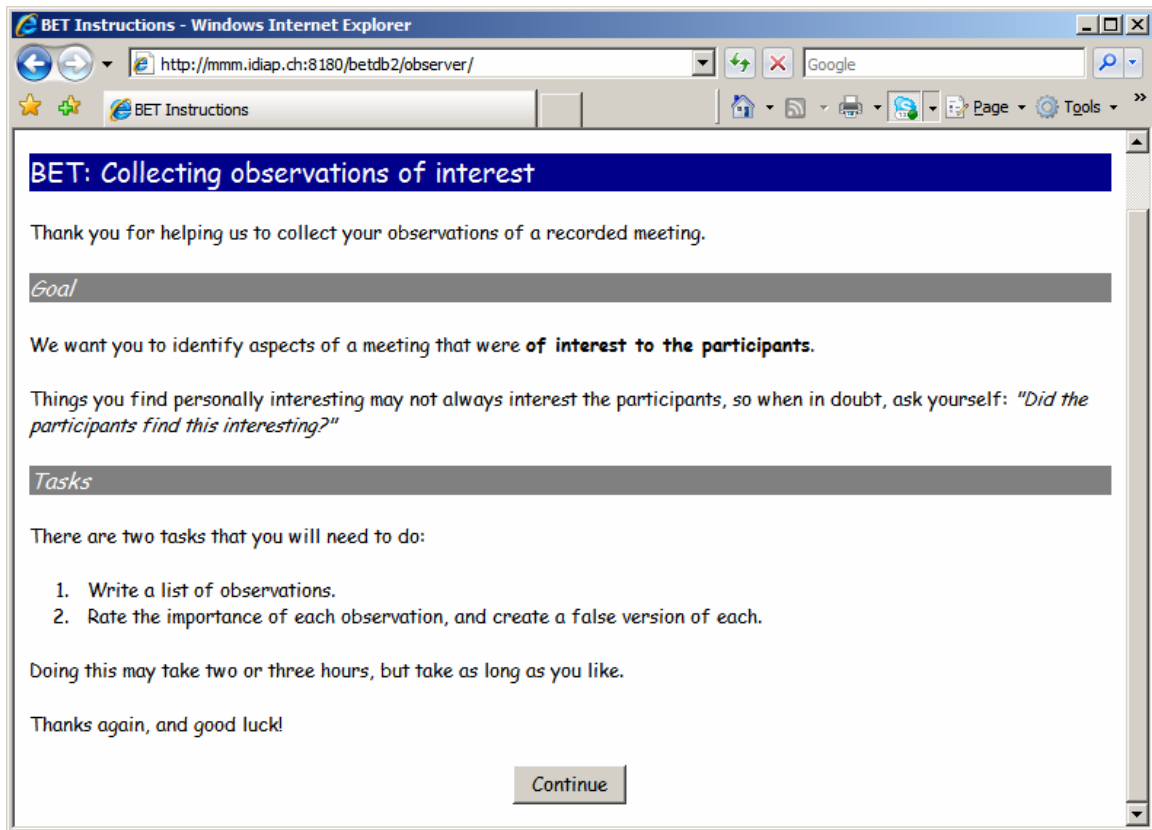


Figure 20: Observer Introductory Instructions

Observer registration confirmation

You are now registered as: **example@idiap.ch** observing meeting IS1008c.

Observer Instructions for Task 1: Writing a list of observations

1. Watch the meeting all the way through (about 50 minutes), perhaps making notes on paper.
2. Try to decide upon 20-40 points of interest in the meeting.
3. Replay any part of the meeting as often you like.
4. For each point of interest that you find, make sure the player is in the correct position, and select one of the three radio buttons:
 - o "Here" if your observation is pertinent to that particular moment;
 - o "Around" if your observation covers at least a minute of the meeting around the point you have selected;
 - o "Throughout" if your observation broadly covers the whole meeting.
5. Review your list of observations. You may delete incorrect entries at any time.
6. When you are happy with your list of observations, click on the 'Go to task 2' button.

Example observations

- The group decided that the Honda minivan was better than the Dodge minivan.
- Mike left the room in the middle of the meeting to answer a phone call.
- Four cars were discussed during the meeting.
- Three articles about cars were used as references during the meeting.
- The date for the next meeting is January 21st.
- Mike and John agreed that the Lotus Elan is not as good as the Ferrari.
- The main topic of the meeting is the top five cars of the year.
- John laughed as the meeting began.
- Mary thinks that the Ferrari is the best sports car.
- The Ferrari Spider was one of the top two choices.

Figure 21: Observer Task 1 Instructions

BET observations by demo@amiproject.org on meeting IS1008c

Remember:

- Interesting to participants.
- Player positioned correctly.

Review the [full instructions](#).

True statement:

According to the manufacturers, the casing has to be made out of wood.

Scope:

Here
 Around
 Throughout

[See all my observations](#) | [Go on to Task 2](#)

Figure 22: Observer Task 1 screenshot

Task 2: Creating false statements and ranking importance

You will now create an *false* complement for each of your observations. In the future, other people will be shown both of your statements and asked to find which is true, by browsing the meeting.

Before submitting your false statement, you will also rank the importance of the observation, from low to high, relative to the whole meeting and your other observations.

Some hints and guidelines

- Both statements should be expressed as **positive** phrases whenever possible. Avoid just adding the word "not".
- Make sure that your false statement is in fact not true at any time.
Example:
True statement: Peter asked a question.
False alternative: Mike asked a question.
Problem: Either of these statements may be true at different times.
- Remember to check your negative sentences for "guessability". Example:
True statement: John laughed as the meeting began.
False alternative: John cried as the meeting began.
Problem: Too easy to guess without viewing the meeting.

Acceptable examples

Here are some example **true statements** from Task 1, followed by several possible alternative **false statements**, which are all acceptable. The words that differ are **bold**, and the replacement is underlined in the false version.

You will only create a **single** false version of each statement.

True statement: **The whole group** decided that the **Honda** minivan was a better car than the Dodge minivan.
False alternatives: Only one person decided that the Honda minivan was a better car than the Dodge minivan.
The whole group decided that the Toyota minivan was a better car than the Dodge minivan.

True statement: **Mike** left the room in **the middle** of the meeting to **answer a phone call**.
False alternatives: Mike left the room in the middle of the meeting to get a drink.
Mike left the room at the beginning of the meeting to answer a phone call.
Mary left the room in the middle of the meeting to answer a phone call.

True statement: **Four** cars were discussed during the meeting.
False alternative: Two cars were discussed during the meeting.

Figure 23: Observer Task 2 instructions

True statement: **Four** cars were discussed during the meeting.
False alternative: Two cars were discussed during the meeting.

True statement: **Three printouts** of articles about cars were presented.
False alternatives: Four printouts of articles about cars were presented.
Three electronic versions of articles about cars were presented.

True statement: The date for **the next** meeting is **January 21st**.
False alternatives: The date of the previous meeting was January 21st.
The date of the next meeting is February 22nd.

True statement: **Mike** and John **agreed** that the Ferrari is better than the Lotus.
False alternatives: Mike and John disagreed whether the Ferrari is better than the Lotus.
Mary and John agreed that the Ferrari is better than the Lotus.

True statement: The main topic of the meeting is the **top five family cars**.
False alternative: The main topic of the meeting is Ferrari cars.

True statement: **John laughed** as the meeting **began**.
False alternatives: Peter laughed as the meeting began.
John coughed as the meeting began.
John laughed as the meeting ended.

True statement: **Peter** thinks that the **Ferrari** is the **best** sports car.
False alternatives: Peter thinks that the Lotus is the best sports car.
Peter thinks that the Ferrari is the worst sports car.
John thinks that the Ferrari is the best sports car.

True statement: The **Ferrari Spider** was one of the top five choices.
False alternative: The Lotus Elan was one of the top five choices.

Note that in some of the examples above, it's possible for *both* statements to be true (not mutually exclusive). You must make sure that in the actual recording, your true statement is always true and your false statement is always false because subjects will not know where they refer to in the meeting.

Note also that it's ok to edit your true statement if you notice a problem with it, or to help clarify your observation pair.

Rating importance

You will also rate the **importance** of each observation from low to high (one to five stars). Remember, this is a rating of importance **to the participants**, not just to you.

Figure 24: Observer Task 2 instructions (continued)

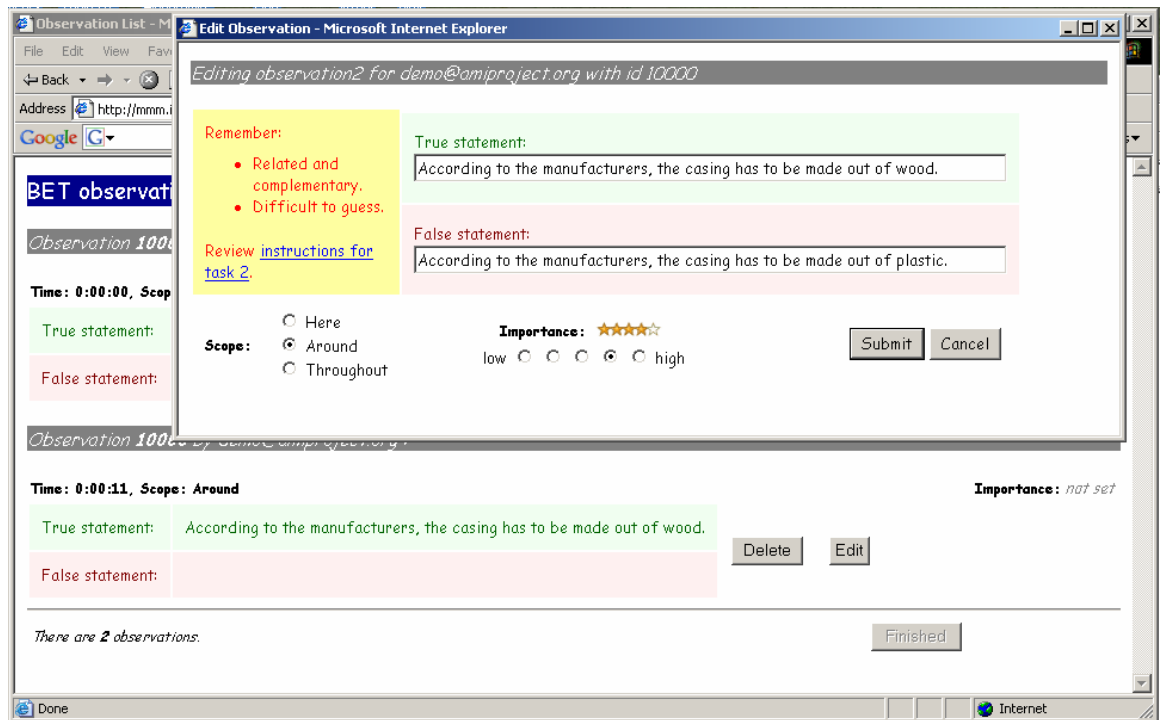


Figure 25: Observer Task 2 screenshot

Appendix 2: Observation Editing pages

BET Admin Observations for meeting IS1008c - Windows Internet Explorer

http://mmm.idiap.ch:8180/betdb2/observer/adminObservationsByImportance.js

BET Validated observations for meeting IS1008c, by group size and adjusted importance

Group Size: 6

#463 Group: woodReq Status: A Median adjusted importance: 0
Time: 0:04:53, Scope: Around **Importance: ★★★★★**

True statement:	According to the manufacturers, the casing has to be made out of wood.
False statement:	According to the manufacturers, the casing has to be made out of rubber.

originalTrue: *according to the manufacturers, the casing has to be made out of wood!*
 originalFalse: *according to the manufacturers, the casing has to be made out of rubber*

#345 Group: labourEthics Status: A Median adjusted importance: 0
Time: 0:20:11, Scope: Around **Importance: ★★★★★**

True statement:	Christine is considering cheaper manufacture in "other countries" before backtracking and suggesting the remote could support a premium price by promising to pay a decent wage to the workers producing it.
False statement:	Ed is considering cheaper manufacture in "other countries" before backtracking and suggesting the remote could support a premium price by promising to pay a decent wage to the workers producing it.

There are 2 observations of group size 6.

Group Size: 5

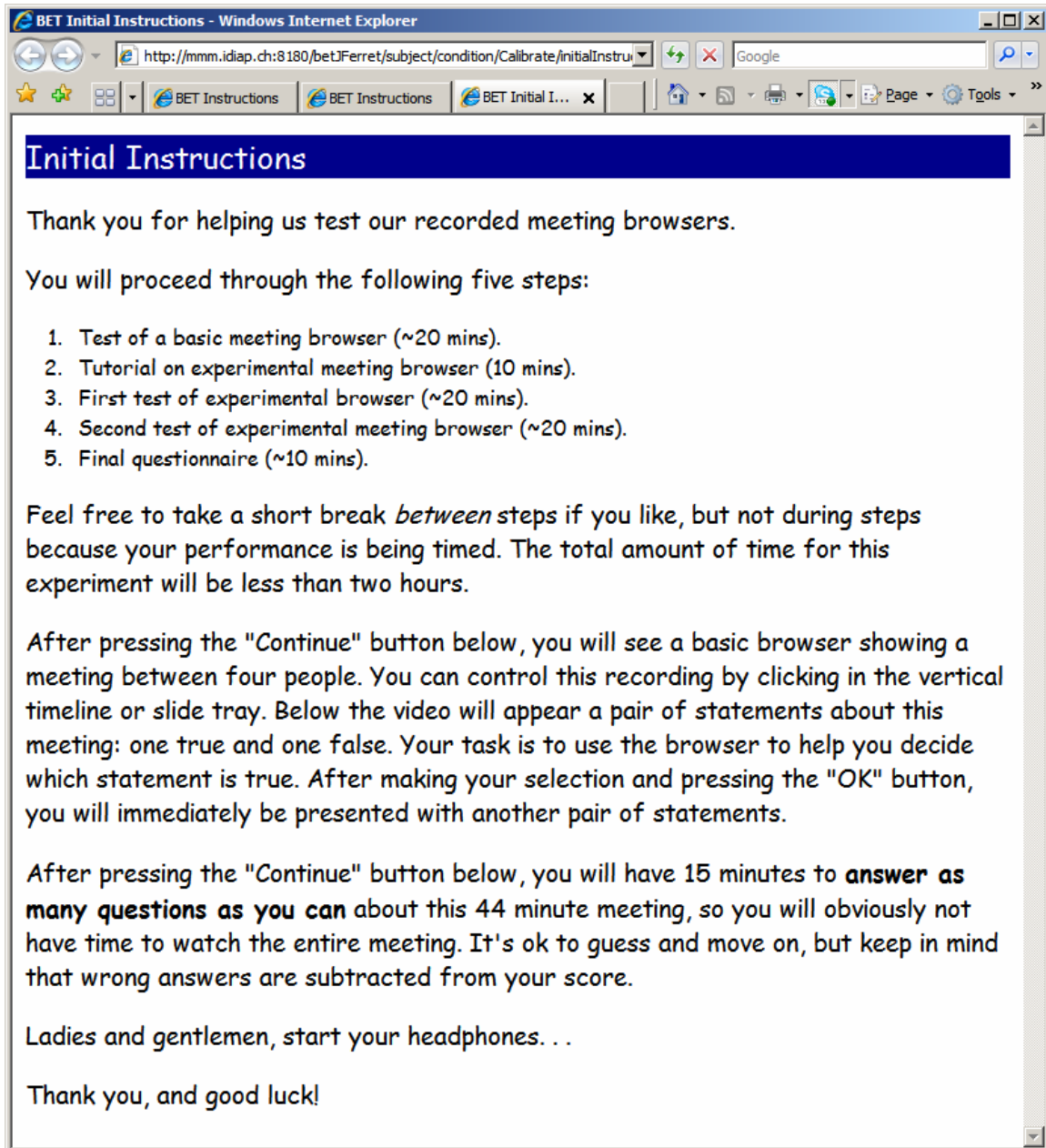
#367 Group: lifeTime Status: A Median adjusted importance: 1
Time: 0:08:00, Scope: Here **Importance: ★★★★★**

True statement:	The product is expected to last over several hundred years.
False statement:	The product is expected to last more than 5 but less than 15 years.

#446 Group: customizable Status: A Median adjusted importance: 0
Time: 0:06:25, Scope: Here **Importance: ★★★★★**

True statement:	Christine suggested that people/customers might want to submit their own design via the internet as custom orders.
False statement:	Christine suggested that people/customers would not be interested in custom design and prefer off-the-shelf products.

Appendix 3: Subject Instruction pages



The screenshot shows a Windows Internet Explorer browser window with the title 'BET Initial Instructions - Windows Internet Explorer'. The address bar contains the URL 'http://mmm.idiap.ch:8180/betJFerret/subject/condition/Calibrate/initialInstru'. The page content is as follows:

Initial Instructions

Thank you for helping us test our recorded meeting browsers.

You will proceed through the following five steps:

1. Test of a basic meeting browser (~20 mins).
2. Tutorial on experimental meeting browser (10 mins).
3. First test of experimental browser (~20 mins).
4. Second test of experimental meeting browser (~20 mins).
5. Final questionnaire (~10 mins).

Feel free to take a short break *between* steps if you like, but not during steps because your performance is being timed. The total amount of time for this experiment will be less than two hours.

After pressing the "Continue" button below, you will see a basic browser showing a meeting between four people. You can control this recording by clicking in the vertical timeline or slide tray. Below the video will appear a pair of statements about this meeting: one true and one false. Your task is to use the browser to help you decide which statement is true. After making your selection and pressing the "OK" button, you will immediately be presented with another pair of statements.

After pressing the "Continue" button below, you will have 15 minutes to **answer as many questions as you can** about this 44 minute meeting, so you will obviously not have time to watch the entire meeting. It's ok to guess and move on, but keep in mind that wrong answers are subtracted from your score.

Ladies and gentlemen, start your headphones. . .

Thank you, and good luck!

Appendix 4: BET Database Schema

The *observations* table is a list of all observations, whether used or not, redundant or not, important or not. It contains several fields:

id:	unique identifier for this observation
meeting:	name of the meeting (IB4010/IS1008c/... etc)
observer:	email of the observer
observationTime:	real time when the observation was made (milliseconds since 1970).
mediaTime:	time of the media player when the observation was made, in milliseconds.
scope:	extent of the observation, as given by the observer:

Here/ Around/Throughout.

importance:	importance of the observation, as given by the observer on 5 point scale.
trueStatement:	statement after minor editing (by us, collectively).
falseStatement:	statement after minor editing (by us, collectively).
bunch:	logical group, as determined by us (collectively), or empty if a singleton.
reject:	coded status, A=accepted, r=redundant, B=... other reasons for rejection.
originalTrueStatement:	statement exactly as entered by the observer.
originalFalseStatement:	statement exactly as entered by the observer.
editAuthorReason:	reason for editing, if any.

The *questions* table is a list of the questions put to subjects. It is simply a ranking of the acceptable observations:

rank:	order to ask the questions.
observation:	id of the observation, from observations table.
meeting:	name of the meeting (repeated for simplicity).
size:	size of the group from which this question came.
medianImportance:	median of the importance field of the group.
meanImportance:	mean of the importance field of the group.

Technically, we cannot take a mean of the importance, as it is a category not a value. We used a linear scale, so each successive category was 1 greater than the previous. However, the final ranking was not overly sensitive to this, so we felt it was an acceptable approximation