ATGENTIVE

IST-4-027529-STP

DELIVERABLE D4.4

Status: PUBLIC

VERSION: 1.0 (Final)

Final Evaluation Report

OXFORD BROOKES UNIVERSITY

Oxford, UK, December 5th, 2007

Paul Rudman and Mary Zajicek

The following people have contributed to this deliverable:

AUP Damien Clauzel Joona Laukkanen Claudia Roda Ayshe Ulgen	**Cantoche** Laurent Ach Benoit Morel	**CELN** Jaroslav Cech Ivana Malá Barbora Parrakova	**INSEAD** Thierry Nabeth (*Coordinator)* Nicolas Maisonneuve Albert A. Angehrn
OBU Paul Rudman *(Author)* Mary Zajicek (*Author)* Antonella De Angeli	**Ontdeknet** David Kingma Inge Molenaar Koen Molenaar Maurice Vereecken	**UTA** Daniel Koskinen Kari-Jouko Raiha Harri Siirtola Veikko Surakka Outi Tuisku Toni Vanhala Kimmo Vuorinen	**STC** Hasse Karlsson

Acknowledgements

We would like to thank the teachers from participating schools in the Czech Republic for their enthusiasm and hard work, which was a crucial part of the success of AtGentSchool.

We would like to thank the participants of AtGentive – students from schools in the Czech Republic and business users from around Europe and beyond – for their time and assistance.

Executive Summary

This document describes the summative and strategic evaluations for the AtGentive project. The project consists of the development of attention-aware versions of the Ontdeknet eLearning platform for children aged 8-15 (to produce AtGentSchool) and INSEAD's collaborative learning platform for adults (to produce AtGentNet).

Summative evaluation began with usability testing of the software, using expert-conducted heuristic evaluations, to ensure a suitable level of usability. The purpose of using heuristics as a software evaluation tool is to validate the software as being to appropriate usability standards separate to its effects on the users. The primary phase of summative evaluation comprises an extended period of direct user testing. AtGentSchool was tested over a six week period by classes in five schools in the Czech Republic (one or two classes per school). The AtGentSchool platform creates virtual learning relationships between subject experts and students, and provides guidance to support individuals to learn together based upon common interests.

AtGentNet was tested over a five month period by business people, across Europe and beyond, enrolled in a business-related distance learning course organised by the Swedish Trade Council (STC). Learners attend seminars and meet with local tutors who provide expert advice and assistance to help learners set and achieve practical goals. The process is supported by the AtGentNet platform which provides both access to information and a social forum.

The criteria for success were based upon three scenarios, implemented for each system (AtGentSchool and AtGentNet). These scenarios relate to the guidance of learning, the (re)direction of attention and user-notification. Research questions were generated from these scenarios, leading to five key indicators: Attention, Performance, Satisfaction, Learning and Collaboration. These indicators are used as part of the summative evaluation to assess the benefits gained by incorporating attentive agents.

In terms of overall success, the data indicate that use of an animated agent for children can successfully promote Performance, Satisfaction and Collaboration, with no apparent detriment to other factors. The use of scenarios as design elements proved very effective in designing for children. For adults, careful and subtle perceptual enhancements appear to be a better approach than animated agents. The data here indicate that performance and collaboration may be enhanced, but only where the motivation pre-exists. Perceptual-based attention support does not engender motivation, compared to the motivation provided to children by the animated agent in AtGentSchool

Strategic evaluation examines the real contribution of AtGentive in the outside world. The primary project outputs are identified, along with their strengths and weaknesses. Overall, the project approach of attacking the problem from several perspectives corresponding to the strengths and expertise of individual participants has proved very fruitful. AtGentive has met its principle objectives to successfully design and run two pilot studies which advance the support and understanding of attention with regard to educational software. While the support is situated in particular applications, knowledge gained may have implications in a wider context. This work has contributed across several disciplines in the research community – teaching and learning, collaborative systems, human- computer interface, and intelligent agents to mention but a few – and offers a basis for continuing research in this new and expanding area.

Contents

1. Introduction

1.1 General introduction to AtGentive

Attention represents one of the key factors of learning performance. The most effective learners are not necessary the most intelligent or the brightest ones, but those who are able to (1) organise efficiently their time; (2) sustain concentrating on their key activities and to complete them, and (3) have the psychological strength to mobilise all their energy for the last miles that will really make a difference.

This situation is aggravated in an online setting, where learners are left on their own, have fewer points of reference to situate themselves, do not receive any direct pressure from a tutor or from their peers, and can more easily procrastinate or engage in learning activities that are very ineffective.

The objective of this project is to investigate the use of artificial agents for supporting the management of the attention of young or adult learners in the context of individual and collaborative learning environments. This project comprises the modification of existing learning-support software to incorporate attention-enhancing features identified within AtGentive's conceptual framework (see deliverable D1.3 – "AtGentive conceptual framework and application scenarios") and found to be desirable by the formative evaluation. Such features range from relatively simple enhancements to facilitate perception of more relevant information, to direct intervention with the user by embodied agents. Overall, the aim has been to enhance the learners' effectiveness by directing their attention in more appropriate directions. This was approached in three broad ways: implicitly, by direct intervention, and by interventions to proactively coach the learners in the management of their own attention (assessment, guidance, stimulation, etc.).

Interventions are controlled by agents that profile the (short or long term) state of the attention of the learners by observing their actions, to assess, to analyse and to reason on these states of attention and to intervene as suggested by the conceptual framework. Where agents need to communicate directly with the learners, they may do so simply by changing the information on the screen, adjusting what is available to the user. More directly, they are able to appear as cartoon-style characters, embedded in the application and its interface. Thus, embodied agents are an important interface element for AtGentive.

Interventions have been designed and tested as part of two different learning infrastructures / contexts. The initial focus for one context, AtGentSchool, is selected schools in the Czech Republic. It supports students aged between 8 and 15 years of age collaborating with subject experts. AtGentSchool is built on the Ontdeknet eLearning platform, created by the Dutch company Ontdeknet. This platform is an electronic learning environment that makes knowledge and skills in society accessible to educational institutions in general and individual students in particular. Virtual learning relationships between subject experts and students are established in this virtual learning environment. The Ontdeknet environment provides guidance to support individuals to learn together based upon common interests. This platform originally used an embodied agent ("Onty", a cartoon fish) to guide learners around the learning environment. The software was adapted to incorporate the AtGentive interventions, and the existing embodied agent changed in accordance with the results of formative evaluation to "Honza" (a cartoon boy) (see Figure 1 and deliverable D4.2 – "Result of the Formative Evaluation".

The second context, AtGentNet, focuses initially on use by adult learners enrolled in business-related courses organised by the Swedish Trade Council (STC). It will support

adult learners, located individually but collaborating using an internet-based system. AtGentNet is built on the ICDT virtual community platform, created by the INSEAD Business School's Centre for Advanced Learning Technologies. This platform is a web-based virtual environment aimed at supporting distributed groups and communities. In terms of functionality, the ICDT Platform integrates features aimed at providing efficient Information, Communication, Distribution and Transaction channels used by the community of users. This platform was adapted to incorporate AtGentive interventions, along with an embodied agent "AtGentiGirl" (a cartoon woman), based upon the results of formative evaluation (see Figure 1 and deliverable D4.2 – "Result of the Formative Evaluation").

Figure 1 – Agents AtGentiGirl (AtGentNet) and Honza (AtGentSchool)

1.2 Introduction to this document

This document comprises the final report of the evaluation component of AtGentive. Two main areas of evaluation are reported. Summative evaluation details the results of two pilot studies – AtGentNet (an adult collaborative learning platform) and AtGentSchool (an interactive guided learning system for children). This summative evaluation looks at the effects of modifying existing software to make it "attention-aware" and respond accordingly.

Strategic evaluation examines the wider effects of AtGentive - its potential for reuse, adaptation and influence of future attention-related research and systems' development.

1.3 Introduction to the Conceptual Framework

The starting point for design of AtGentive's attention-based enhancements is the Conceptual Framework. The framework breaks attention-supporting interventions into four main forms (see Figure 2):

- Perceptual – "Bottom-up" processes (e.g. a flashing image attracts attention)
- Deliberative – "Top-down" processes (e.g. the user may decide to check their email every hour)
- Operational – Managing interruptions (e.g. the user may disconnect a telephone)
- Meta-cognitive – Self-support (e.g. the user may learn which emails are "junk" and can be ignored)

Interventions may be seen as relating to three types of problem:

- Procedural interventions for Regulative problems
- Content interventions for Cognitive problems
- Process interventions for Meta-cognitive problems

A key aspect of the Conceptual Framework is that such interventions may be driven by the monitoring of events. Such events will be discerned from as wide a variety of sources as possible. The main categories of event are as follows:

- Application events (e.g. The user has started a task in the application, new information is available for the user)
- User events (e.g. The user indicates that (s)he wants to be notified about certain events, or that a task should have a high priority)
- Tracking events (e.g. The user has been idle for some time, a resource has been used by other users)

The relationship between these different elements of the Conceptual Framework is illustrated in Figure 2. Support is broadly categorised as that of the user's immediate focus of attention (what they are concentrating on at this moment) and their broader voluntary attentional choices (what they may choose to attend to next).

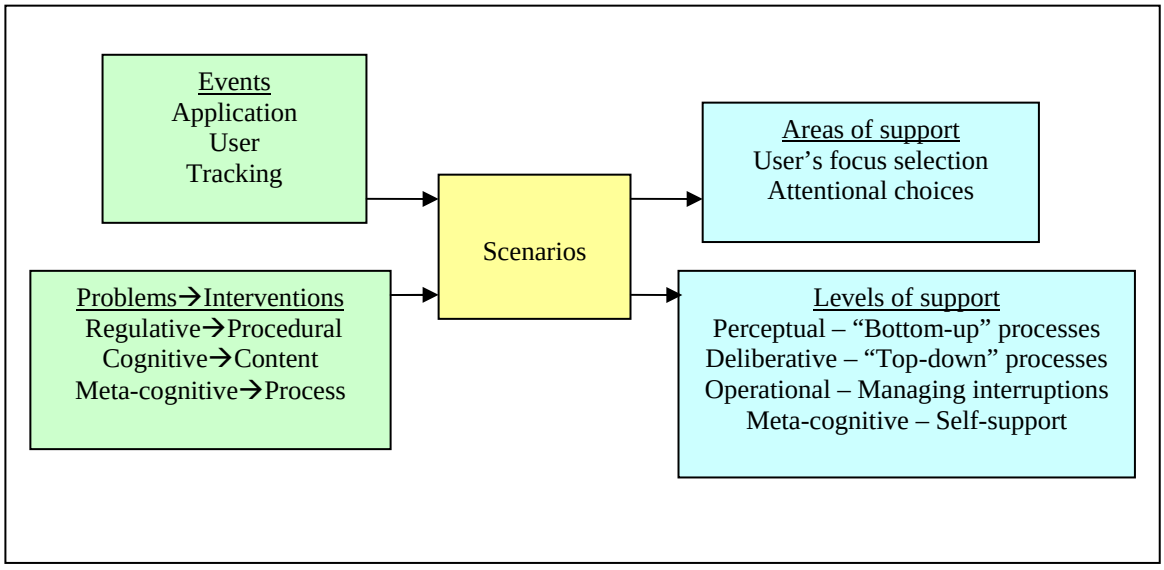


Figure 2 - Relationship between areas of the Conceptual Framework

It is hypothesised that these events, which may be captured and analysed by AtGentive agents, will reveal the interaction paths between the essential elements of User(s), Application (Ontdeknet, ICDT), Environment (external events) and the AtGentive agent(s). Further, this observation will reveal the user(s)' attentional choices, preferences, and possible future foci. It is this analysis that results in the agent's interventions. These interactions are exemplified by a number of user scenarios (see deliverable D1.3 – "AtGentive conceptual framework and application scenarios" – for further details). A number of these scenarios are implemented within AtGentive, as described in Section 2.

1.4 Introduction to Platforms

1.4.1 AtGentSchool

AtGentSchool is an e-learning system that allows students to work on assignments in collaboration with experts outside the school. The system revolves around the “Project screen” (see Figure 3). The project was designed in collaboration with the Czech teachers, especially for the AtGentSchool pilot. The student’s overall assignment is to compare two countries – New Zealand to Czech Republic – before deciding in which country they would prefer to live. In order to achieve this, the students have eight possible learning activities (tasks), which they work on in order – in collaboration with the expert – to acquire the knowledge necessary to make their final decision.

The students have a significant degree of control in their selection of learning goals and learning activities. Previous tests of the Ontdeknet software have shown that students with good regulation skills learn successfully with experts in this environment. However, many students will benefit from extra support in their collaboration with the expert. The project screen provides a script to support this collaboration.

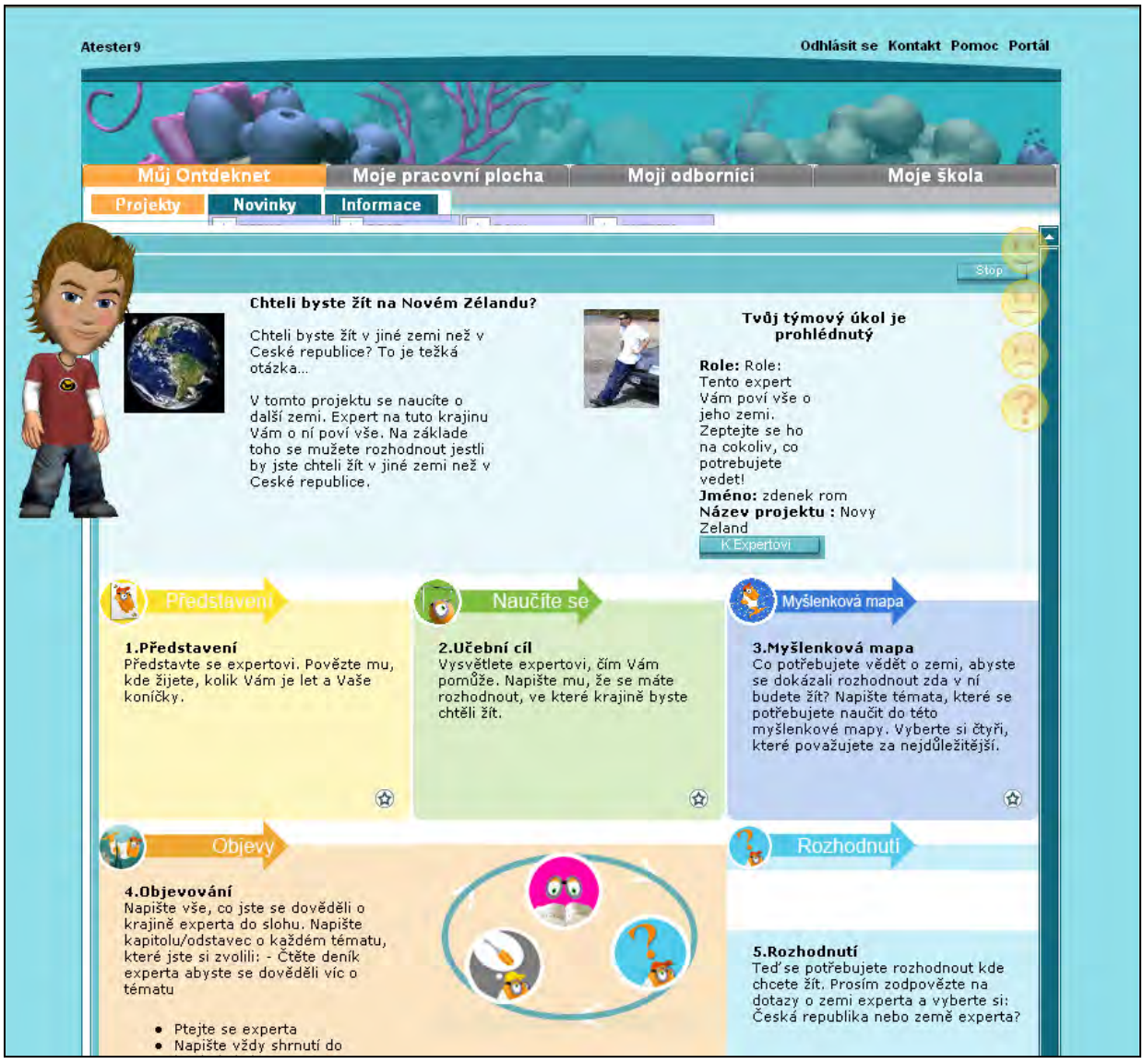


Figure 3 - Screenshot of AtGentSchool home page, showing the animated agent

The project-screen displays all learning activities a student needs to perform to successfully finish the main assignment by collaboration with the expert.

The project screen describes the stages as follows (see deliverable D3.2 – "The Prototype" for further details):

Main assignment

"Would you like to live in another country then Czech Republic? That is a difficult question! In this project you will learn about another country. An expert from that country will tell you everything about his country. Then you can decide if you want to live in Czech Republic or in his country!"

1. Introduce

"Introduce yourself to the expert.

Tell where you live, your age and your hobbies."

2. Learning goal

"Explain your expert what his is going to help you with.

Tell him that you have to decide which country you would like to live in."

3. Mind Map

"What do you need to know about the other country to decide if you like to live there?

Write the topics you need to learn more about in the mind map. Select the four topics that you find most important."

4. Discover

"Write everything you discovered about the country of the expert in a paper.

Make a chapter of every topic you selected:

- Read the dairy of the experts to learn more about the topic
- Pose questions to the expert
- Write a summary in your chapter"

5. The decision

"Now you need to decide where you want to live

Please answer the questions about the country of the expert and make your choice: Czech Republic or the country of the expert!"

1.4.2 AtGentNet

AtGentNet is a platform aimed at supporting the online interaction of groups of people engaged in an offline training programme in which they can only meet physically during short periods of time (a few days every several weeks). In particular, AtGentNet aims to help this group stay "in touch" while they are physically dispersed, and to contribute to helping them know more about each other, stimulate their interaction and knowledge exchange about the programme, and keep them motivated (see deliverable D3.2 – "The Prototype" for further details).

Figure 4 shows a typical home page, comprising a number of “portlets”. Many of these window-style screen areas may be relocated within the main window as the user prefers. The portlets are listed below. (These portlets are described in more detail in Section 2.2.4.)

Note that most agent “interventions” are shown via a portlet, not using the embodied agent. (These agent-related aspects of the platform – only available to users in the Experimental group – are annotated in the following list with a *).

- Chat – permanently-visible text discussion area for all users
- Control – shortcuts to general frequently-used pages (e.g. “News”)
- Personal – shortcuts to frequently-used pages relating to the user (e.g. “My messages”)
- News, highlight – features a recent news item (defined by the system administrator)
- News, latest – lists the most recent news items
- Personal – lists recent unread entries
- Last visitors – lists visitors over the last six hours
- Knowledge – lists the recent most popular postings
- *Watch – Shows an overview of items (postings / people) that the user has previously selected to keep track of
- *Agent On / Off – allows the user to request a number of general “interventions”, such as help in using the platform. These are delivered by the embodied agent
- *Agent, Pending interventions – the main initial communication point between agent (not embodied agent) and user, detailing “interventions” (i.e. suggestions, such as specific postings the user may benefit from reading)

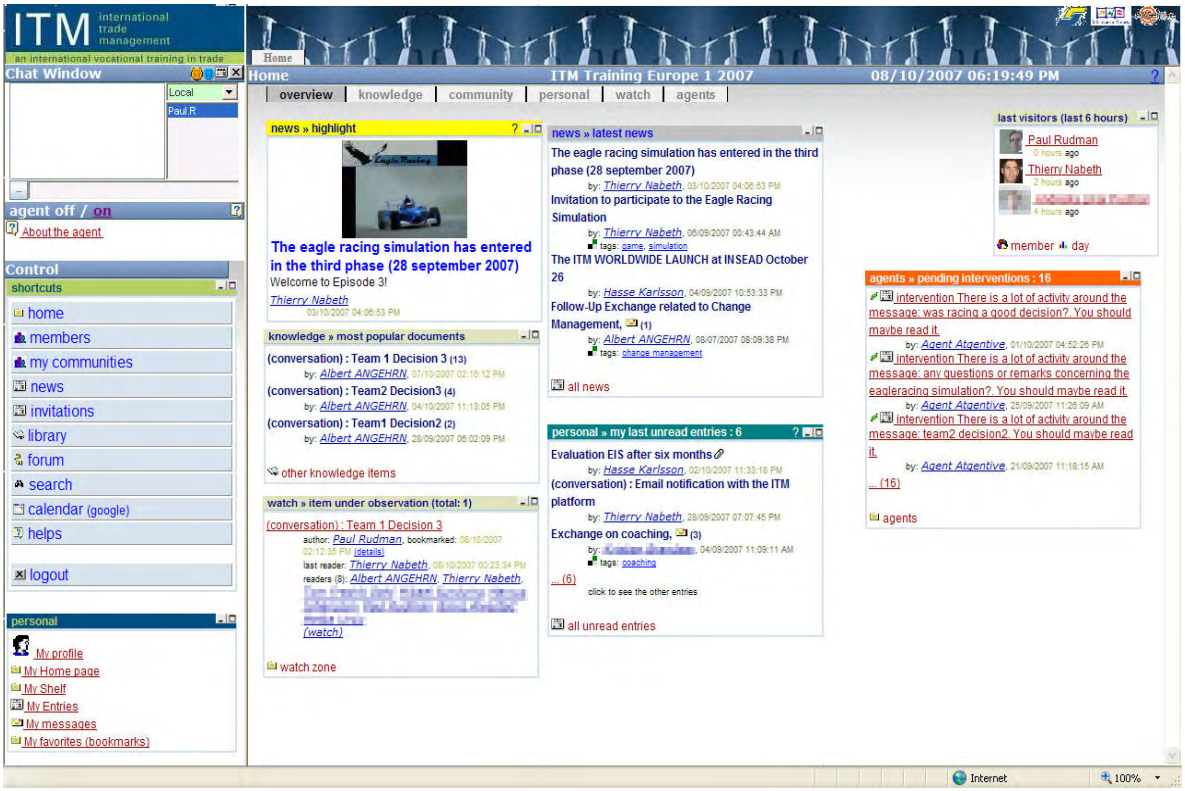


Figure 4 - Screenshot of AtGentNet, showing the home page portlets

1.5 Introduction to Summative Evaluation

The objective of the summative evaluation is to evaluate the success of the attention-related modifications to the collaborative e-learning platforms, as implemented in the two pilot systems. Evaluation methodologies employed are wide ranging, employing appropriate groups of users and longer, more sustained and realistic user tasks than those used within the earlier formative evaluation. Control groups who are exposed to the same material without the benefit of attentive agents also play a vital role in evaluation.

The summative evaluation described in this document evaluates the AtGentive interventions, the overarching aim of which is to assist collaborative learning through the direction of attention. This phase of AtGentive began at month 10 (September 2006) with the actual pilots taking place between months 17 and 23 (April-October 2007).

It is important to note that extensive formative evaluation took place in the early part of the project (see deliverable D4.2 – “Result of the Formative Evaluation”). This both defined the development of the software and included initial tests by the developers, both internally and with representative users, to ensure the software was of suitable quality for external evaluation. In particular, AtGentSchool conducted a number of “Test-runs” with 10-11 year-olds, while the project partners gave extensive feedback for AtGentNet (see deliverable D3.2 – “The Prototype”).

The current section (1.5) of this document detail the approach and methodology employed for the summative evaluation. Section 2 describes the evolution process from original user Scenarios to user pilots, including the timetable followed. Section 3 describes the tools used for the summative evaluation, including a description of the five Key Indicators, the measurement of which provides the means to verify the efficacy of AtGentive in achieving its stated aims.. Section 4 then describes and analyses the results obtained and their meaning in terms of the Key Indicators.

1.5.1 Approach

For the summative evaluation, two concepts are considered, usability and usefulness. Usability measures the ease with which the user is able to pursue their activities using the system, while usefulness is the beneficial effect the system has on what the user achieves. So for example, being able to obtain today’s date is useful, while being able to see today’s date just by looking at the bottom-right of the screen is usable.

In practice, the two concepts are usually linked. The more usable a system is, the more useful it is likely to be. For example, if the system proffers advice to the user, but that advice requires a difficult combination of button presses to access, the advice itself may be less acceptable (and therefore less helpful) because of the context of frustration its accessing generated. Less acceptable advice may reduce the likelihood of its being acted upon, and thence the usefulness of the system as a whole. Therefore, while usability and usefulness are considered separately, the same measures may contribute to the evaluation of both concepts. (See Section 3.3 for further discussion.)

The overall approach taken has been a combination of heuristic evaluation and field observation and measurement. Heuristic evaluation generates rapid results which may be used immediately to improve the product, with 3 evaluators being expected to find over 60% of usability problems (Nielsen & Mack, 1994). Laboratory and field experiments, on the other hand, take much longer to set up and produce results that often require interpretation, but can access the more complex and hidden usability problems. The two approaches work well together:

> 'By identifying obvious or clearcut usability problems, heuristic evaluation "harvests the low-hanging fruit and provides a focus for laboratory testing" ' (Kantner & Rosenbaum, 1997, pg. 2).

Heuristic evaluation took place as soon as the software became available. Where necessary, feedback from this evaluation was used to improve the software prior to the next stage. Field tests then examined the software in a naturalistic environment. This user pilot provided a good insight into the main benefits, advantages, problems and difficulties offered by the software. As a first pilot, these insights will be used to improve the software further and ensure that any future comprehensive large-scale testing of the software focuses on the main benefits and problems the software exhibits.

1.5.2 Methodology

The methodology employed for the user evaluation pilot has been to develop a set of key indicators against which the performance of the software may be measured. These key indicators are based upon the six scenarios used as a basis for the systems' design (three per system). The key indicators were then expanded to define the means by which they may be measured for each platform.

1.5.3 Heuristics

The use of heuristics is a well established methodology in the area of usability evaluation of computer software. The procedure (Nielsen, 1993) consists of several evaluators who separately examine the software for problems, using predefined criteria (heuristics). It offers the ability to identify a large proportion of usability problems quickly and at low cost. Further, the use of "Extreme Programming" as a development methodology within AtGentive requires an evaluation methodology capable of offering rapid feedback as to the quality, usability and relevance of each released software iteration. Heuristic evaluation offers this rapid feedback, especially at the early stages of development.

By the use of appropriate heuristics, described later, this methodology has been adapted for use with the summative evaluation of the AtGentive system. Heuristic evaluation has been used as a first-line test of the AtGentNet and AtGentSchool systems. By identifying problems at this stage, the effectiveness of the user evaluations was maximised. Most importantly, the use of heuristic evaluation allowed the separation of primary usability issues from those usability issues which, while not problematic as such, could nonetheless be relevant to other aspects of the software's effectiveness and usefulness.

1.5.4 Evaluation Framework

The design for the summative evaluation has incorporated an adaptation of an evaluation framework developed for the evaluation of Computer Aided Learning packages. The CIAO! Framework (Jones et al., 1999) is based on the interaction between Context, Interaction, Attitudes and Outcomes. Context comprises the original aims and goals of the system designers. Interaction is the way in which the software is used – an interplay between usability and activity. Attitudes and Outcomes are the results of interaction – effects on the students (such as frustration or being supported) and desirable outcomes (such as improvements in learning or collaboration). These concepts are described in more detail in section 3.3.

1.5.5 User evaluation

We carried out a rigorous summative evaluation using both quantitative and qualitative measures. For example a quantitative measure of learning could be a student test, while a qualitative measure of learning could be a semi-structured interview with an adult learner.

It has been important not to place too high a load on the participants. This was taken into account in designing the evaluation tools themselves. For example, the adult learner questionnaires were kept to a length deemed acceptable by a representative of the participants, while for the sample schools, direct evaluation of the students (such as questionnaires) took into account the teachers' view of how many questions were acceptable and when and how often a questionnaire could be administered.

The evaluation process was modularised (see Figure 5) in order to address specific aspects of the AtGentive software and in order to maximise the data collected and ensure smooth-running of the evaluation process – given the involvement of multiple countries and researchers with a variety of backgrounds and experience.

The first evaluation phase took place during user training and familiarisation. As this took place before completion of the software, the software did not have the AtGentive interventions enabled at this time. This phase did, however, yield valuable results and these results were used as additional formative evaluation, since the systems were not completed at that time. Also, of great importance was the use of this evaluation phase as a trainer for evaluators to give them the opportunity to appreciate the full implications of the process and what it was hoped to achieve, as well as to ensure that they were operating to the necessary standard and that they were properly supported in their learning. For AtGentSchool, this phase was undertaken with the teachers; for AtGentNet, this phase took place with customers of STC and with their representatives at STC.

The second evaluation phase took place with the final software. A formal heuristic evaluation was used to ensure that the software itself was working to an acceptable standard for the final test.

The third and final evaluation phase of AtGentive was the main pilot evaluation. For AtGentSchool, the six weeks of system use was broken into two parts, each with their own evaluations. This allowed for comparison between initial use and more extended use of the system. For AtGentNet, the trial evaluated both general use over a five month period and a "business simulation exercise", where each week for three weeks participants watched an on-line video describing a business problem and then collaborated using the AtGentNet platform to decide upon the advice they would give in this situation (see Section 2.2.7).

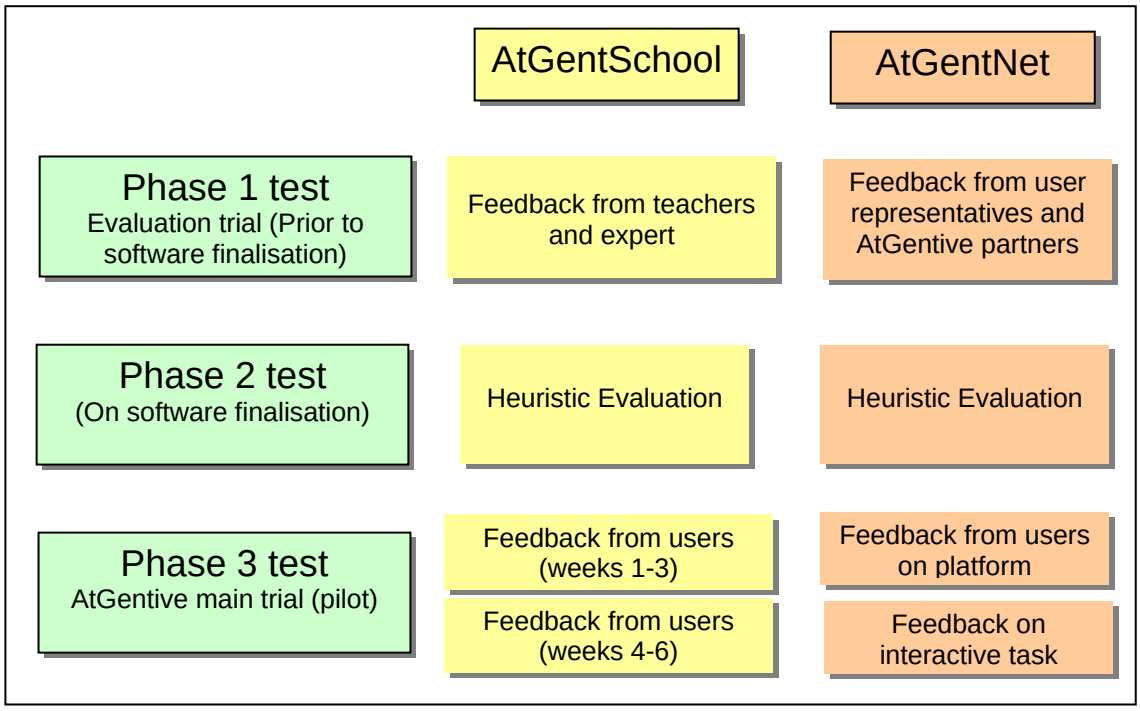


Figure 5 - Diagram showing the three summative evaluation phases

1.5.6 Ethical Issues

Both AtGentSchool and AtGentNet accrue information about the systems' users. The information collected is kept in a secure and confidential manner. It will not be used in a way that could identify individuals by those outside the AtGentive research partners without those individuals' explicit permission.

It should be noted that neither system intends to build a comprehensive user profile, in terms of maximising the acquisition of information about a user and creating a discrete profile that may be accessed, shared and used in its own right. Profiling information generated remains within the software for use only within that context. In addition, within AtGentSchool each computer (running AtGentSchool) was be used by two students simultaneously. It was therefore not possible to accrue detailed information on individual children.

1.5.7 Additional experiments

In addition to the piloting of AtGentSchool and AtGentNet, experiments were conducted of a more general nature into specific issues considered to be of critical importance in further development of the AtGentive concept.

- Experiments to investigate learners' likely compliance with agent instructions / advice
- Experiments to investigate the effects of embodied agents' gestures
- Psychophysiological tests using eye-tracking equipment to establish the background for future advanced versions of AtGentive that could incorporate physiological input.

1.6 Introduction to Strategic Evaluation

The purpose of the strategic evaluation is to document and evaluate the potential of the AtGentive outputs to make ongoing contributions in the outside world after completion of the AtGentive project. This is irrespective of specific post project activities planned (see deliverable D6.4 – “Assessment and Consolidation report in the perspective of further exploitation and final exploitation plan”).

Three stages of strategic evaluation have been identified:

- Project Objectives – the project’s goals and desired objectives, with reference to the original Description of Work
- Project Outputs – the tangible results generated by the project – concepts, knowledge and artefacts – as identified by the individual responsible project partner(s)
- Project Impacts – potential value and effects of the project outputs. This will be assessed using key indicators, as described in section 6.3

Section 6 of this document describes the approach, methodology and results of the strategic evaluation.

2. Scenarios to Pilot study

Each software platform (AtGentSchool and AtGentNet) is based upon existing software with the implementation of three scenarios of use, with reference to the conceptual Framework (see deliverable 1.3) and user-based formative evaluation. The scenarios to be implemented for each platform are listed below, along with a description of their implementation.

2.1 AtGentSchool

Scenarios for AtGentSchool were created as a result of the formative evaluation – specifically the "Wizard of Oz" study where a suitably knowledgeable "Wizard" selects the interventions manually (see deliverable D4.2). The scenarios for AtGentSchool were developed in relation to the scenarios described in the Conceptual Framework document (see deliverables D1.3 – "AtGentive conceptual framework and application scenarios", and D2.2 for further details) but represent the circumstances of use within AtGentSchool.

2.1.1 Scenario S1 – Guidance of Learning

This scenario is dealing with the effective support of the learning process of the user. A predetermined intervention scenario will guide the selection of the intervention. Based on the attentional state, the context of the user and the user model, it will be determined if an intervention is to be provided to the user:

2.1.2 Scenario S2 – Idle User

When the user is judged to be idle, the appropriate intervention is selected from the intervention model. The primary challenge for this scenario is to select the most appropriate intervention given the preceding user events:

2.1.3 Scenario S3 – External events

This scenario deals with external events that occur within the AtGentSchool application. The right moment has to be determined to communicate the external event to the learners. The focus lays on establishing *if* and *when* an intervention should be communicated to the user. Based on the attentional state of the learner(s) and the current context of the learner(s), the *right intervention moment* is assessed:

2.1.4 Implementation

In order to implement the three scenarios described above, a theoretical approach was established, using "scaffolding" as the primary concept. Scaffolding is a dynamic process, whereby the learner is provided with continual assistance specific to the next step in their learning process. It has been likened to the continual extension of scaffolding as a building progresses (Bruner, 1983), but the emphasis is on a continual adjustment of assistance, rather than a fixed structure as the scaffolding analogy might otherwise imply.

In practice, learners are encouraged to carry out the parts of tasks that are within their ability, while the "teacher" (in this case, AtGentSchool) "fills in" or "scaffolds" the rest. The scaffolding involves recruiting the learner's interest, reducing their choices, maintaining their goal orientation, highlighting critical aspects of the task, controlling their frustration, and demonstrating activity paths to them (Wood, Bruner, & Ross, 1976).

The scaffolding process within AtGentSchool is dynamic, and provides four main types of support: meta-cognitive, cognitive, behavioural and motivational. Examples of these four support types are given below:

- meta-cognitive – i.e. help with self-organisation
- cognitive – i.e. direct help with the current task
- motivational – i.e. interventions to motivate the learner
- behavioural – i.e. interventions to change immediate behaviour

The interventions include those used to implement the three scenarios described above, but go further towards a general support for scaffolding, as now described:

Meta-cognitive interventions

Meta-cognitive interventions support learners in their understanding of the meta-cognitive activities that can be performed during the learning process. Learners with low self regulation skills do not perform meta-cognitive activities during their learning process. The agent's interventions are dynamically provided to the learner at an appropriate moment. This allows the students to become aware of the meta-cognitive activities they could use to help them regulate their learning. The following meta-cognitive interventions were implemented in the AtGentSchool pilot (see Figure 6). These interventions also address **Scenario S1** – Guidance of Learning, **Scenario S2** – Idle User, and **Scenario S3** – External events.

Orientation interventions

Experts in specific fields are known to spend more time on orientation to a task than novices. A better orientation on the task allows for a better comprehension of the task elements. The agent's introduction of the task can lead to better comprehension of the task, which can influence the time and performance of that task. These interventions are provided to the learners in the project screen overview just before the learner is about to commence on the task. For example, the intervention to offer orientation for the task "learning goal" is:

> "Your expert would like to know what your learning goal is, could you tell him? Please click here to write your learning goal."

Explanation interventions

In the explanation of the learning task the agent models the task execution to the learner. This is expected to help students in the accomplishing the task effectively. These interventions are provided to the learner right after the task page is opened. For example, the agent's explanation intervention for the activity "introduction" is:

> "Here you will introduce yourself, I will give an example: "My name is Honza, I live in Prague, I am 16 years old. My hobbies are skating and chatting. I have one older brother named Karl."

Monitoring interventions

The monitoring statements of the agent clearly indicate to the learner that the current task is finished and explain again what the system or the expert will do with the information that task has provided. The clear closure on the task should help the student to continue on the next task. The monitoring interventions are provided to the student immediately upon completion of a task. An example of a monitoring intervention is the one given after the completion of "filling in the learning goal":

"I'll directly go to your expert and explain him what you would like to learn."

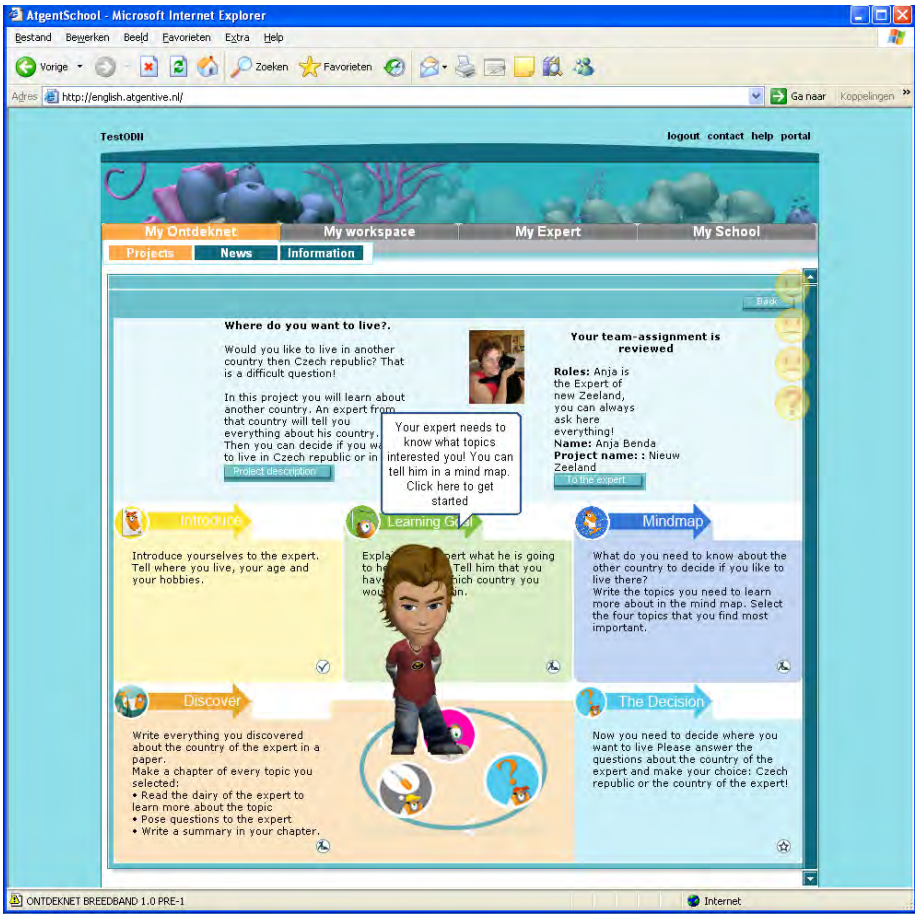


Figure 6 - Example of a meta-cognitive intervention within AtGentSchool

Cognitive interventions

Cognitive interventions support the student's learning process during the execution of a learning activity. Cognitive interventions provide the knowledge and skills necessary to perform the task (Garner, 1987). They support actions with respect to both the content and context of the learning task and are direct at the level that Nelson referred to as the object level. These interventions are specifically adjusted to the learning activity at hand. The triggers for cognitive interventions are:

- Idle user – as tracked by the ASKME module These interventions also address **Scenario S2** – Idle User
- User request – namely the student clicks on the "question mark" button. These interventions also address **Scenario S3** – External events

There are two types of cognitive interventions: cognitive support interventions and cognitive resources interventions. Cognitive support is directed at helping the current learning activity whereas a cognitive resource provides students with a link to a resource in the learning environment that can help them perform the task. For example, a cognitive support intervention for the activity "mind-map" (where student have to write down all topics that are related to the subject that they are studying) would be:

"What do you already know about the subject you are going to study?"

while a cognitive resource for the same learning activity would be:

"Need some ideas? You can read the introduction diary of the expert"

See Figure 7 for an example of a cognitive intervention in practice.

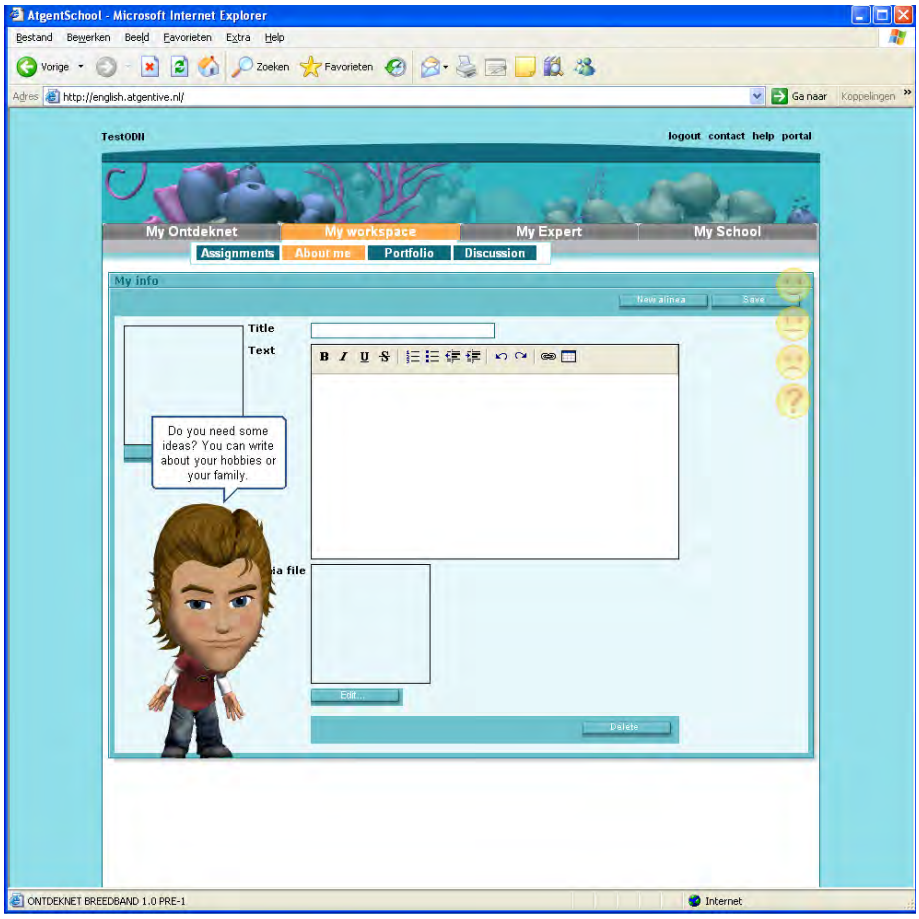


Figure 7 - Example of a cognitive intervention within AtGentSchool

Motivational interventions

Motivational interventions support the learner's motivation to work on the task and they are directed at increasing the motivation of the students. Motivation influences the activity of the learner to a large extent.

Motivational interventions are triggered by two events:

- Idle user – as tracked by the ASKME module. These interventions also address **Scenario S2** – Idle User
- Emotional indicators of the learner – i.e. the student clicks on one of the four "smiley" buttons ("happy", "sad", "neutral" or "confused")

Where the user has become idle in a task and there are no more cognitive interventions for this user the motivational interventions are shown. An example of a motivational intervention is:

> "You can do it! Just start writing"

When an emotional indicator is generated, the agent mirrors the state of the learner showing an animation and expression resembling the state indicated. Figure 8 shows the mirroring of a "sad" intervention.

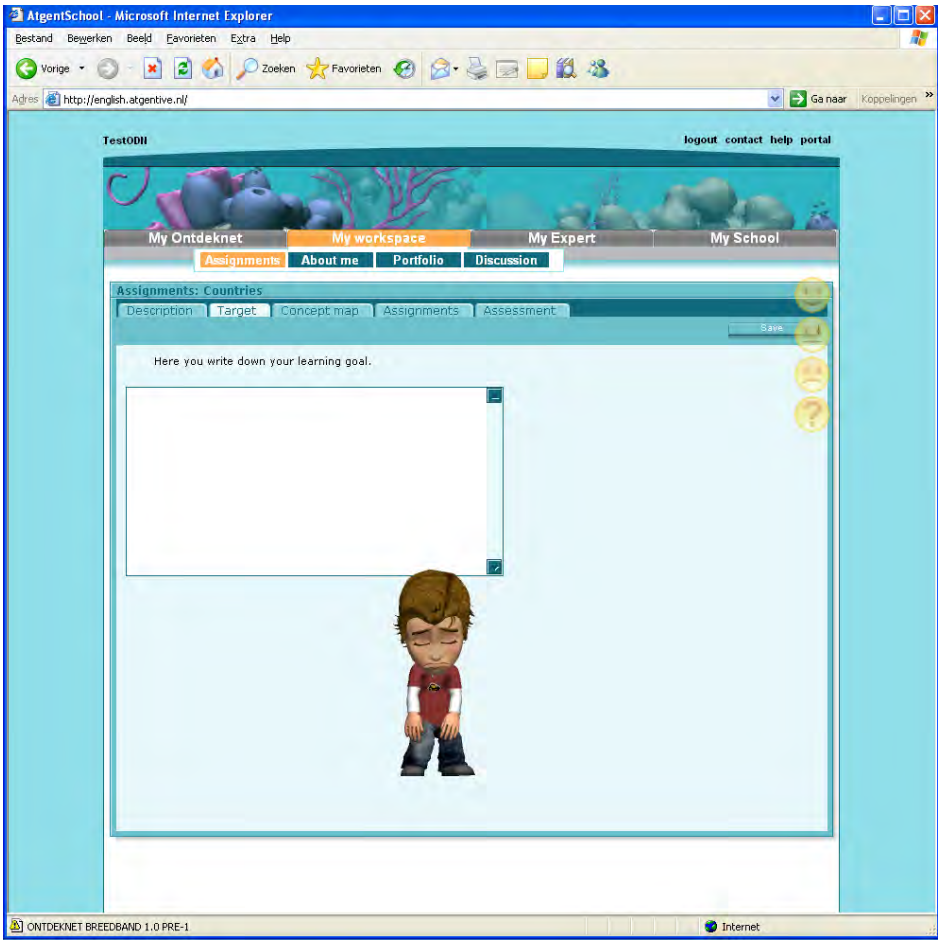


Figure 8 - Example of a motivational intervention within AtGentSchool. (Note the feedback buttons ("smileys"), far right)

Behavioural interventions

Behavioural interventions support the learner in working more effectively with the environment. These interventions, which target the learner's actions in the environment, are directed at supporting the learner in effectively moving between activities. There are two types of behavioural interventions:

- external events notifications
- navigational support

The external events are events that are caused by another user in the learning application which have relevance to this specific user. The system evaluates how important the information is in relation to the user's current task and activity, and will decide to notify or to postpone the notification for a later time. For example, when an answer is posted to a question on the forum, the agent would tell the current user by saying:

> "Your expert has answered your question"

Navigational support consists of simple navigational statements to direct the user to certain elements in the system. For example:

> "Click here to go back to the project screen"

2.1.5 The animated Agent

In AtGentSchool, the animated agent ("Honza" – see Figure 1) is used for two separate purposes. All users see Honza as present on the screen. When the student clicks on a "smiley" to give feedback to the system about their mood, Honza reciprocates by adopting a related body state, along with a short narrative.

For the experimental group, Honza also delivers the interventions described above by proactively making suggestions intended to assist the student at that moment, given their (assumed) current attention. Interventions relate to the above scenarios. For example, in the case of Scenario S3 – External events, if an email is received for the student the system will wait until it deems the student to be at a suitable break point before informing them of the new email.

2.1.6 Continual development

AtGentSchool underwent refinements at the pre-pilot stage. In particular, improvements were made to the agent's rules (see deliverable D2.2, section 2.4). These changes were based upon further analysis of the system by Ontdeknet, teachers' feedback during the "overview training" and understandings gained from the formative evaluations. However, once the pilot study began only essential changes were made to the software.

2.1.7 User Evaluation Pilot

The user pilot took place in three stages:

- Teacher training (pre- pilot)
- Pilot (part 1)
- Pilot (part 2)

The user pilot took place over a six week period beginning 3rd May 2007. Students typically used AtGentSchool for either six or seven 45 minutes sessions within this period. (School timetabling did not allow for exact weekly sessions, and included some classes that took a one week school trip during the pilot.) Students used computers in pairs (with a few exceptions of single or triple use), working together on the computer at the same time. Half the participants used the original AtGentSchool software, without the interventions, while half the participants used the new full version of AtGentSchool. Students were randomly assigned into groups (within the requirement of having the same number of student pairs per class in each group). The children themselves were not explicitly made aware of the difference in the two systems, or the split into groups.

2.1.8 Timetable of events

Below is shown the timings for evaluation activities related to AtGentSchool. Further details are shown in deliverable D5.1 – "Specification of the implementation of the pilots".

November 2006

Introductory meeting held in Prague on 9th November 2006 to introduce teachers to AtGentSchool developers and researchers and the AtGentSchool system

January – February 2007

"Overview training"— introduction to Ontdeknet / AtGentSchool for teachers

May 2007

Final training for teachers using Czech version of AtGentSchool in Prague on 2nd May 2007

User evaluation pilot began in five schools in and around Prague on 3rd May 2007

June 2007

User evaluation pilot ended on 14th June 2007

Teachers and Expert focus group in Prague on 27th June 2007. Details of the workshop are given in "5.4 – Report from the Workshop arranged to Analyse the Results of the Pilot"

August 2007

OBU and UTA completed a detailed heuristic evaluation of AtGentSchool, which was distributed within AtGentive on 6th August 2007. It was necessary to delay the start of this heuristic evaluation until an English version became available.

2.2 AtGentNet

Scenarios for AtGentNet were created as a result of the formative evaluation (see deliverables D4.2 and D4.3). In particular, they take into account the partners' use of the ICDT platform, discussions with a representative of the users (Hasse Karlsson from STC) and direct user feedback in the form of questionnaire study. This study is reported here, followed by the three scenarios.

2.2.1 Scenario N1 – (Initial) learning guidance

This scenario is intended to support self-directed learning. A new user is guided through the platform by being given tasks to work on, ensuring they have completed the tasks correctly. The guidance is in three steps: (1) create a user profile, (2) learn about the features of the platform and (3) learn how to supply information and communicate with other users. This process may take place over several logins to the platform (the system will remember where the user was).

Note that this is an exception to the usual user experience of the ICDT platform (and thus AtGentNet) where the user has free reign over his or her work and creates / selects his or her own tasks. It also contrasts with AtGentSchool where the user experience is one of being guided throughout by an embodied agent.

Note also that this is similar to the new-user introduction in Ontdeknet (and therefore AtGentSchool). However, the Ontdeknet system already contains this new-user guidance, so the guidance will appear both in the control and modified versions. With AtGentNet, the new-user guidance is implemented as a scenario, and will only appear in the modified version, allowing evaluation of its usefulness.

2.2.2 Scenario N2 – Notify tutor of user (in)activity

This scenario allows the tutor to encourage students who have not read important documents or messages.

When the tutor logs in, if (s)he has posted documents which (s)he has requested that students read or respond to, (s)he will be notified of individual students that have not read / responded.

2.2.3 Scenario N3 – Notification of events

The purpose of this scenario is to assist students to identify the most useful, relevant and urgent documents to attend to.

When the students logs in, (s)he is notified about documents based on their social-network, interests, etc. The notification criteria are controlled by the tutor.

2.2.4 Implementation

Unlike AtGentSchool, the agent is used only to deliver specific help information. Implementation of the scenarios in AtGentNet is based upon support for the learner's perception, via entries in a number of additional "portlets" – small window-style display areas within the main display area – that may be repositioned by the user (see Figure 9). In addition, AtGentNet incorporates changes in the software to enhance perception, as this is an attention-related factor identified by the Conceptual Framework. The additional portlets made available to AtGentNet users are:

- "Watch" portlet – allows the user to select specific postings for which changes / access details are displayed in the Watch portlet on the Home page
- The Watch tab – calls up an expansion of the Watch portlet, listing a greater number of watched items and allowing statistical details to be displayed via the "analyser agent". **Scenario N2** – Notify tutor of user (in)activity – is implemented here, in that the user (in this case a tutor) may select a posting to watch and then obtain a list of users who have read that posting
- "Agent On / Off" portlet – allows the user to request a number of general "interventions", such as help in using the platform. These are delivered by the embodied agent. **Scenario N1** – (Initial) learning guidance – is implemented here, in that the user may request the agent to "walk around" the platform and describe the most important elements, their use and usefulness.
- "Agent", Pending interventions portlet – the main initial communication point between agent (not embodied agent) and user, detailing "interventions" (i.e. suggestions, such as specific postings the user may benefit from reading). This portlet is shown on the Home page. **Scenario N3** – Notification of events – is implemented here, in that the user is shown a list of the most important postings for them to read
- The Agent tab – calls up an expansion of the Agent portlet, listing a greater number of "interventions" and allowing details to be displayed of a selected intervention

- "Knowledge" portlet – lists the recent most popular postings
- The "Knowledge" tab – calls up an expansion of the Knowledge portlet, listing a greater number of popular items, along with statistical details of popularity over the last seven and 30 days
- "Community" portlet – lists details of recent visitors to the platform
- The "Community" tab – calls up an expansion of the Community portlet, showing more lists, such as "Most prolific authors", along with statistical details of platform use over the last seven and 30 days

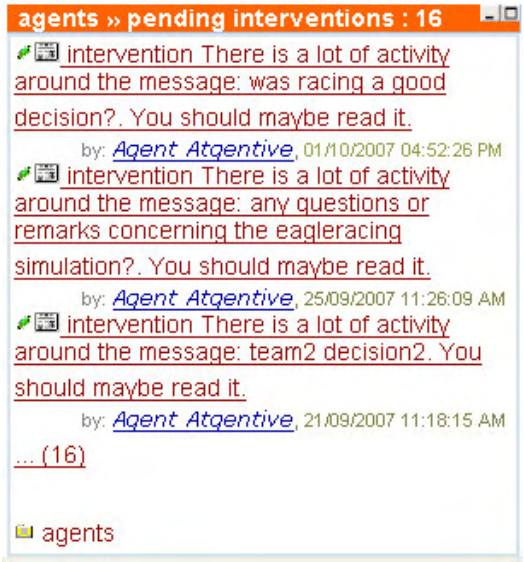


Figure 9 - Example of an AtGentNet "portlet"

2.2.5 The animated Agent

The formative evaluation phase of AtGentive showed that an animated agent in the form of a young woman (see Figure 1) would be the most acceptable character for AtGentNet. This character was named "AtGentiGirl", after "Elastigirl" of popular culture (a "superhero" who appears in DC Comics[1] and the feature file "The Incredibles"[2]). This name was chose due to the character's similarity of appearance to this fictional character and the AtGentNet user demographics (a median age group of = 26-35) fitting with the target audience for "Elastigirl".

For AtGentNet, the animated agent only appears as a result of an explicit user request for help (and only for the experimental group). As described above, the agent is used to deliver Scenario N1 – (Initial) learning guidance. The embodied agent "walks" around the screen, pointing at relevant items and giving an explanation.

[1] Elasti-Girl: http://en.wikipedia.org/wiki/Elasti-Girl

[2] The Incredibles, the movie : http://www.pixar.com/featurefilms/incredibles/chars_pop2.html

2.2.6 Continual development

AtGentNet underwent refinements at the pre-pilot stage. These changes were based upon further analysis of the system by INSEAD, AtGentive partners' feedback and understandings gained from the formative evaluations. However, once the pilot study began only essential changes were made to the software.

2.2.7 User Evaluation Pilot

All students on the 2007 ITM training course took part in the user pilot, which ran for the whole of the training course (May to October 2007). One country was excluded from the study (and used a separate version of the ICDT platform) as it was felt that the internet infrastructure in that country was not sufficiently mature to support the pilot. This gave 27 participants for the pilot.

All students from any one country were allocated the same group (Control or Experimental). This ensured that students most likely to discuss the platform in the initial stages were seeing the same interface. Countries were allocated randomly into groups. 13 students were in the Experimental group (using the new AtGentNet version of the ICDT platform) and 14 students were in the Control group (using the same version of the platform, but with the advanced features disabled or not visible). Participants were not explicitly made aware of the difference in the two systems, and each participant could see and interact with all other participants, regardless of group.

The user pilot took place in two stages:

- General use – all participants, 24th May to 4th September 2007
- Simulation Exercise – self-selected participants, 5th September to 10th October 2007

The second stage comprised a business simulation exercise. Participants were sent an email asking them to take part in the simulation by watching an online video presentation describing a business situation in which the fictional Eagle Racing company attempts to find sponsors for their motor racing team. After the first video participants were asked to discuss the dilemma and vote secretly on their choice of action (one of two available). Participants were then allocated into one of two groups, according to their decision. For the next three weeks each group then discussed future dilemmas using a separate private discussion area on the platform.

AtGentNet was developed using the Extreme Programming methodology (Beck, 1999). This requires an iterative process of develop-test cycles. However, it is crucial to bear in mind the limited time available for AtGentNet users to use and report on the system. The business professionals involved (see Section 3.1.3) were not able to make multiple tests of the system. Therefore, the partners and user representatives at STC offered feedback on AtGentNet during the development stage. The platform then remained effectively stable during the user pilot.

2.2.8 Timetable of events

Below is shown the timings for evaluation activities related to AtGentNet.

May

User evaluation pilot began with an introduction to AtGentNet for TRIM participants at meeting in Lidköping, Sweden on 24th May 2007

September 2007

"Eagle Racing" simulation began on 5th September 2007 with interested (self-selected) TRIM participants

Initial questionnaire – users' background and use of and attitudes towards the ICDT platform and the animated agent – sent out 20th September 2007

October 2007

"Eagle Racing" simulation completed on 10th October 2007.

3. Tools for Summative Evaluation

This section describes the development of suitable tools – evaluation instruments – for the summative evaluation of AtGentSchool and AtGentNet. Before development of these tools begins, it is necessary to consider the user-groups and the tasks they will be asked to perform.

3.1 Description of users and tasks

3.1.1 Description of users - AtGentSchool

Five elementary schools in or near Prague in the Czech Republic participated in the user trial, with either one or two classes per school. In some cases, only a portion of the regular class participated in the trial. Number of students participating per class therefore ranged from 6 to 14 per class, with 6 classes. The modal age of students was 11 years. Students had previous experience of working in groups, though not necessarily while at the computer. The teachers assessed students as to their Ability, using the criteria of Communication Ability and Interest in the Outside World. Students were allocated into pairs for working on AtGentSchool. This was done by friendships, but in practice tended to pair students of similar ability.

The students worked in their native language (Czech). All teachers either spoke good English or had access to the appropriate translation / interpretation facilities. Facilities and infrastructure at the schools (such as computers and network connections) generally met those anticipated of a contemporary UK school.

3.1.2 Description of tasks - AtGentSchool

AtGentSchool is built on the Ontdeknet eLearning platform, created by the Dutch company Ontdeknet. This platform is an electronic learning environment that makes knowledge and skills in society accessible to educational institutions in general and individual students in particular. Virtual learning relationships between subject experts and students are established in this virtual learning environment. The Ontdeknet environment provides guidance to support individuals to learn together based upon common interests.

The students were set the task of examining their home country (Czech Republic) and a foreign country (New Zealand) using AtGentSchool, with a view to deciding in which country the students would prefer to live. A colleague teacher with experience of New Zealand acted as expert-traveller.

3.1.3 Description of users - AtGentNet

The Swedish Trade Council (STC) runs a distance learning course for business people entitled TRIM (Trade Management Implementation). The TRIM project implements the new International Trade Management (ITM) concept, devised by STC. This concept aims to help SMEs (Small and Medium-sized Enterprises) to grow internationally. Participants of the distance learning TRIM course are based in Namibia, South Africa, Slovenia, Norway, Iceland, Hungary and Sweden. The TRIM project ran until the end of October 2007. The participants continue to have access to a “Virtual Trade Network” offering life long learning, career and business opportunities.

Trainees within these seven countries comprised 46 managers who followed classes in their own country and internationally as a group. The classes were on subjects about international business issues; examples include cross cultural training, change

management, market research, international seminars and technical expert knowledge. Trainees were supported by locally-based coaches / tutors who provided expert advice and assistance.

Each trainee collaborated with a tutor, working together for 0.5 days per month on what was termed the “coaching material”. This material consisted of questionnaires in different sections, which the tutor and the manager selected and discussed together. These questionnaires relate to the individual situation of each manager and his / her position on different issues with respect to exporting; for example, the selection of the market and appropriate products. The tutors are trained in coaching skills and additional background knowledge.

3.1.4 Description of tasks - AtGentNet

Prior to AtGentive, the IDCT platform had been used by the tutors and project administrators only to post information and to provide trainees with presentations of the lectures and supportive information, i.e. the platform was mostly used to store information for the participants. While the AtGentive-modified platform does support this form of use, through enhancement of perception, interventions added in relation to Conceptual Framework-related scenarios require interactive use of the platform. It was therefore necessary to create an extra task for the participants to work on that required interaction (such as posting replies to documents).

It is important to note that the managers involved in the TRIM project are very busy with their main work, and have difficulty finding extra time to participate. Therefore this task consisted of a business simulation exercise (see Section 2.2.7) which offered additional experience of business decision-making. Participants who chose to take part in this extra activity were required to collaborate with each other, interaction necessary for the testing of AtGentNet interventions.

3.2 Derivation of the Key Indicators

The evaluation process assesses the effectiveness of the AtGentive interventions against a set of Key Indicators. The generation of the Indicators is described in detail in deliverable D4.3, and is summarised here. (See Figure 10 for diagrammatic example.)

The process began by analysing the scenarios to be implemented in AtGentSchool and AtGentNet (see Section 2) to isolate their essential elements. For example:

AtGentSchool - Scenario S1 – Learning guidance
AtGentSchool - Scenario S2 – Re-attracting an idle-user attention

Meta-level scenarios were created in order to amalgamate conceptually comparable scenarios. This produced three meta-level scenarios as follows:

Meta-scenario a) propose a task [S1 and N1]
Meta-scenario b) recover user’s attention [S2 and N2]
Meta-scenario c) notification of an event [S3 and N3]

The next step was to operationalise these meta-level scenarios by identifying question(s) that best evaluate each scenario’s success in usefully controlling the users’ attention – taking into account the intended outcome(s) of each scenario. These are referred to as the Strategic Questions, intended to be representative of those implied by each scenario, rather than a comprehensive list of all questions that may be asked. For example:

Scenario (a): Propose a task

Strategic questions:
"How is the user's work affected by being proposed a task?"
"How do users feel about being proposed a task?"

The scenarios, then, inform the creation of strategic questions—those questions that best interrogate the effectiveness of a particular scenario. The Key Indicators are essentially abstractions of these strategic questions. In order to achieve this abstraction, the most critical phrase(s) from each strategic question are taken as indicative of the effects that question is investigating. For example:

"How is the user's work affected by being proposed a task?"
work affected by
"How do users feel about being proposed a task?"
"How do users feel about having their attention directed?"
users feel[ings]

The minimum number of key indicators were then identified that would measure the concepts behind the identified critical phrases. The Key Indicators identified at this stage are:

Performance, Attention and Satisfaction

These Key Indicators are important for the measurement of anticipated change brought about by implementation of the scenarios. It is also important to look for improvements in the primary functioning of the platforms and to ensure that other elements at work for the users are not impacted detrimentally by this implementation. AtGentSchool is primarily an educational tool, for use by school children in a classroom setting (see section 2.1). Therefore, the primary goal of that situation – learning – must also be examined. For AtGentNet, collaboration is a key element in its effectiveness and it is considered to enhance directly its overall learning effectiveness by providing "communication channels for 'learning groups' operating within or across companies aiming at different forms of synchronous or asynchronous knowledge and social exchanges" (Angehrn, 2004). Therefore Learning and Collaboration are additionally required. This completes the generation of Key Indicators:

Performance, Attention, Satisfaction, Learning and Collaboration

The generation of these Indicators is described in detail in deliverable D4.3, and in diagrammatic form (in part) in Figure 10 below.

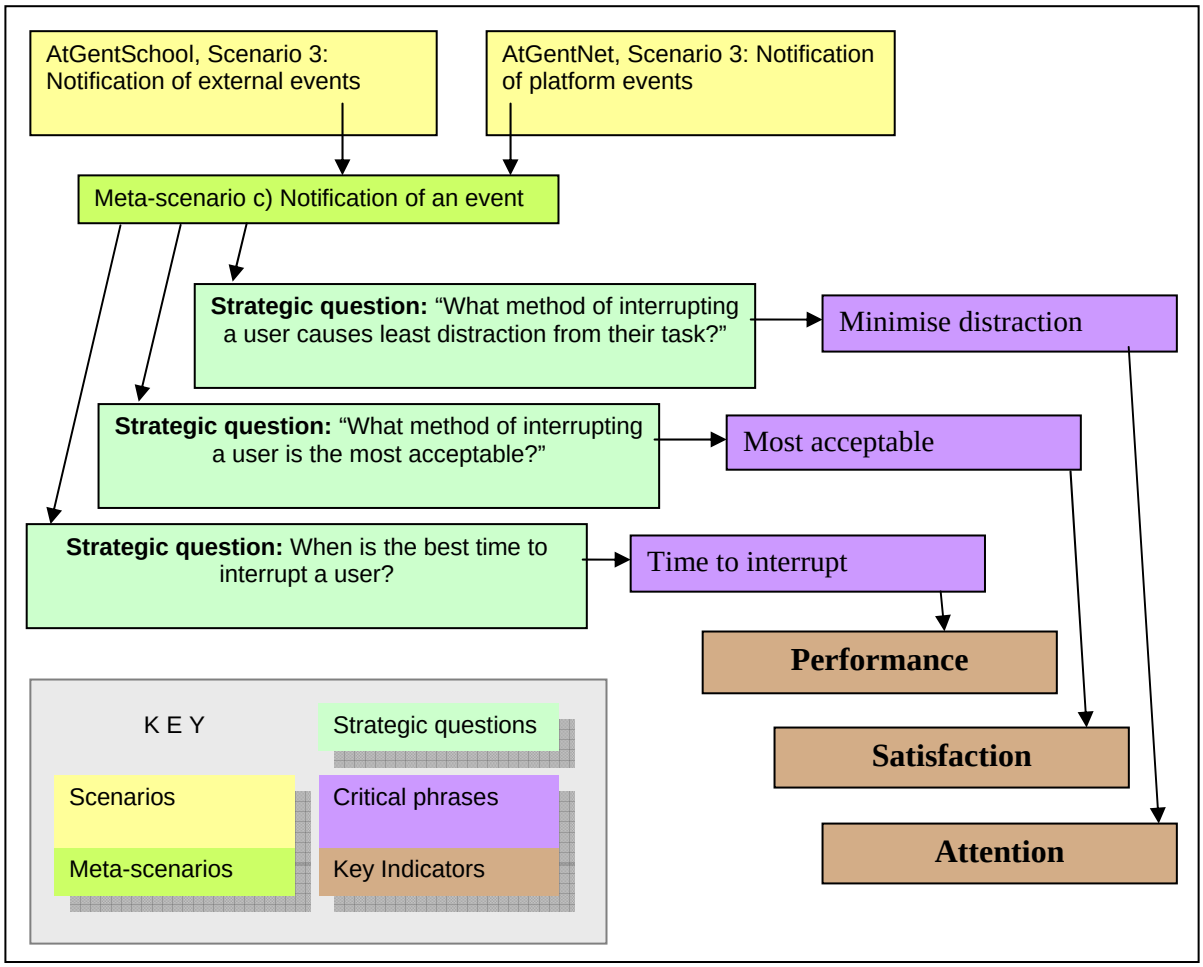


Figure 10 - Derivation of the Key Indicators from the Scenarios (part)

3.3 Context, Interaction, Attitudes and Outcomes

The design for the evaluation plan incorporated an adaptation of an evaluation framework developed for the evaluation of Computer Aided Learning packages. The CIAO! Framework (Jones et al., 1999) is based on the interaction between Context, Interaction, Attitudes and Outcomes. Context comprises the original aims and goals of the system designers. Interaction is the way in which students use the software – interplay between usability and activity. Attitudes and Outcomes are the results of interaction – effects on the students (such as frustration or being supported) and desirable outcomes (such as improvements in learning or collaboration) (see Figure 11).

- **Context**: the original aims and goals of the system designers. In the case of AtGentive it is primarily the aims of interventions in relation to the scenario being enacted.
- **Interaction**: the way in which students use the software. This is an interaction between the software's affordance (Gibson, 1979) (and thus usability) and the user's activity.
- **Attitudes and Outcomes**: the results of using the software. Attitudes refer to the students' perceptions, such as frustration or a feeling of being supported.

Outcomes refer to students' achievements (such as improvements in learning or collaboration).

Note that although Usability was specifically tested prior to the user trial (using heuristic evaluations) it remained relevant during the user pilot, not so much as an element for measurement in itself, but as a factor that may affect other measurables. For example, students may have generally liked an embodied agent, but may see one particular interaction as impolite within their circumstances of use. This may have affected their interaction with the agent at that point and thus the outcome of that specific intervention.

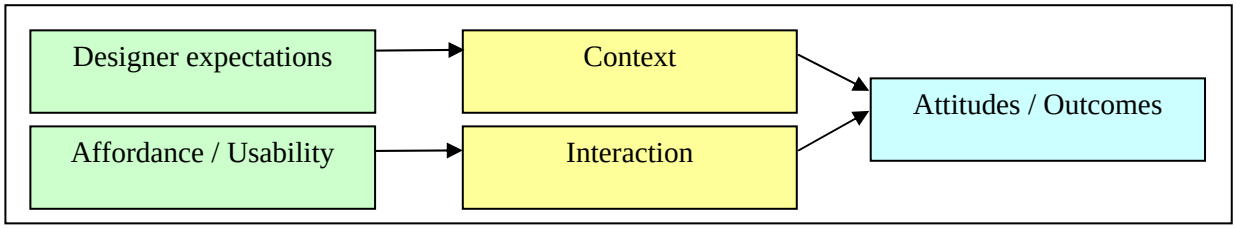


Figure 11 - Interconnection between Context, Interaction, Attitudes and Outcomes

The application of the key indicators has been with reference to this framework. Examples are shown below:

- Context: e.g. Comparison of expected and actual student activity following an intervention (or, for the control condition, where an intervention would be)
- Interaction, e.g. Proportion of correct / incorrect responses to an intervention (taken from the system log file)
- Attitudes: e.g. Rating scales for emotive response (student questionnaire)
- Outcomes: e.g. Expert assessment of learning success

3.4 ISO 9241 for Satisfaction and Performance

3.4.1 Overview

The EU body International Organization for Standardisation (ISO) has created an standard for usability: ISO 9241-11:1998 (ISO, 1998). This standard discusses the evaluation of computer software and gives a comprehensive set of indicators which may be employed for this purpose.

To begin, ISO 9241-11 defines usability as the

> "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use".

In the context of AtGentive, this may be expanded to:

> Usability is the extent to which AtGentSchool can be used by the school children and AtGentNet can be used by the business learners to achieve their learning and collaboration goals within their school or business community, in terms of effectiveness (accuracy and completeness), efficiency (resources expended) and satisfaction (freedom from discomfort – physical and psychological – and with a positive attitude).

3.4.2 Effectiveness, Efficiency, Satisfaction

"Effectiveness" refers to the results the user is able to achieve by using the software. This is a measure of what may be achieved and the quality of the finished result, in relation to what was expected. "Efficiency" refers to the user effort required to achieve the finished result. This is a measure of the process of achieving a result, the amount of time and effort required of the user, the problems encountered along the way and the software's success in assisting the user. Together, "Effectiveness" and "Efficiency" are taken here as indicators of "Performance".

"Satisfaction" is a subjective concept and as such it is difficult to use empirical measures. It is therefore usually measured by asking the user to estimate their satisfaction on a Likert-style rating scale, as well as asking satisfaction-related questions, for example, "I feel in control when using AtGentSchool". Other, less subjective but none-the-less indirect, measures may be used, such as "number of complaints". There is also the possibility of measuring some elements of satisfaction empirically. For example, how much software is used when use is voluntary?

Finally, there is evidence to suggest that satisfaction and dissatisfaction are separate factors and may not correlate (Jokela, Aikio, & Jounila, 2005). This is because "satisfaction" is built from a number of elements – the individual perceptions and experiences of an artefact. For example, one may be satisfied with a cup of coffee due to benefits of its caffeine content, yet dissatisfied with the same coffee because of its flavour. Therefore questionnaires ask about both satisfaction and dissatisfaction.

The following sections detail the forms of measures for each of the Key Indicators. "Satisfaction" and "Performance" incorporate elements from ISO 9241 as these are seen as Key Indicators that could be particularly affected by subtle usability issues. However, in accordance with the CIAO! Framework (see Section 3.3) the full range of system use is considered in a holistic manner, from usability and user activity, to user attitudes and outcomes.

3.5 Heuristics

The purpose of using heuristics as a software evaluation tool, prior to testing with users, is to trap and remove a large proportion of usability problems before user testing begins, and in so doing to validate the software as being to appropriate usability standards for the user pilot.

The most-used heuristics for usability evaluation were developed primarily by Jacob Nielsen during the 1990s, with particular emphasis on the World-Wide Web (Nielsen & Mack, 1994). They were developed originally from an examination of usability problems found in one piece of software by a large number of "usability specialists" (Rolf & Jakob, 1990). The list was later refined using a formal factor analysis of a larger number of usability problems (Nielsen, 1994). For a full list of Nielson's heuristics see Appendix 1. However the heuristics most appropriate to this study are:

- Visibility of system status
- Match between system and the real world
- User control and freedom
- Consistency and standards
- Error prevention
- Recognition rather than recall
- Flexibility and efficiency of use
- Aesthetic and minimalist design

- Help users recognise, diagnose, and recover from errors
- Help and documentation

AtGentive interacts with the user, and as with any human-computer interaction situation, these “standard” heuristics would be expected to apply. For example, where the “interaction” is an animated character suggesting that the user open an additional document, the words used should be understandable by the intended users (“Match between system and the real world”), if the suggestion is accepted it should be reasonably easy to close that document (“User control and freedom”), the document should not be opened using unfamiliar software (“Consistency and standards”) and so forth.

In addition, new heuristics have been created to assess AtGentive-related interactions. As with the “standard” usability heuristics, these additional heuristics are based upon examination of the area (in this case, the proposed scenarios)

These heuristics are derived from the Key Indicators (see Section 3.2). As with the “standard” usability heuristics described above, the Key Indicators – and thus the additional heuristics – are based upon examination of the area (in this case, the scenarios to be implemented).

Key Indicator One: Attention

Success in attracting attention

Where the system attracts the user’s attention, it should do so in a manner that will not be accidentally overlooked or misinterpreted

Distraction is minimised

The user should not be interrupted in their task, unless the interruption assists that task significantly or is justified by the importance of the interruption. Where appropriate, interruptions should be delayed until the user is less busy.

Any animated agent should not be unduly distracting

Key Indicator Two: Performance (Effectiveness and Efficiency)

Task is performed well

Interventions should not cause a task to be performed less well overall. Where the intervention is intended to improve the performance of a task, it should do so.

Key Indicator Three: Satisfaction

Overall satisfaction

All suggestions / interventions made by the system should appear to the user to have at least some effective purpose. The user should not consider any suggestion to be “pointless” or “stupid”.

Positive image of the animated character

The user’s immediate reaction to seeing any animated character should be at least neutral and preferably positive. The user should anticipate that the character’s appearance will make their task easier, not more difficult. The user should not have negative feelings about the animated character (threatened, humiliated, etc.)

User control and freedom

This is an extension to the “standard” heuristic. The user should feel in control of the AtGentive interventions. The user should not be worried that they will be interrupted at any moment, or that they are likely to miss something important

Key Indicator Four: Learning

Learning experience is supported

Interventions should not cause the learning experience to be degraded. Where the intervention is intended to improve the learning experience, it should do so.

Key Indicator Five: Collaboration

Collaboration is supported

Interventions should not discourage collaboration. Where the intervention is intended to improve collaboration, it should do so.

3.6 Questionnaires

The methodology followed for AtGentive questionnaire development was:

> Determine clearly what you want to measure
> Generate an item pool
> Determine the format of measurement
> Have the initial item pool reviewed by experts
> Consider inclusion of validation items
> Administer items to a development sample
> Evaluate the items
> Optimise scale length
> (DeVellis, 2003. Chapter 5)

We wished specifically to measure the Key Indicators (see Section 3.2) along with general demographic information. The “item pool” (individual items to be measured) comprises those already identified in deliverable D4.3, a number of papers from previous studies, general texts on surveys and further contemplation of the evaluation requirements. This created a list of concepts to be measured. An example is shown below, with an abstract concept followed by a specific question. A full list is given in Appendix 2 and in Appendix 3.

> General (own) satisfaction with AtGentSchool [First day] [Bi-Weekly]
>
> **I am completely satisfied with AtGentSchool** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)

3.6.1 AtGentSchool

Questionnaires were administered to the students at the end of a class using the computers they used for AtGentive. The school environment restricted the number of questionnaires that could be fitted into class time. Two questionnaires were administered per class, one around mid-way through the trial and one at the end. (It was not possible to administer one in week one as intended due to technical problems with the software and lack of class time.)

In addition, a questionnaire-format pre- and post-test was administered to assess learning gains per group.

Questionnaires for the teachers were self-administered around the time of the children's questionnaires (each individual teacher chose when to complete their questionnaires).

In addition, the teachers were invited to provide feedback in the form of a written weekly log.

3.6.2 AtGentNet

Questionnaires were administered to the business users at the beginning of the Simulation exercise. An email was sent to all users on each occasion requesting their participation and including a link to the on-line questionnaire.

3.7 Log files

Both systems generated log files detailing the system's use. Log files are respectively in text format, and in XML Atom format. Typical examples are shown below for AtGentSchool and AtGentNet respectively:

Atgentive event: (text format)

2007-06-04 21:26:03,870 DEBUG - New application event: EVENT:16727 BREAKPOINT screen=st_assignment,module=ASSIGNMENT > DESCRIPTION (make - assignment)

Atgentnet event: (Atom format)

```
<entry>
  <title>Thierry Nabeth  logged into: the Community</title>
  <author><name>Thierry Nabeth</name></author>
  <link href="" />
  <id>urn:uuid:1E1D6</id>
  <updated>2007-12-04T16:41:25+02:00</updated>
  <summary>Title: the Community, by: Thierry Nabeth,
                  04/12/2007 04:41:25 PM CET</summary>
  <category scheme=http://atgentnet.com/resource/type/1.0
                  term="community" />
  <category scheme=http://atgentnet.com/resource/name/1.0
                  term="theCommunity" />
  <category scheme=http://atgentnet.com/event/type/1.0
                  term="logged into"/>
  <category scheme=http://atgentnet.com/event/user/1.0
                  term="person:Thierry Nabeth"/>
 </entry>
```

In the case of AtGentnet, tagging is also used to add metadata describing more precisely the event (such as the type of the event).

3.8 Interviews and Focus Groups

Following the AtGentSchool pilot, a semi-structured interview was held with the participating teachers, the expert and members of AtGentive.

Towards the end of the AtGentNet pilot, telephone interviews were conducted with a sample of six participants of the TRIM course.

3.9 Pilot Evaluation criteria

The effects of AtGentive modifications on the Experimental group were assessed according to the Key Indicators (See Section 3.2). The application of these indicators is described along with the results in Sections 4.2 and 4.3).

4. Summative Evaluation - Results

4.1 Heuristic Evaluations

4.1.1 AtGentSchool

The full results of the heuristic evaluation are included in Appendix 4. In total, 39 suggestions were made for improvements. A summary is given below. It was not considered that any of the issues found would prevent the successful conduct of the pilot study.

- Possible problems with parts of the system available prior to login
- Slow response when clicking on the “smileys”
- Difficulty with navigation outside of the home screen
- Improvements suggested for the agent’s behaviour

4.1.2 AtGentNet

The full results of the heuristic evaluation are included in Appendix 5. In total, 77 suggestions were made for improvements. A summary is given below. Some changes were deemed necessary and carried out to the software during the early stages of the pilot.

- A number of important usability improvements needed for the interface
- General lack of transparency between screen options and system function
- A number of words and concepts of a technical nature – may cause problems for the general user
- The agent is not enough proactive, and takes no action without a direct user request

4.2 Pilot study – AtGentSchool

Data collected as part of the AtGentSchool pilot, and available for analysis, is as follows:

- assessment of students’ work
- log files
- questionnaire responses
  - students pre-trial test
  - students feedback
  - students post-trial test
  - teachers feedback
- teachers’ diaries
- post-trial workshop

4.2.1 Teachers’ experience

4.2.1.1. Teacher Questionnaires

Teachers were asked to complete an attitude survey as soon after the students completed theirs as practicable. This provides an interesting insight into the practicalities of using AtGentSchool in a classroom setting. Two questionnaires were completed; the first soon after the pilot began, the second towards the end. (The exact dates of

completion varied between teachers, but the first represents thoughts about the first weeks of the pilot while the second was after many initial problems had been resolved and the teachers had had time to reflect on the pilot.)

First questionnaire

Initially, the students needed to get used to the AtGentSchool paradigm, i.e. that instructions, information and students' work are all mediated by the computer. Students 11 years of age in Czech schools are used to receiving instructions from the teacher. At first, students would read instructions from AtGentSchool, but would not act upon them, expecting them to be confirmed / reiterated / given by the teacher:

> "The children are not used to work[ing] individually without clear instructions or they are not used to be given the instructions through the computer like this."

> "The children didn't want to read and listen to the instructions and follow them. They wanted all the time the clear instruction how to fulfil the task without working on it."

The teachers had been asked not to explicitly explain the details of using AtGentSchool, but to let the students explore the software for themselves. However, as the students did not properly understand that their tasks were allocated by the software, and not the teacher, there were many problems at the beginning:

> "Mostly children don't know what to do."

> "For them it was a kind of game because there was Honza's figure and … they were clicking and clicking without any reason. They were only trying what happens if....I am not sure that they caught the idea of AtGentSchool."

> "I am afraid that they didn't catch the main idea and they lost themselves in following the step -by-step instructions."

This made the teachers task at first very difficult, as they had not been trained to support students in the use of AtGentSchool at a step-by-step level (since it was expected that the students would simply follow the on-screen instructions):

> "I didn't feel the control over the lesson because the children were unconcentrated trying to ask me all the time what to do and give them the final solution what to do … how to operate the programme. And I wasn't able to answer their questions."

In addition, the "overhead" of running a pilot was very significant at first:

> "There were many jobs to do. Filling questionnaires, failing computers..."

Despite these problems, the teachers remained optimistic, realising that the first lesson or two were not representative of using AtGentSchool overall:

> "Children were interested to use platform."

> "They need much more time to feel comfortable with the programme because the way of work like this is completely new for them. They need to get used to it then it should work OK."

> "The second lesson they were successful in filling in the first activity - write about you and your hobbies."

Teachers rated the students' overall use of time early on in the pilot using a rating scale of 1-7 on four activities, as shown in Figure 12.

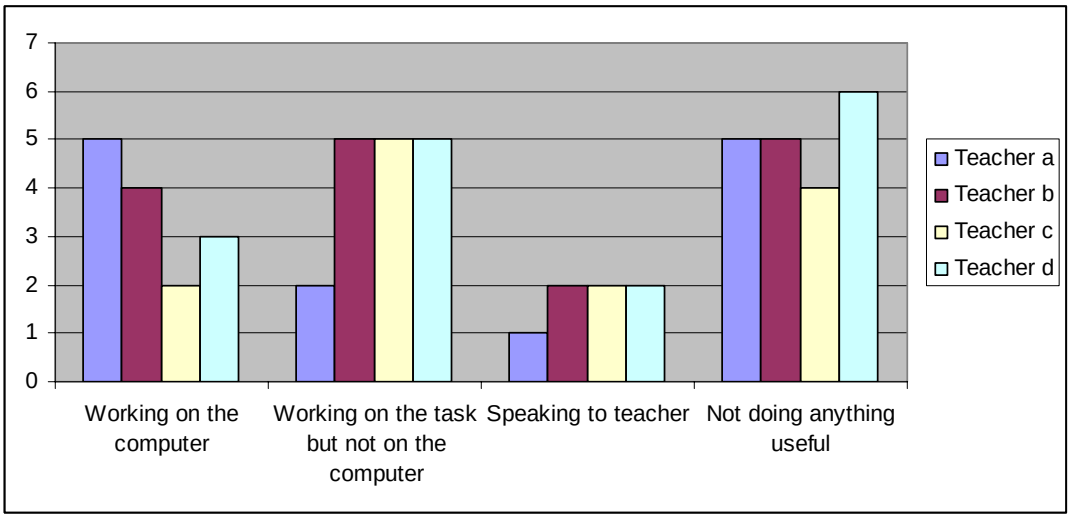


Figure 12 - Teachers' rating of students' activity during the early part of the pilot

Second questionnaire

By the second questionnaire, teachers were able to reflect on the whole pilot, giving a more balanced view. As regards the new way of working – taking instructions from the screen – it took time to adjust:

> "For children it is easier to ask teacher then to try to understand how the platform works or what words in questionnaire means."

> "The students were asking at first me (because it's their usual system of work) then when I turned their concentration to the system they worked on the platform"

> "When the students get used to [AtGentSchool] they were able to overcome difficulties (but not everybody)."

Generally, the teachers would have liked more time in which to run the pilot:

> "I think that they liked the personality but I am not so sure that they were fully satisfied - but it was mainly different system of working with the computer (they are some how used to search the internet and [AtGentSchool] was completely new for them and I think they needed much more time to feel comfortable when using this new system)"

Although students had experience of working in groups (pairs), it seems that the type of group-work required by AtGentive was different, creating problems for the students:

> "The students are not used to work independently in such way they are not used to study by reading texts themselves. The system there leads them to summarize but not select the information from such a big amount of data (they are not used to study in groups and cooperate in this way)"

> "[AtGentSchool] was completely new system of work so the students had to invent a new way how to be successful or how to overcome difficulties (some of them were successful). Some of the couples had so many problems in social cooperation that they failed in project work."

One teacher found particular problems with students that missed one or more classes:

> "Problem was when children were missing at school … For me was difficult to 1) explain how to use platform 2) explain what is their schoolwork 3) answer their

questions (words in questionnaire) 4) control children 5) fill my questionnaire in one time"

One observation in particular was noted, and considered throughout this report:

"in my class in experimental group were girls. They work much more they have more patience and endeavour."

Generally, the experience was seen as positive for the students:

"In the beginning it was for children very difficult but after few lessons it was better."

"For children it was something new they like to use computer new type of project."

Teachers again rated the students' overall use of time using a rating scale of 1-7 on four activities, as shown in Figure 13.

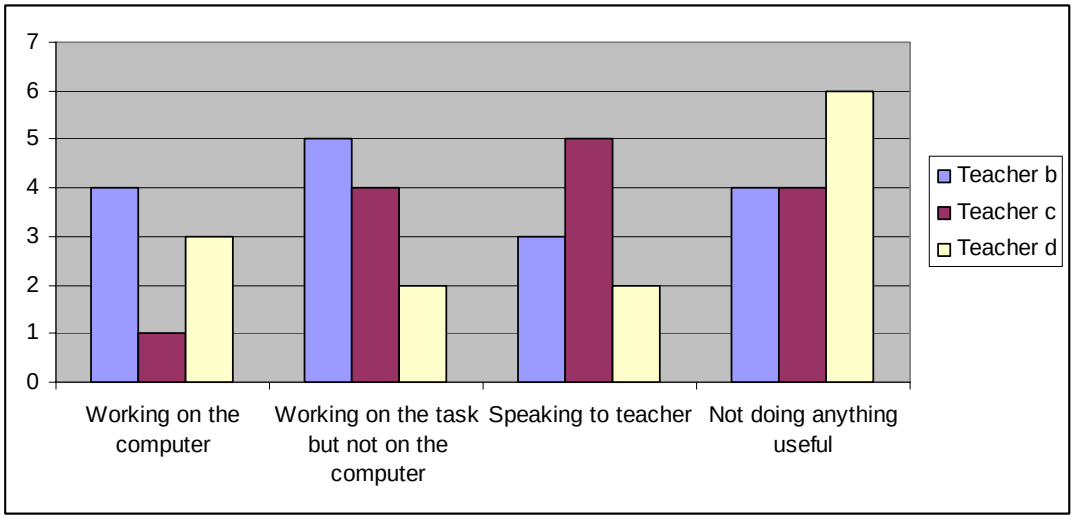


Figure 13 - Teachers' rating of students' activity during the latter part of the pilot

Comparing the first and second questionnaires, Figure 14 shows how the teachers' view of student activity changed over the course of the pilot. (Note that the data is taken from a seven-point rating scale and thus percentages are of limited accuracy – they should be used as a guide only.) Initially, teachers felt students spent more time not doing anything useful than any other single activity. Later, the ratings become more even, with the highest rating for working on the task (not on the computer) and an increase in the amount of time speaking to the teacher.

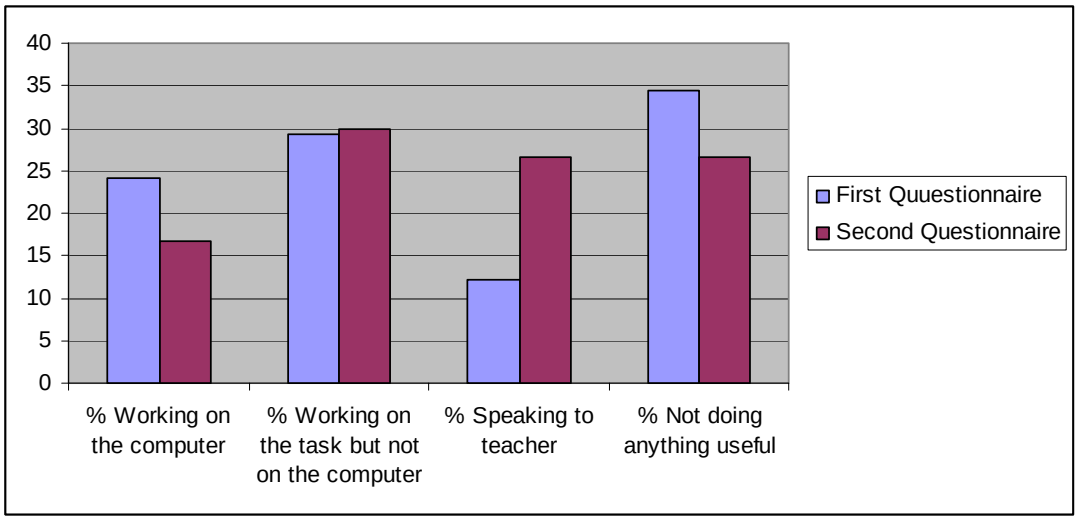


Figure 14 – Change in mean - teachers' rating of students' activity

4.2.1.2. Teachers' diaries

A number of comments were made regarding the software, which have been passed to the development group within Ontdeknet for consideration. It is important to consider the comments in terms of AtGentSchool as a whole. It is important to consider the comments in terms of AtGentSchool as a whole. For example, allowing two windows to be open at one time could create new usability problems, may not be technically practical, or could change the whole nature of the software:

> "The problem was, that it is not possible to have open two windows at the time, therefore they had to return to the diary repeatedly which was time inefficient."

> "They did not know where to write the answer, this had to be instructed. Buttons in application are not clear."

Comments on the students' initial expectation of instructions from the teacher, rather than the screen:

> "This way of communication is very unusual and distant for them - they cannot replace computer for teacher - cannot accept my 'idle' role - and keep on turning to me and ask me questions. I tell them to read the screen and look for the answers there, but they are not interested very much. They are not used to that."

> "Children often ask for help – mostly they would be instructed to read what is on the screen, sometimes the individual instruction is given."

By midway through the pilot, the students were settling in to the lessons:

> "[AtGentSchool] is operated by children fairly confidentially, they also use circle icon (ASK-READ-WRITE)."

> "They do not pose too many questions to me, it is mainly concerning the content of the paper"

By the last lesson, most of the earlier problems had been resolved:

> "The last lesson … ran smoothly. Children worked on the platform without any problems."

> "As it was the last lesson, children had no problem to sign in, orientate on the platform (mainly) and to fill in the last task."

The time available for students was problematic:

> "The biggest problem is the time, 45mins is not enough for them, mostly they use also break – so end up working for 60mins on the pilot a week."

> "I think 6 lessons are not enough for them, in one week sequence they tend to forget everything … In order to be aware of the platform functioning they would have to use it more often and longer then set in the pilot."

> "I think, if children could [work] with the platform for unlimited time, they could get used to it more easily … The biggest enemy of the project is in my opinion the length of the pilot."

Differences were observed between the Control and Experimental groups:

> [Earlier] "Experimental group works faster. They were done with work at least 10mins in advance. Other control group used whole 45min session … The children who had so called "stupid Honza" were working more slowly and were not so certain. Otherwise they were working with pleasure and enjoyed the time."

> [Earlier] "There is a huge difference between control and experimental group"

> [Midway] "Control group often asks for help. Children do not seem to realize the fact, the most of the instructions they also can find on the screen. Experimental group works fine."

> [Later (Second Questionnaire)] "The differences between the two groups - control and experimental were not so big as I thought at first."

4.2.1.3. Teachers' overall experience

Amalgamating comments from the teachers' questionnaires, diaries, and discussions from the end-of-pilot teachers' workshop, we see a clear picture of the teachers' experience. We present a summary here, along with recommendations for future inter-country projects.

Schedule

Scheduling the pilot just before the end of the school year meant that three of the classes took a one-week school trip during the pilot. This meant they had forgotten much of what they had learned about AtGentSchool by their next lesson.

Generally, it is easy to forget the very constrained time-frame in which schools must work. Every requirement for class or teacher time must be planned well in advance if it is to be made available from other commitments.

Conclusion: Work very closely with schools well in advance, taking account of their existing timetable. Fix dates and times at least three months before the event and do not expect to make any changes or additions.

Technical

There were significant technical problems encountered at the beginning of the pilot. In particular, the Czech schools prefer a more secure and complex firewall than is usual for schools in several of the other participating countries. This was unanticipated and therefore made implementation more complex.

Conclusion: Different technical background and settings must be considered in advance.

The Czech language uses a character set that was not easy to implement in some areas of the software and the electronic questionnaires. The teachers found a “workaround” of asking students to use “SMS writing” until the problems could be solved, but this was not an ideal solution.

Conclusion: Detailed technical requirements and tests are needed before the pilot. Once the pilot starts it is extremely difficult to improvise or to be flexible within the school environment. In particular, a (computer systems) administrator is often not available at a moments notice; the teachers may not have the technical knowledge to report bugs accurately. In addition, they are not able to call / email immediately a problem occurs – they usually have no technical support on the spot.

Teaching practice

The usual method of teaching computer classes in the participating schools was for students to take their instructions from the teacher. Therefore, while the AtGentSchool software itself provides instructions, the students were reluctant to follow them, since their experience was that they should wait to be told by the teacher. In addition, the teachers were used to instructing the students. The software did not appear to instruct students to the extent teachers were anticipating. This encouraged teachers to give instruction themselves, rather than stand back and watch the students not understanding what they should do.

Conclusion: It must be remembered that the teachers first priority is to enable the students to learn, not to conduct an experiment. The software – and the experimental situation – must assure teachers that their students are able proceed with their learning. If not, the “experiment” will be adapted to ensure that the students are not disadvantaged.

While the concept of group work was not new to any of the students, some did not have experience of working in pairs when using the computer. The teachers ameliorated this situation very effectively by choose pairs of students who were friends.

The students usually work more by collecting and filtering information than coming up with ideas of their own in the way AtGentSchool expects. Teachers addressed this by focussing on concepts such as “Compare”.

Conclusion: Teachers know their students well and can often address problems that may seem difficult to, or not anticipated by, the experimenters.

Students

Students were very motivated to participate in the AtGentive pilot. It was seen as a high-prestige project and very different from their usual school work and allowed them to learn about another country from someone who had visited the country.

The teachers’ view of student activity changed over the course of the pilot. Initially, teachers felt students spent more time not doing anything useful than any other single activity. Later, the ratings become more even, with the highest rating for working on the task (not on the computer) and an increase in the amount of time speaking to the teacher.

Embodied agent

Generally, the teachers concluded that female students were more likely to listen to the agent than male students.

It was suggested that the agent would be more effective if it were to show the students what to do (walk to the appropriate screen location, point, type and have words appear as they would for the students, etc.) as well as describe.

Control group: Students in this group were upset at first that the agent wasn't helping them, but later they accepted that David was not going to help them and got to like having him there.

Experimental group: Less notice was taken of the embodied agent after mid-way through the pilot. The probable reason was because "they realised Honza was not solving their problems". In particular, it was thought that the agent was very helpful at first, but the help they needed changed over time while the agent did not.

The agent did not always leave enough time for the students to read, understand, discuss (between themselves) and take action before giving another instruction.

Conclusion: Further improvements to the embodied agent are possible.

AtGentSchool

In the English version of AtGentSchool the terms "Concept map" and "Mind map" are used interchangeably. When translated into Czech, the two translations had significantly different meanings. This was not picked up by the heuristic evaluation as that was conducted on the English version. During the pilot, the difference caused confusion for the students.

Conclusion: Any translation work needs to be completed well in advance. Time needs to be allocated for a native speaker who also understands the system to check it in detail.

The "smiley" (feedback) buttons tended to be seen as statements, rather than input (i.e. they were smiling at the students).

Conclusion: If possible, test software with one or two representative students from the target schools before a main (or pilot) study.

Difference between groups

Teachers observed that the Experimental group worked noticeably faster then the Control group at the beginning of the pilot. By mid-way, this was still the case, but the Control group had settled in to a steady, though slower, pace. By the end of the pilot, the differences between the two groups did not seem anywhere near as pronounced.

Time

The time available for the pilot placed very high demands on both the teachers and students. In the initial lessons, teachers needed to understand the software, deal with technical problems, explain to their students how to use the system and what to do, new ways of working (taking instructions from the screen, working in pairs at the computer), implement student questionnaires, complete diaries and their own questionnaires, along with the rest of their usual teaching duties. Teachers who were able to allocate more time from their schedule in the short period between availability of the software and the first few lessons of the pilot were in a much better position to assist the students.

Students had to assimilate a new working environment in addition to becoming productive within that environment. With each lesson only 45 minutes in length, and only around six lessons, that significantly reduced the time actually working on their learning task.

Conclusion: Consider the "overheads" – familiarisation with the software and the new working environment and practices, and collection of experimental data – and allow time for these in addition to the actual pilot.

4.2.2 Results from Pilot

Firstly, we note that two possible confounding variables were identified. The first is Ability; the Experimental and Control groups were deliberately balanced for Ability ("Low", "Medium" or "Smart") so this would not be expected to influence the main effects. The second is Gender: there were more girls in the Experimental group, and more boys in the control group (see Table 1). Both these variables are checked throughout the analysis and where necessary their effects are described and discussed.

Control	n		Experimental	n
Male	17		Male	11
Female	9		Female	14
Mixed	1		Mixed	3

Table 1 - Distribution by gender for treatment groups (pairs)

4.2.2.1. Assessment of students' work

Each student's work was originally assessed by their individual teacher. While each teacher was consistent in their marking, discussions with the teachers revealed that there was no overarching marking scheme within the school. It was therefore not possible to compare students' marks across classes. In order to allow this comparison, all the students were remarked by the AtGentive team (Barbora Parrakova and Inge Molenaar) at a meeting in Prague on 28-Jun-07 using an agreed marking scheme. The remarking was "blind", in that papers were taken at random, without reference to any student or teacher information. It was not considered useful to compare this remarking with the teachers' original marks, since all the marking schemes involved are different.

The items marked by the AtGentive team and considered reliable are listed in Table 2. Other items have been excluded due to large numbers of missing data points. The results are shown in Figure 15.

Name	Description
Qns2Xp	Number of questions the student pair asked of the expert
Status	4=Did not do questionnaire / 5=Did questionnaire
Intr.10	Quality of their introduction (Number of topics mentioned, 0-10)
GdGoal	How good their stated goal is (0=not good 1=good)
CMap	Concept map (Number of topics mentioned)
PPara	Number of paragraphs in their paper (students divide text into blocks so it's possible some will place more text per block but examination of the data shows reasonable consistency)
PQual	Paper quality – 1=low / 2=medium / 3=good

Table 2 - List of marked data considered reliable

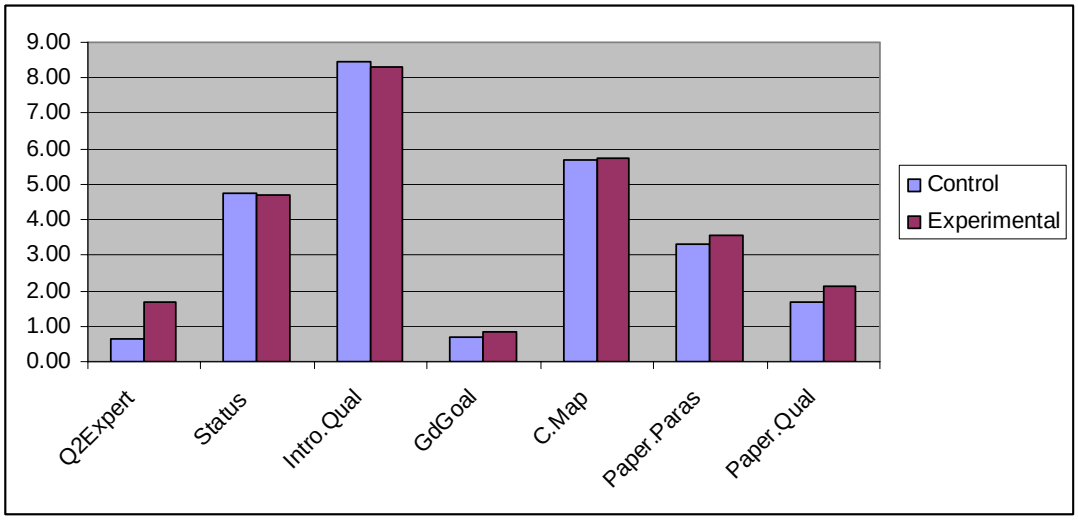


Figure 15 - Students' marks - mean values by treatment

T-tests were the basis for examining the Interval data (such as Qns2Expert) and Mann-Whitney tests for categorical data (such as Status). A significant difference (at the 0.05 level) was found for:

- number of questions asked to the expert (Qns2Expert) (F = 4.659, Significance = 0.035)[3]
- quality of the paper (Paper.Qual) (Z = -1.963, Significance = 0.050) [4]

[3] ANOVA; equality of Variances not assumed (Levene's Test result: F = 11.027, Significance = 0.002)

[4] Kruskal-Wallis test, looking for significance at the 0.05 level

That is, the Experimental group asked more questions and wrote a better quality paper. The other variables were not found to differ significantly between the Control and Experimental groups.

Regarding the possible confounding variables (see the beginning of Section 4.2.2), none of the variables tested (see Table 2) showed a significant difference related to either gender[5] or ability.

4.2.2.2. Log files

The log files are text-format files with one line for each "event" (such as the user changing screens, or being sent an intervention). There is one file per student pair (i.e. one file per computer) comprising Date, Time, Event type, Event details. The main events available are:

- User changes of screen
- Interventions
- Feedbacks (clicks on the "smileys")

A typical log file would contain data of the form:

```
2007-05-17 11:06:36,461 - Sending intervention: INTERVENTION:16747 TEXT_ES_HAPPY1

2007-05-17 11:21:33,787 - New application event: EVENT:16747 START_TASK
diary#3879_18071#16747

2007-05-21 12:32:27,446 - New application event: EVENT:16734 FEEDBACK  userstate=confused
```

OBU created an analysis program that reads in all the log data and picks out selected events, outputting totals for further analysis. (For example, "Find the total number of "happy" feedback clicks for those who were assessed as stating a good goal compared to those who did not.)

Length of time spent on each system element

Eight system elements were available to the students. Each element was selectable and displayed its own screen. The options were:

- projectmanager – The main "Home" screen
- persinfo – Students describe themselves and their interests to the expert
- assignmenttarget – Students describe their goal in using the system
- conceptmap –Students create a concept map of main points to investigate
- paper – Students describe the two countries in textual form
- diary – The expert provides the students with real life information and experiences in dairies
- forum – Student's pose questions to the expert who answers the questions in a forum
- question – Students pose a question to the expert

[5] Jonckheere-Terpstra test, looking for significance at the 0.05 level. This test was used instead of Kruskal-Wallis in order to take account of the Ability (Low→Medium→Smart) comprising an increasing scale

The time spent "on" a system element (i.e. with that system element on-screen) was calculated as the time between "START_TASK" events on the log files. (It was also necessary to end the task timing at the end of each session, since no "LOGOUT" events were recorded. In this case, the last event prior to an "INIT_APPLICATION" event was taken as the end of the previous session / system element.) Figure 16 and Table 3 show the results of a one-way Analysis of Variance (ANOVA).

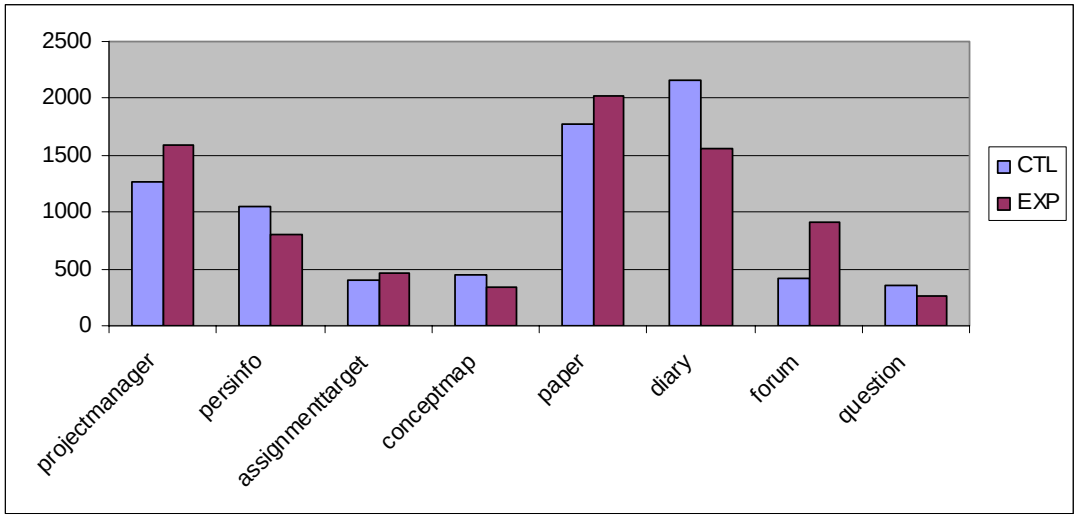


Figure 16 - Mean times per system element by treatment group

		Mean	Std. Deviation	Significance
projectmanager	CTL	1261.78	959.276	0.244
	EXP	1594.54	1123.284	
persinfo	CTL	1054.56	356.840	0.010
	EXP	806.50	332.670	
assignmenttarget	CTL	407.74	388.301	0.533
	EXP	470.29	350.566	
conceptmap	CTL	452.30	370.487	0.245
	EXP	346.82	291.960	
paper	CTL	1771.04	1022.395	0.429
	EXP	2021.96	1290.749	
diary	CTL	2159.96	1245.122	0.041
	EXP	1557.11	857.935	
forum	CTL	416.89	494.254	.019 [a]
	EXP	913.00	951.263	
question	CTL	351.85	483.097	0.457
	EXP	264.71	374.496	

[a] Levene's test for Homogeneity of Variances gives 8.339 for "forum", which is significant (0.006),; this has been taken into account in the Significance value

Table 3- Times spent on system elemnts by treatment group

The results show three significant differences between the treatment groups:

- persinfo: The Control group spent significantly more time here
- diary: The Control group spent significantly more time here
- forum: The Experimental group spent significantly more time here

Regarding the possible confounding variables (see the beginning of Section 4.2.2), ability was not found to be significantly related to any of the differences in time spent. Gender, however, was related to one element (persinfo) as shown in Figure 17 and Table 4. Females spent significantly more time on this element than males.

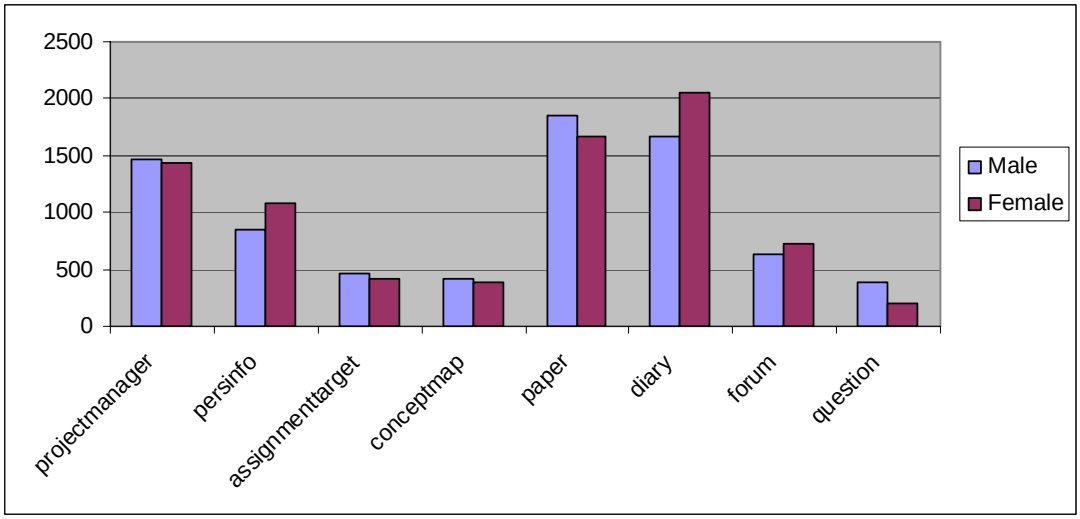


Figure 17 - Mean times per system element by Gender

		Mean	Std. Deviation	Significance
persinfo	Males	853.89	329.35	0.018
	Females	1080.96	332.12	

Table 4 - Mean times per system element by Gender

However, a Multivariate Analysis of Variance shows that Gender does not account for sufficient of the variance of persinfo to cause the significant difference between treatment groups (see Table 5).

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
persinfo	Type	.968	1	.968	5.081	.029
	Gender	.039	1	.039	.092	.763

Table 5 - MANOVA result for persinfo by Treatment Type and Gender

Feedback events (i.e. user clicks on a" smiley")

Students were able to click on a set of four "feedback" buttons in the form of "smileys":

- Happy
- Neutral
- Sad
- Confused

Not all feedback events were included in the analysis. Most importantly, clicks less than 20 seconds apart are discounted (there are many instances where a student clicks many times over a few seconds; this is assumed not to represent a changing state of mind). 20 seconds was chosen by examining the log files to be long enough to exclude multiple clicks. Also, clicks outside 8am – 3pm (scheduled classes) are discounted to remove tests.

The "Confused" smiley was removed from the Control group after week one (i.e. lessons on the 3rd and 4th of May), as the students expected it to give help in using the system (which it did not, and was not designed to do) and were unhappy that no help was given. Therefore, the analysis of "Confused" events only include such events from either group that occurred in week one.

Results for the whole pilot show that the Control group clicked more often on "Neutral", "Happy" and "Sad" feedbacks (see Figure 18), with both "Neutral" and "Sad" found to be significantly different at the 5% level (see Table 6). Table 6 also shows the same test for "Confused" (means, CTL= 1.19, EXP= 2.36) – this was not found to be significant; however, while this contains data from both treatment groups, it only relates to the first week of the pilot (as explained above).

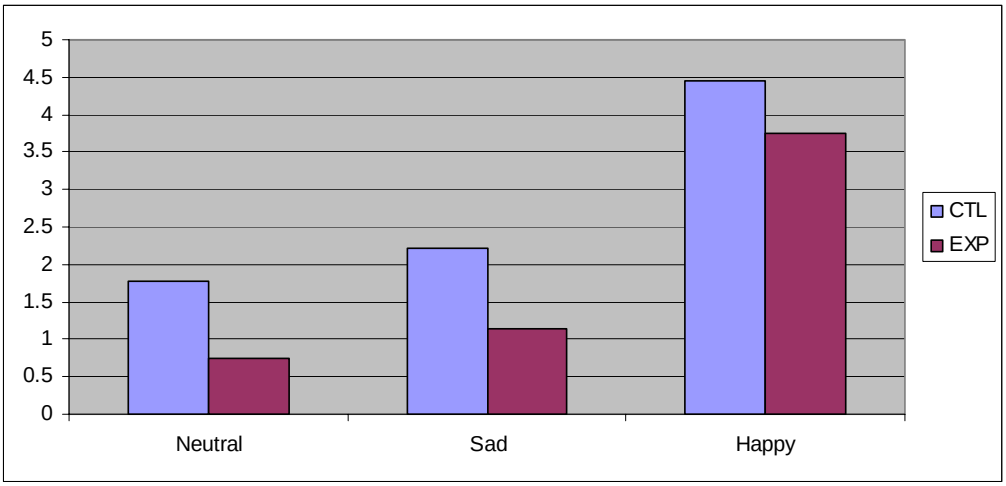


Figure 18 – Mean number of feedbacks for each student pair over whole of pilot, by treatment group

	Neutral	Sad	Happy	Confused (Week1)
Mann-Whitney U	237.000	252.500	349.500	320.000
Wilcoxon W	643.000	658.500	755.500	698.000
Z	-2.499	-2.183	-.484	-1.069
Asymp. Sig. (2-tailed)	0.012	0.029	0.629	0.285

Table 6 – Analysis of feedbacks by treatment group (Mann-Whitney tests)

Regarding the possible confounding variables (see the beginning of Section 4.2.2), none of the feedback ("Neutral", "Happy" and "Sad") variables tested showed a significant difference related to either gender[6] or ability.

However, while "Confused" (for week one) was not significantly different by gender, it was found to be significantly different for the Control group by Ability, as shown in Figure 19 and Table 7.

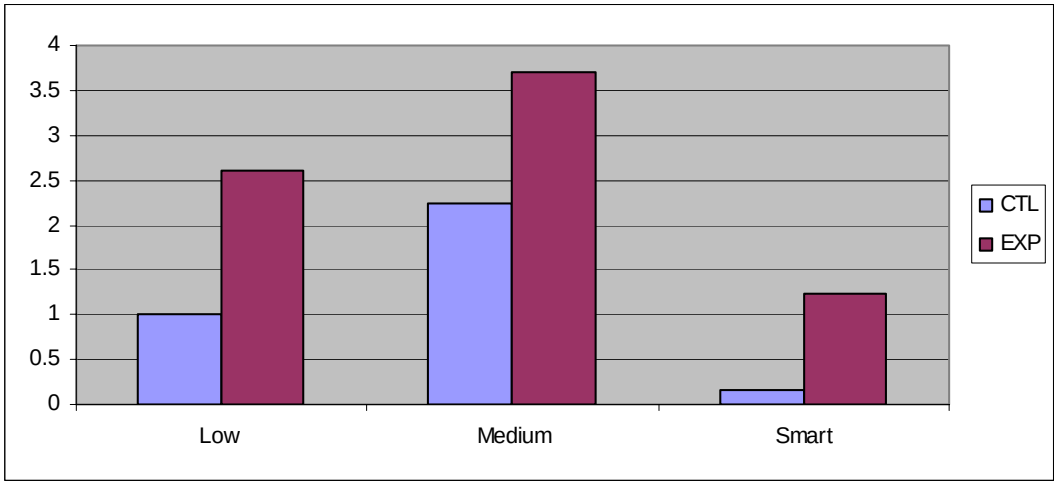


Figure 19 - "Confused" feedbacks during week one; mean values by treatment group and Ability

[6] Jonckheere-Terpstra test, looking for significance at the 0.05 level. This test was used instead of Kruskal-Wallis in order to take account of the Ability (Low→Medium→Smart) comprising an increasing scale

Confused, wk1	Control group	Experimental group
Number of Levels in Ability	3	3
N	27	28
Observed J-T Statistic	58.500	85.500
Mean J-T Statistic	108.000	122.500
Std. Deviation of J-T Statistic	19.184	21.633
Std. J-T Statistic	-2.580	-1.710
Asymp. Sig. (2-tailed)	.010	.087

Table 7 - Jonckheere-Terpstra tests, showing significance of "Confused" for week one by treatment group and Ability

Interventions

Reconstruction of interventions for the Control group

It was intended to reconstitute the interventions that the Control group would have received if they had been using the AtGentive-enhanced version of the software (as the Experimental group). That way, we could see how the interventions were related to events. For example, if the experimental group went to the mind map on average 36 seconds after a certain intervention, and the control group did the same thing 112 seconds after an intervention would have been given to them, then we could say the intervention had an effect.

In order to ensure that the reconstruction was correct, the following test was applied:

1. Note that the Experimental group log files have lines of "Sending intervention" and the Control group log files do not
2. Reconstruct the "Sending interventions" lines for both the Control group log files and the Experimental group (even though the Experimental group already have these lines)
3. Check that the reconstructed "Sending intervention" lines for the Experimental group are the same as the real lines for the Experimental group. If so, it provides an assurance that the reconstructed lines for the Control group are correct

None of the reconstitution methods tried passed this test. This was because an exact definition of the algorithms used by the reasoning module to initiate interventions for the experimental group is required. Unfortunately, although the reasoning module delivered appropriate interventions, in practice it did not exactly follow its original specification in a way that is very difficult to reverse-engineer. This made it impractical to reconstruct “Sending intervention” lines for the control group that would be directly comparable with the lines on the Experimental group’s log files. The intervention analysis therefore looks exclusively at interventions sent to the Experimental group.

Interventions received and Marks obtained

An analysis was made to look for any correlation between the number of interventions students received and the marks they were awarded. Two significant correlations were found (see Table 8), which are:

- a **positive** correlation between the number of meta-cognitive interventions and the number of questions to the expert
- a **negative** correlation between paper quality and both cognitive and meta-cognitive interventions

In other words, receiving more meta-cognitive interventions relate to asking more questions to the expert, and receiving less interventions of either type relates to a higher paper quality.

Also, this difference in interventions is not directly related to differences in either Ability or Gender.

		TOTAL COG interventions	TOTAL MC interventions
TOTAL COG interventions	Correlation Coefficient	1.000	.419(*)
	Sig. (2-tailed)	.	.027
	N	28	28
TOTAL MC interventions	Correlation Coefficient	.419(*)	1.000
	Sig. (2-tailed)	.027	.
	N	28	28
Us.Qns2Xp	Correlation Coefficient	.489(**)	.460(*)
	Sig. (2-tailed)	.008	.014
	N	28	28
Us.Status	Correlation Coefficient	-.227	.039
	Sig. (2-tailed)	.246	.843
	N	28	28
Us.#Partic	Correlation Coefficient	-.024	.203
	Sig. (2-tailed)	.903	.301
	N	28	28
Us.Intr.10	Correlation Coefficient	.066	.014
	Sig. (2-tailed)	.740	.945
	N	28	28
Us.GdGoal	Correlation Coefficient	-.312	-.278
	Sig. (2-tailed)	.106	.151
	N	28	28
Us.CMap	Correlation Coefficient	-.124	.005
	Sig. (2-tailed)	.528	.981
	N	28	28
Us.PPara	Correlation Coefficient	-.162	-.197
	Sig. (2-tailed)	.411	.314
	N	28	28
Us.PQual	Correlation Coefficient	-.388(*)	-.318
	Sig. (2-tailed)	.042	.099
	N	28	28

Table 8 - Correlations between Interventions by type and marked data (Expe. group only)

Additional Analysis

The log data was additionally analysed using a visualisation technique. While no further significant results were found using this method it adds to the comprehensiveness of the analysis as the technique has the ability to pick out unexpected effects. See D5.4 – "AtGentSchool and AtGentNet Workshops – Experiences from the Pilots" for further details.

4.2.2.3. Students' Pre- and Post-pilot knowledge tests

The purpose of the students' pre-test is to ensure that there is no significant difference between the Control and Experimental groups' knowledge of New Zealand prior to the trial. The post-test was to investigate whether a significant difference in learning between the groups had occurred during the pilot.

The students' pre-test contained 17 true / false questions, such as "The favourite sport in New Zealand is rugby" (true) – see Appendix 6 for a complete list. The students' post-test contained 15 true / false questions, such as "Moa is a big animal resembling a tiger" (see also Appendix 6).

The tests were self-administered (with the teacher's assistance where necessary) during class time using the same computer as that used for AtGentSchool. While AtGentSchool, and the Attitude survey questionnaires, were completed as a pair (by both students collaboratively), the pre- and post-tests were completed individually, one student then the other. This improved the accuracy of the tests.

For each of the pre- and post-test data sets, a Chi squared test was performed to compare the expected frequencies of True/False answers with the observed frequencies, per question and between the Control and Experimental groups. The results of the pre-test are shown in Table 9, and the post-test in Table 10.

Question number	Sig.	Pre-test Question
1	0.692	The favourite sports in New Zealand is rugby (true)
2	0.073 [a]	There are three main islands in New Zealand (false)
3	0.255 [b]	People speak German in New Zealand (false)
4	0.211	The Queen of England is also Queen of New Zealand (true)
5	0.266 [b]	Kiwi is a lizard on NZ (false)
6	0.196	The closest country to New Zealand is Australia (true)
7	0.139	The capital of New Zealand is Wellington (true)
8	0.043 [a]	When it's winter in Czech Republic, it's winter in NZ as well (false)
9	0.546	The original inhabitants of NZ are the Maori people (true)
10	0.574	There are fjords on NZ as in Norway (true)
11	0.055 [a]	There are no volcanoes on New Zealand (false)

12	0.111 [b]	The Maori people give kisses with their noses (true)
13	0.081 [a]	New Zealand is connected to Australia by a tunnel (false)
14	0.710 [b]	There is almost no nature on New Zealand (false)
15	0.123 [a]	There are no whales or dolphins around New Zealand (false)
16	0.125 [a]	The Kiwi eats only kiwi (false)
17	1.000 [a]	People commemorate the execution of Jan Hus in NZ (true)

[a] Fisher's Exact Test is quoted for this statistic as one or more expected frequencies is <=10
[b] Yates Continuity Correction is quoted for this statistic as one or more expected frequencies is <5

Table 9 - Results of Chi Squared test for students' pre-pilot knowledge test

Question number	Sig.	Post-test Question
1	0.723 [b]	New Zealand has two main islands (true)
2	0.528	Moa is a big animal resembling a tiger (false)
3	0.756 [a]	The first settlers were people of Greenland (false)
4	10.000 [a]	Mt. Cook is the highest mountain on New Zealand (true)
5	0.175 [b]	Kiwi-national bird of New Zealand (true)
6	0.772 [a]	You can go by car from South to West Island (false)
7	0.065 [a]	The biggest pest of New Zealand is possum (true)
8	1.000 [b]	Original inhabitants were English (false)
9	0.022	Most people live on South Island (false)
10	0.798 [a]	Maori have their own language (true)
11	0.944 [b]	The capital of New Zealand is Auckland (false)
12	0.284 [b]	The closest continent is Australia (true)
13	0.142 a	The expert travelled around New Zealand by bike (false)
14	0.036 a	The highest building in the Southern hemisphere is in NZ (true)
15	0.374 b	Mt. Taranaki is also a volcano (true)

[a] Fisher's Exact Test is quoted for this statistic as one or more expected frequencies is <=10
[b] Yates Continuity Correction is quoted for this statistic as one or more expected frequencies is <5

Table 10 - Results of Chi Squared test for students' post-pilot knowledge test

16 of the 17 pre-test questions, and 13 of the 15 post-test questions were found **not** to have answers that differed significantly in correctness between the Control and Experimental groups.

For the pre-test, question eight (“When it's winter in Czech Republic, it's winter in NZ as well”) gave a significant result, with the Experimental group scoring better than the Control group.

For the post-test, questions nine (“Most people live on South Island “) and 14 (“The highest building in the Southern hemisphere is in NZ”) gave significant results. The Experimental group scored better than the Control group for question nine, whereas the Control group scored better than the Experimental group for question 14.

4.2.2.4. Student Attitude Survey Questionnaires

Students were asked to complete questionnaires twice during the pilot – mid-way and towards the end (see Section 3.6.1). These questionnaires contained rating scales to assess the participants’ emotive response to AtGentSchool.

The statements that students were asked to rate are as follows:

1. Honza, the agent, helped me a lot
2. The software does exactly what I want
3. The software does not do anything that I want
4. Honza looks great
5. Honza is really friendly
6. Honza is very helpful
7. Honza is very annoying
8. I think Honza likes me a lot
9. When I use the keyboard or mouse, the software does something in response straight away
10. I understand everything the software tells me I should do
11. I like the look of the software
12. I know what I’m doing when I use the software
13. I feel in control of the software
14. The instructions on the screen are really helpful
15. I understood what the teacher told me to do
16. I could quickly get hold of the teacher to ask questions (translated as “I have no problem to ask teacher a question”)
17. The teacher knew all about the software
18. I really enjoyed the lesson
19. How much time did you spend learning how to use AtGentSchool, compared to actually using it?
20. We decided to divide the work equally
21. We shared equally the typing on the computer
22. We shared equally deciding what to type on the computer

Almost twice as many student pairs in the Control group completed the questionnaire twice compared to the Experimental group (22 compared to 14). The date at which students completed these questionnaires varied considerably, both between and within classes. Since attitudes are affected by continued use of the system, the date each student pair completes each questionnaire will be likely to affect the result. Comparisons between student groups are more likely to be affected than comparison within student groups. Therefore a repeated measures approach was chosen.

Students in the Control and Experimental groups were analysed separately. A Wilcoxon Signed Ranks Test was used to assess whether a significant change in attitude had taken place between the first and second questionnaire. Significant results (at the 5% level) are reported below:

Control Group	Mean (Q1)	Std. Dev (Q1)	Mean (Q2)	Std. Dev (Q2)	Z	Signif-icance
1.Honza helped me a lot	4.05	.653	4.09	1.109		not sig.
2.Software does what I want	3.05	1.253	2.73	.985		not sig.
3.Software does nothing I want	3.27	1.202	3.27	1.032		not sig.
4.Honza looks great	2.50	1.102	3.05	1.495	-1.999	.046 (+)
5.Honza really friendly	2.50	1.012	2.45	1.262		not sig.
6.Honza very helpful	3.86	.941	4.23	1.066		not sig.
7.Honza very annoying	2.95	1.174	3.09	1.477		not sig.
8.I think Honza likes me a lot	3.32	1.129	3.32	1.427		not sig.
9.Immediate response	3.32	1.041	3.45	1.371		not sig.
10.Understand what SW tells me	4.18	.958	4.05	1.090		not sig.
11.I like the look of the software	2.32	.780	2.27	.985		not sig.
12.I know what doing with SW	2.45	1.101	2.55	1.011		not sig.
13.I feel in control of the software	3.14	1.207	2.59	1.098		not sig.
14.Screen instructions helpful	2.41	.854	2.64	1.177		not sig.
15.Understood teacher's instruct's	2.68	1.359	2.36	1.255		not sig.
16.OK to ask teacher a question	2.32	1.086	2.05	1.133		not sig.
17.Teacher knew all about SW	3.05	1.090	2.91	.921		not sig.
18.I really enjoyed the lesson	3.05	1.290	2.91	1.231		not sig.
19.Time learning AGS vs using	2.05	.653	2.23	.612		not sig.
20 Divided the work equally	2.59	.959	2.23	.752		not sig.
21.Shared typing equally	2.41	1.054	2.27	1.077		not sig.
22.Shared deciding/type equally	2.50	1.144	2.45	1.335		not sig.

Table 11 - Wilcoxon Signed Ranks Test comparing first and second questionnaires for the Control group

Experimental Group	Mean (Q1)	Std. Dev (Q1)	Mean (Q2)	Std. Dev (Q2)	Z	Signif-icance
1.Honza helped me a lot	2.71	.914	3.43	1.284		not sig.
2.Software does what I want	2.43	.756	3.29	.994	-2.070	.038 (+)
3.Software does nothing I want	4.00	.784	3.14	1.167	-2.235	.025 (-)
4.Honza looks great	2.71	1.383	3.00	1.414		not sig.
5.Honza really friendly	2.14	1.027	3.14	1.351	-2.038	.042 (+)
6.Honza very helpful	2.50	1.019	3.36	1.151	-2.165	.030 (+)
7.Honza very annoying	3.64	.929	2.50	1.225	-2.073	.038 (-)
8.I think Honza likes me a lot	3.07	.997	3.64	1.008		not sig.
9.Immediate response	3.07	1.141	3.93	.997	-1.996	.046 (+)
10.Understand what SW tells me	2.36	.929	3.14	1.292		not sig.
11.I like the look of the software	1.79	.699	2.86	1.406	-2.223	.026 (+)
12.I know what doing with SW	2.07	.475	2.86	1.292		not sig.
13.I feel in control of the software	2.71	.914	2.86	1.231		not sig.
14.Screen instructions helpful	2.57	.938	3.43	.852	-1.997	.046 (+)
15.Understood teacher's instruct's	1.57	.756	1.86	.949		not sig.
16.OK to ask teacher a question	1.43	.646	1.64	1.151		not sig.
17.Teacher knew all about SW	2.36	.929	2.71	.726		not sig.
18.I really enjoyed the lesson	2.14	1.512	3.07	1.492	-2.214	.027 (+)
19.Time learning AGS vs using	2.50	.650	2.14	.535		not sig.
20 Divided the work equally	1.79	.802	2.36	1.336		not sig.
21.Shared typing equally	2.00	1.359	1.86	1.099		not sig.
22.Shared deciding/type equally	1.86	1.231	2.07	1.141		not sig.

Table 12 - Wilcoxon Signed Ranks Test comparing first and second questionnaires for the Experimental group

As the results show, there was virtually no change in the Control group between the two questionnaires. The only significant change was an improvement in the students' view of the agent's appearance ("Honza looks great").

For the Experimental group, however, nine of the same measured attitudes changed between questionnaires:

2. The software does exactly what I want
3. The software does not do anything that I want
5. Honza is really friendly
6. Honza is very helpful
7. Honza is very annoying
9. When I use the keyboard or mouse, the software does something in response straight away
11. I like the look of the software
14. The instructions on the screen are really helpful
18. I really enjoyed the lesson

All changes were effectively positive. That is, all positive statements (such as “Honza is really friendly”) increase between the first and second questionnaires, whereas negative statements decrease (such as “Honza is very annoying”).

4.2.3 Analysis

4.2.3.1. Summary of significant results

Table 13 shows the differences found between the Experimental and Control groups (in terms of the effects on the Experimental group).

Source	**Significant difference observed in the Experimental group**
Marks	Asked more questions of the expert
Marks	Wrote a better quality paper
Time on screens	LESS time describing themselves
Time on screens	LESS time looking at the expert’s diary
Time on screens	MORE time in the forum
Feedbacks	Less clicks on “Neutral”, and “Sad”
Feedbacks	Similar rate of “Confused” clicks (for week one) by Ability (In Control group, “Smart” students clicked less on “Confused”.)
Interventions	a POSITIVE correlation between the number of meta-cognitive interventions and the number of questions to the expert
Interventions	a NEGATIVE correlation between paper quality and both cognitive and meta-cognitive interventions
Pre-test	Scored BETTER on one question
Post-test	Scored BETTER on one question and LESS WELL on one question
Post-test	Less students completed the post-test questionnaire
Attitude change – improvement over time (not compared to the Control group)	2. The software does exactly what I want (+) 3. The software does not do anything that I want (-) 5. Honza is really friendly (+) 6. Honza is very helpful (+) 7. Honza is very annoying (-) 9. When I use the keyboard or mouse, the software does something in response straight away (+) 11. I like the look of the software (+) 14. The instructions on the screen are really helpful (+) 18. I really enjoyed the lesson (+)

Table 13 - Summary of effects seen in the EXPERIMENTAL group (in comparison to the Control)

These effects are now arranged in terms of the Key Indicators of Performance, Attention, Satisfaction, Learning and Collaboration (see section 3.2).

Key Indicator	Significant difference observed in the Experimental group
Performance	Wrote a better quality paper
Satisfaction	Less clicks on “Neutral”, and “Sad”
Satisfaction	Similar rate of “Confused” clicks (for week one) by Ability (In Control group, “Smart” students clicked less on “Confused”.)
Satisfaction	Attitude change – improvement: 2. The software does exactly what I want (+) 3. The software does not do anything that I want (-) 5. Honza is really friendly (+) 6. Honza is very helpful (+) 7. Honza is very annoying (-) 9. When I use the keyboard or mouse, the software does something in response straight away (+) 11. I like the look of the software (+) 14. The instructions on the screen are really helpful (+) 18. I really enjoyed the lesson (+)
Learning	Scored BETTER on one pre-test question Scored BETTER on one post-test question and LESS WELL on one post-test question
Collaboration	Asked more questions of the expert
-	LESS time describing themselves
-	LESS time looking at the expert’s diary
-	MORE time in the forum
-	a POSITIVE correlation between the number of meta-cognitive interventions and the number of questions to the expert
-	a NEGATIVE correlation between paper quality and both cognitive and meta-cognitive interventions
-	Fewer students completed the post-test questionnaire

Table 14 - Summary of effects seen in the EXPERIMENTAL group (in comparison to the Control) by Key Indicator

4.2.3.2. Analysis

The general hypotheses for these tests are:

- **H0:** (Null) There is no difference between the Control and Experimental groups
- **H1:** (Experimental) There is a significant difference between the Control and Experimental groups

We begin by assuming that H0 describes the results. The object of the statistical tests is to demonstrate beyond reasonable doubt that H1 is, in fact, the case, and we should reject H0 in favour of H1 – that we should accept that there is a significant difference between the groups.

It is important to note that a non-significant result merely allows us to continue assuming that H0 describes the results (Chalmers, 1994). A non-significant result does not demonstrate that this (no significant difference) is the case; it only allows us to maintain our unproven assumption. A significant result, however, does allow us to say that we have evidence for a hypothesis, and that this hypothesis is H1 – that there is a significant difference between the groups.

By this mechanism, one significant result carries greater importance than a number of non-significant results. It is with this in mind that we now examine the results in detail.

Performance

Paper quality (low / medium / good) was significantly better overall for the experimental group. The paper is one of the last assessed items that the students complete. They begin with writing an introduction and stating their goal, then create the concept map and asking questions of the expert. It is possible that effects of being in the experimental group take time to be visible. This would explain the finding of a significant difference only towards the end, in the quality of their written paper.

Satisfaction

A large number of significant results were found for satisfaction, all showing that the Experimental group were more satisfied with their experience than the Control group. Firstly, at the start of the project, the Control group clicked more on the “Neutral”, and “Sad” smileys, suggesting that they were less happy than the group who had an active agent. Also, it is interesting to note that within the Control group the “Smart” students clicked less on “Confused” than the “Medium” and “Low” ability students. This difference was not found in the Experimental group, suggesting that the active agent specifically supported the less able students.

Examining changes in attitude during the pilot, the only significant change found within the Control group was an improvement in the students’ view of the agent’s appearance (“Honza looks great”). This may well be that they “got used” to the agent over time. For the Experimental group, however, a large number of changes were found.

Firstly, we propose discounting the improvement of “When I use the keyboard or mouse, the software does something in response straight away”. There were a number of technical difficulties in installing and setting up the software at the start of the pilot. Many of these were solved after a period of time. It is possible that an improvement in response time may be as a result of these technical changes. Further, changes of this nature may affect the experimental group differently to the Control group, since the software was not the same for each group. Therefore, an improvement in the students’ perception of response time may be due to an actual change.

“The software does exactly what I want” and “I like the look of the software” improved over time (along with a decrease in the oppositely phrased statement). This suggests that the students’ expectations were more in line with the reality of the software towards the end of the pilot than at the start. The Control group did not experience this change, suggesting that the active agent may have assisted students in understanding the software.

“The instructions on the screen are really helpful” improved over time for the Experimental group only. One possibility is that as the students used more parts of the software (more screens) they encountered screens with better instructions. However, this was not something noticed during the heuristic evaluation. We propose an alternative explanation. We suggest that the students conflated the agent’s instructions with those that were part of the screen itself. The improvement would then relate to use of an active agent over time, in a similar way to that described in the previous paragraph.

“Honza is really helpful” (and Friendly, and Less annoying) improved over time for the Experimental group only. This suggests that interacting with an active agent over time was a positive experience. Conversely, the Control group – who interacted much less with a relatively passive agent – did not experience this change.

Finally, perhaps the clearest statement “I really enjoyed the lesson” improved over time for the Experimental group only. The first week or two was a difficult time for both groups, as they were not used to this form of software (where one takes instructions from the screen, not the teacher). This finding suggests that provision of an active agent helped students recover from these difficulties, making later lessons more enjoyable.

Overall, then, we accept the Experimental hypothesis, that the Experimental group were more satisfied than the Control group.

Learning

When evaluating educational technologies, “learning” is very difficult to measure, as much of the intended learning is not specifically factual, but rather at a general or a meta-cognitive level. For example, in comparing two countries, students may learn about the meaning of “compare” and of “contrast”, about how to assemble information, about how to communicate with an expert or with peers, and so forth. This learning is useful, indeed, potentially far more useful than learning facts about New Zealand. This more general learning may, however, not be demonstrated by the students during the pilot. For example, a student may decide on a goal at the start of the pilot that is not suitable (and therefore marked as a “bad” goal), but during the pilot may reflect on the goal and realise their mistake. Although they will have learned how to choose a better goal next time, this learning is not visible during the relatively short pilot study.

In order to assess learning within the confines of the pilot, a number of factual questions were posed to the students. The results were not clear.

The Experimental group scored BETTER on one pre-test question AND on one post-test question. They also scored LESS WELL on one post-test question. This does not provide evidence for rejecting H0 – the null hypothesis – and thus suggests that no evidence for an improvement in learning was found.

Conversely, we must take into account that the hypotheses are two-tailed; that is, that they include the possibility that the Control group learned more than the Experimental group. If the post-test had found only the significant result of the Control group scoring more than the Experimental on one question then this conclusion may be indicated. However, the post-test in fact has one result for each possible direction, that is, in one question the Control group learned more than the Experimental group and in one question the Experimental group learned more than the Control group. The most likely explanation here is that the post-test is not a sufficiently accurate instrument to draw a conclusion from this result, which should be attributed to “noise” (extraneous factors). For example, students in each group sat together. This may have resulted in individual knowledge items that one student may say out-loud being overheard by others in that group, leading to greater “knowledge” for that question without learning that item through use of the software.

Overall, then, we retain the Null hypothesis, that there is no significant difference in learning between the Experimental and Control groups.

Finally, as described above, more subtle learning would be expected to have taken place. In particular, since the Experimental group wrote papers that were, overall, of significantly better quality than those of the Control group, it would be safe to speculate that the Experimental group would be able to write better papers in future assignments, and thus that they have learned, overall, how to write better papers.

Collaboration

The Experimental group asked more questions of the expert than the Control group. This suggests greater interest, desire or ability to collaborate with the expert in their learning.

We accept, then, the Experimental hypothesis, that the Experimental group collaborated more than the Control group.

It is interesting to speculate on collaboration within the student pairs. Ideally, we would have liked to audio-record a number of student pairs, transcribe their speech, translate into English and analyse the dialogue. This would have allowed us to compare intra-pair collaboration across groups. However, this would have been a very resource-intensive analysis, and was not of sufficient priority to justify the large diversion of resources necessary. The project partners are aware of the potential for this form of analysis, and intend to incorporate it into follow-on investigations.

Attention

It did not prove practical to install eye-tracking equipment in schools, or to video-record and transcribe the interaction between individual student pairs and their computer. Therefore no systematic measurement of attention was possible. However, Attention is a primary Key Indicator, which is to say that it is assumed that the AtGentive modifications will affect the user's attention directly; the other Key Indicators are considered to be secondary, in that they are affected by changes in attention. For example, if directing the learner's attention to salient details were to improve results on a test. It is therefore not essential to measure attention directly in order to assess the effectiveness of attention support.

Additional pedagogical Analysis

The AtGentSchool pilot provides a rich composite of data which it was felt warranted further analysis. In addition, Table 14 lists a number of findings that do not directly relate to the Key Indicators. The data were therefore explored and discussed further, taking a pedagogical approach, with the aim of gaining a better understanding of the learning processes for the children who participated in the pilot study.

Since this analysis is complex and additional to the AtGentive remit, the full description appears in Appendix 14. In summary, the findings are discussed in terms of the effects that different types of system interventions had on the learning process. In particular, how self-regulation differed between the Control and Experimental groups, the support provided to the knowledge building process and the community effects on learning.

4.2.3.3. Conclusion

The evidence indicates that the AtGentive modifications to the Ontdeknet software generated improvements in three of the Key Indicators: Performance, Satisfaction and Collaboration. Students accepted the agent and performed better as a result of its assistance. This in turn suggests that the scenarios implemented were effective, and that the Conceptual Framework was successful in its generation of those scenarios.

Student satisfaction was better where the agent acted as assistant, suggesting that as well as being effective in promoting performance, assistance from the agent was liked and appreciated by the students. It is worthy of note that for those using the helpful agent (i.e. the Experimental group) the more able students were less confused during the earlier classes. This suggests that extra support for less able students is indicated for future use of such agents.

Students collaborated more with the expert when assisted by the agent. This is an important finding, as collaboration is a necessary skill in the workplace and one that is not so easily taught in the classroom situation.

The evidence does not suggest that improvements in Learning took place as a result of assistance by the agent (nor was any detrimental effect observed). However, improving the performance and satisfaction of students is very worthwhile. It reduces the load on the teacher, allowing them more time to assist the less able students. It may also be that learning gains would be apparent under other circumstances – for example with less motivated teachers / students.

4.3 Pilot study - AtGentNet

Data collected as part of the AtGentNet pilot, and available for analysis, is as follows:

- log files
- questionnaire responses
    - students feedback
- post-trial interviews

4.3.1 Results from Pilot

4.3.1.1. Log files

The log files used for the evaluation are text-format files (the platform has the ability to export the activities in various formats, such as text or Atom) with one line for each "event" (such as the user amending their profile, or reading a posting). There is one file per student. However, of the 27 participants, one student from each group did not use the ICDT platform, giving data from 12 students in the Experimental group and 13 students in the Control group.

It must be noted that this is a small quantity of data. A further problem is a lack of homogeneity within the data; often a small number of participants have used a function far more than others. However, the others tend to be disparate, making the removal of data as "outliers" highly problematic, since removal of one outlier usually invites the next value in turn to be considered in similar terms. The decision was therefore taken to use the data as obtained.

Logged events comprise Date, Time, Event type, Event details. The main events available are:

- User changes of screen
- Creating / Editing / Reading postings
- Searches

A typical log file would contain data of the form:

```
17/09/2007    09:40:34    visiting    place    'forum exchanges'
17/09/2007    09:40:37    visiting    place    'forum exchanges_exchanges'
17/09/2007    09:40:48    reading     message  'exchange on coaching'
```

OBU created an analysis program that reads in all the log data and picks out selected events, outputting totals for further analysis.

Number of logins per month

Firstly, the number of times participants logged in each month was calculated. This gives a broad view of how the system was used. Figure 20 shows logins for the months May to October, by treatment group (Control and Experimental) and as a total.

As described in Section 2.2.8, the pilot began with an introduction to AtGentNet for TRIM participants at meeting in Lidköping, Sweden on 24th May 2007. However, many of the participants were sent details of the platform prior to this date, allowing them to log in from mid-May onwards.

On the 4th of September 2007 participants were sent emails inviting them to take part in the Eagle Racing simulation (see Section 2.2.7). Therefore logins for September include any specifically for this. (The racing simulation was over by 10th October, but a discussion forum continued throughout October.)

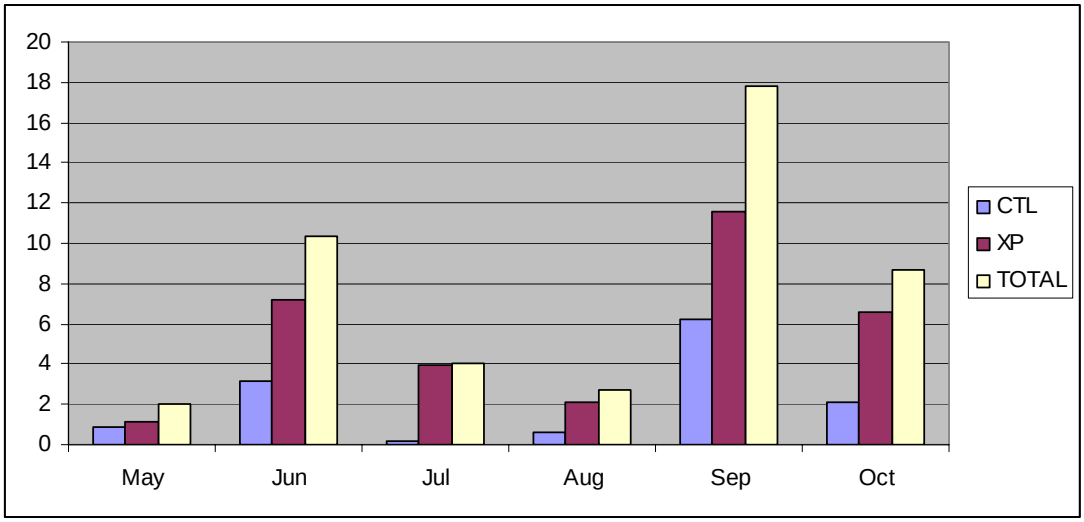


Figure 20 - Total number of logins per month, by treatment group and total

Number of events by type

Data were extracted from the log files under the following headings (in some cases, this meant amalgamating a number of log entry types):

Title	Meaning
En	Logged in (Entered System)
Lv	Logged out (Leave)
Rd Pr OTHER	Read someone else's profile (including any sub-pages)
Rd Ms	Read message (posting)
Vi Pl non-HP	Visited other (non-home) pages
Vi Pl HP-overview	Visited top-level home page
Vi Pl HP-personal	Visited Personal home page
Vi Pl HP-agents	Visited Agents home page

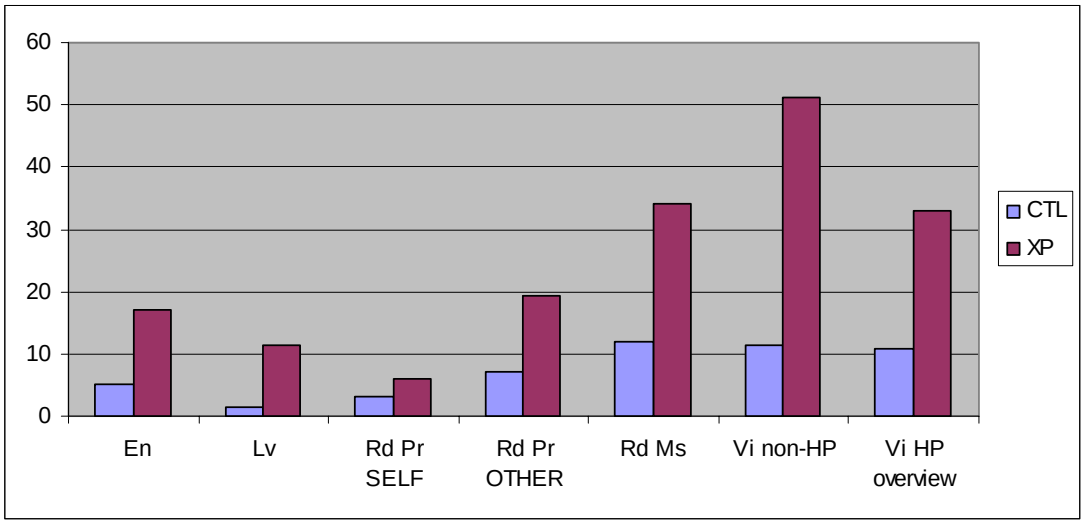
Rd Pr SELF	Read own profile (including any sub-pages)
Vi Pl HP	Alternative access to top-level home page
Vi Pl HP-news	Visited News home page
Vi Pl HP-invitations	Visited Invitations home page
Vi Pl HP-knowledge	Visited Knowledge home page
Vi Pl HP-community	Visited Community home page
Vi Pl HP-watch	Visited Watch home page
Cr Th	Created Thread
Cr Ms	Created Message
Ed Ms	Edited Message
Se query	Searches

Figure 21 - List of log file headings

The pattern of system use changes very noticeably between the initial period (May to July) and the Eagle Racing simulation (September onwards). Therefore, the two time periods have been analysed separately. The data are shown in Figure 22 (pre-Eagle Racing) and Figure 26 (during Eagle Racing).

Pre-Eagle Racing

In the pre-Eagle Racing period, all the measured values appear to differ between groups (see Figure 22), with the Experimental group consistently using the system more than the Control group.



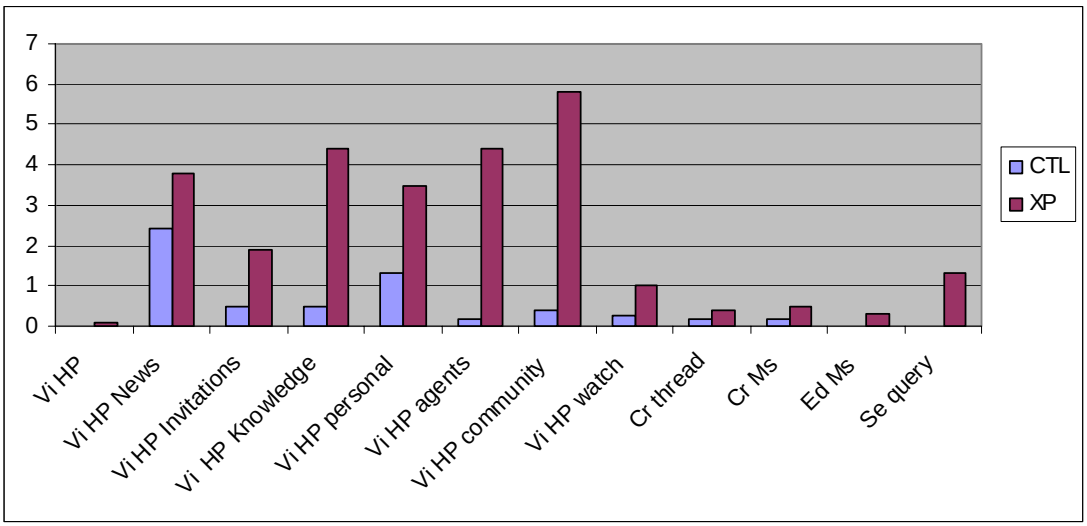


Figure 22 - Pre-Eagle Racing mean system use counts (split over two graphs to allow different scale values)

However, a closer examination of the data shows a more complex picture. Firstly, a Mann-Whitney test [7] shows that most of the differences are not statistically significant see Table 15). (Note that the graphs show mean total values in order to clearly demonstrate differences, whereas the statistical analyses use a ranking-based algorithm.) The two variables that are significantly different between treatment groups (at the 0.05 level) are “Vi place 'homepage - knowledge'” and “Vi place 'homepage - agents'”. (It should be noted that the agent homepage was not very feature-rich for the Control group and was essentially unused, whereas five out of the 12 Experimental group participants used the agent homepage.)

[7] Significance referred to is at the 0.05 level. Exact significance is quoted to compensate for the small sample size. 1-tailed test is quoted as visual inspection of the data shows the Experimental group has consistently greater system use

	En	Lv	Rd profile SELF	Rd profile OTHER	Rd message	Vi place	Vi place 'homepage'
Mann-Whitney U	65.500	77.000	66.000	63.000	65.500	53.000	71.500
Wilcoxon W	156.500	155.000	157.000	154.000	156.500	144.000	162.500
Z	-.682	-.058	-.667	-.819	-.683	-1.362	-1.041
Exact Sig. (1-tailed)	.256	.485	.260	.213	.255	.090	.480

	Vi place 'homepage - overview'	Vi place 'homepage - news'	Vi place 'homepage - invitations'	Vi place 'homepage - knowledge'	Vi place 'homepage - personal'	Vi place 'homepage - agents'
Mann-Whitney U	67.000	68.000	69.500	44.000	77.000	51.000
Wilcoxon W	158.000	159.000	160.500	135.000	168.000	142.000
Z	-.599	-.578	-.527	-2.005	-.059	-1.960
Exact Sig. (1-tailed)	.282	.288	.316	.025*	.491	.034*

	Vi place 'homepage - community'	Vi place 'homepage - watch'	Cr thread	Cr message	Ed message	Se query
Mann-Whitney U	56.500	58.500	73.500	73.500	71.500	65.000
Wilcoxon W	147.500	149.500	151.500	151.500	162.500	156.000
Z	-1.478	-1.347	-.434	-.434	-1.041	-1.502
Exact Sig. (1-tailed)	.082	.118	.531	.531	.480	.220

Table 15- Mann-Whitney U test results comparing Experimental and Control groups

Secondly, it is very noticeable that three participants in the experimental group used the system far more than the other participants, as shown in Figure 23 .

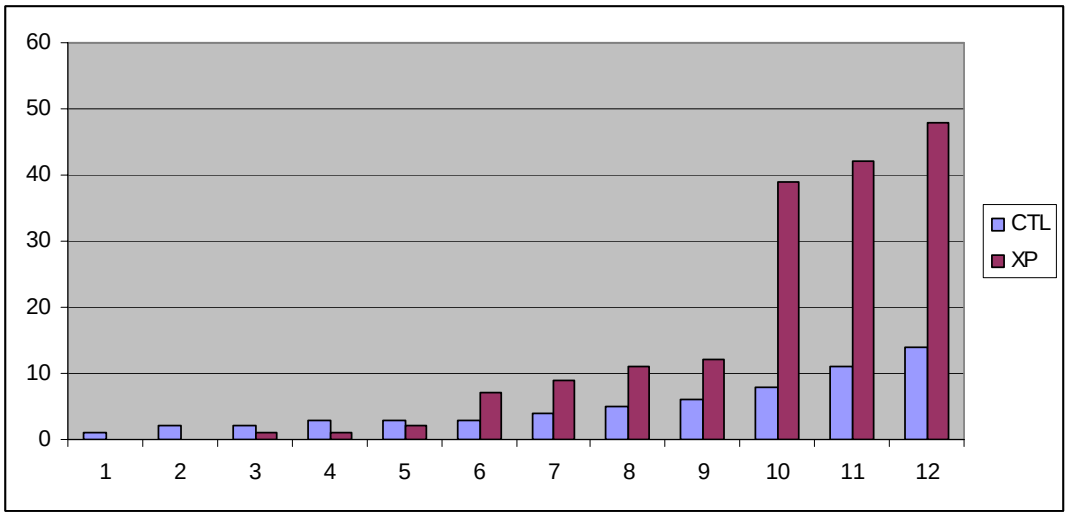


Figure 23 - Number of logins by user, sorted by Frequency [8]

However, in terms of system usage during this period, once logged in, the data show a more general variation. Figure 24 shows accesses to other participants' profiles (taken as an indicator of social interest). In this case there is less of a pronounced difference between the top three participants (the same three people as in the other two graphs of this nature) and the other participants.

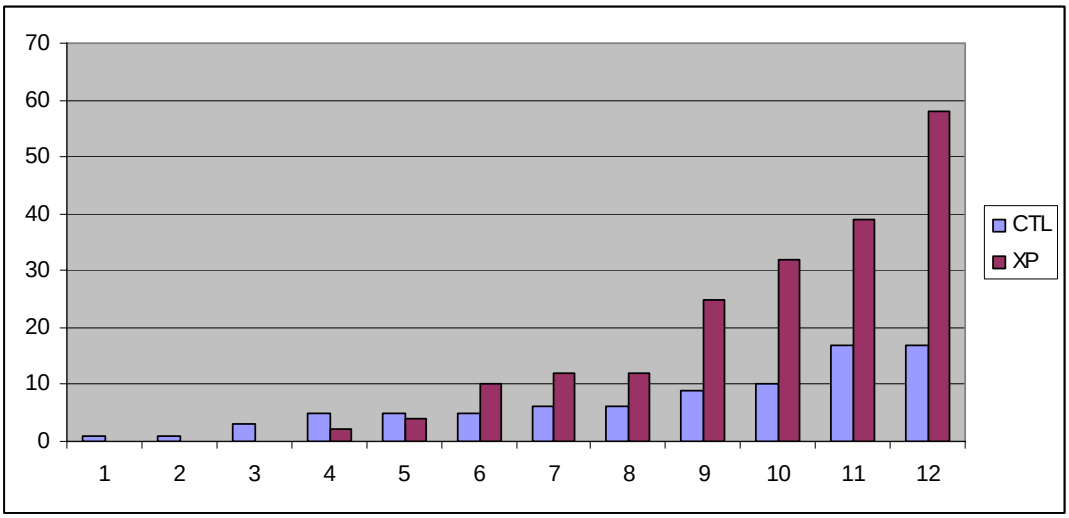


Figure 24 - Number of accesses to others' profiles (social interest) by user, sorted by frequency

Similarly, Figure 25 shows accesses to postings on the platform (e.g. to read about a meeting, download a document etc.). This too shows more of a trend than three outliers.

[8] Note that the Control group comprises 13 participants and the Experimental group 12. For the purposes of this, and the following two figures, one participant form the Control group, who did not log in at all during this period, is not shown.

Given the non-significance of most measured variables (Table 15) and the shape of the profile-access (Figure 24) and read (Figure 25) graphs, it was decided to treat the data from the three participants who logged in the most as valid, rather than outliers.

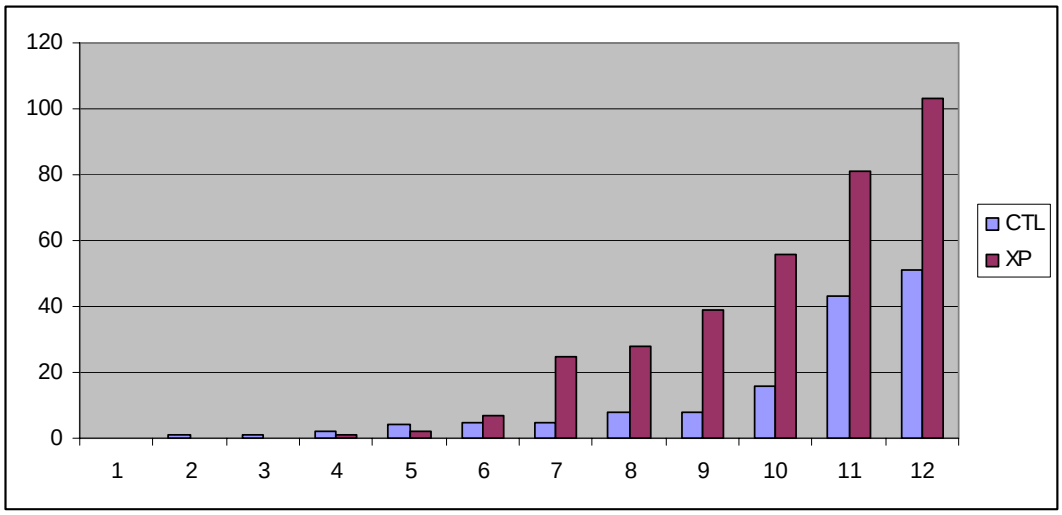
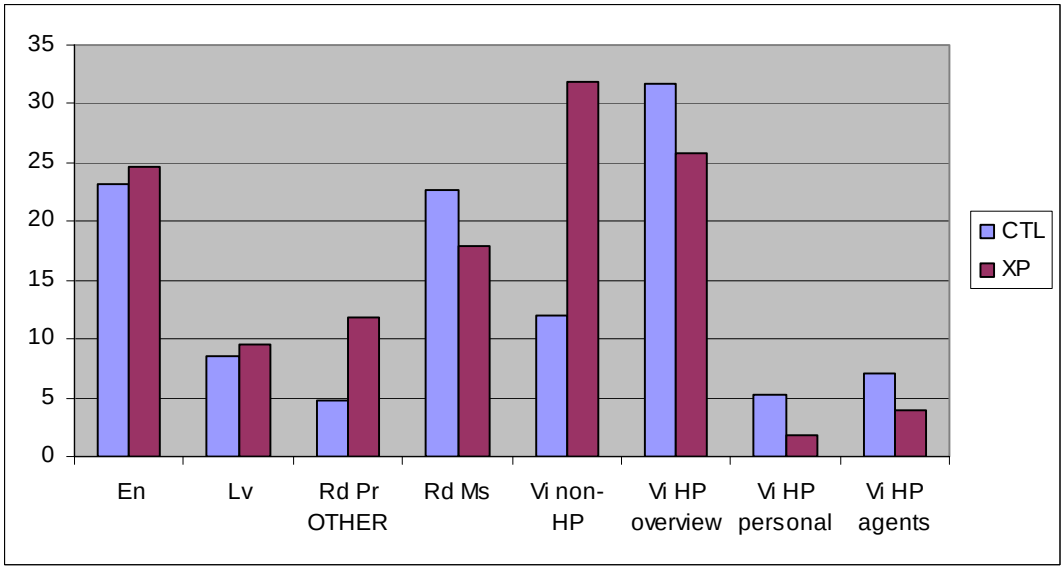


Figure 25 - Number of document reads by user, sorted by frequency

During Eagle Racing

Log data from the duration of the Eagle Racing simulation shows a different picture to the earlier time (see Figure 26), with system use much more evenly distributed.



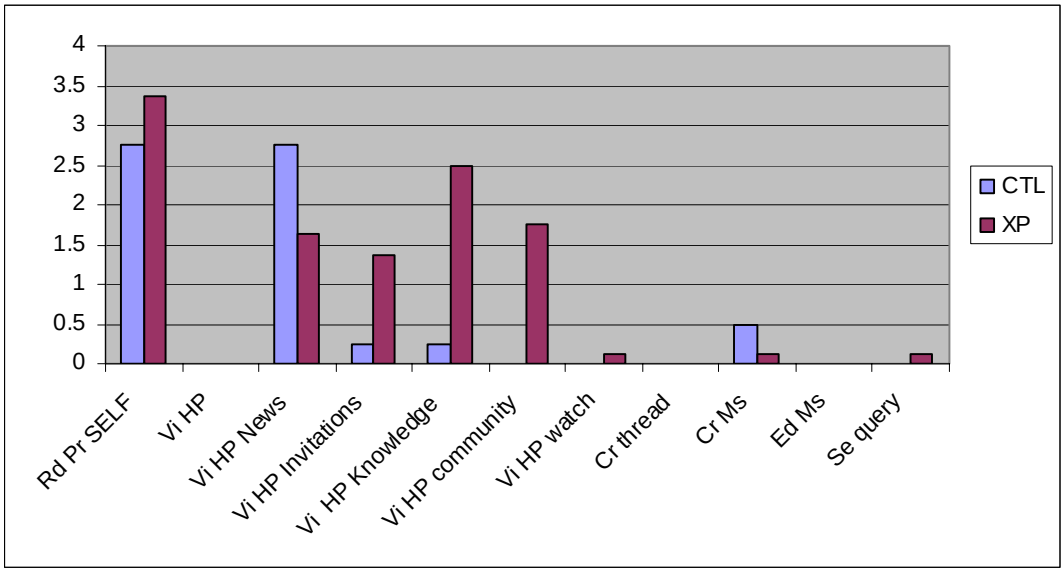


Figure 26 – Mean system counts use during Eagle Racing

A Mann-Whitney test [9] in this case shows no statistically significant differences (see Table 16).

	En	Lv	Rd profile SELF	Rd profile OTHER	Rd message	Vi place	Vi place 'homepage'
Mann-Whitney U	16.000	15.000	12.500	9.000	11.500	8.500	16.000
Wilcoxon W	52.000	25.000	22.500	19.000	47.500	18.500	52.000
Z	.000	-.176	-.606	-1.193	-.766	-1.278	.000
Exact Sig. (2-tailed)	1.000	.939	.588	.263	.489	.228	1.000

	Vi place 'homepage - overview'	Vi place 'homepage - news'	Vi place 'homepage - invitations'	Vi place 'homepage - knowledge'	Vi place 'homepage - personal'	Vi place 'homepage - agents'
Mann-Whitney U	12.500	13.000	9.000	8.500	7.500	11.000
Wilcoxon W	22.500	49.000	19.000	18.500	43.500	21.000
Z	-.595	-.523	-1.295	-1.373	-1.483	-.882
Exact Sig. (2-tailed)	.598	.679	.271	.267	.180	.410

[9] Significance referred to is at the 0.05 level. Exact significance is quoted to compensate for the small sample size. 2-tailed test is quoted as visual inspection of the data does not give a clear direction of differences

	Vi place 'homepage - community'	Vi place 'homepage - watch'	Cr thread	Cr message	Ed message	Se query
Mann-Whitney U	8.000	14.000	16.000	13.500	16.000	14.000
Wilcoxon W	18.000	24.000	52.000	49.500	52.000	24.000
Z	-1.621	-.707	.000	-.653	.000	-.707
Exact Sig. (2-tailed)	.208	1.000	1.000	.758	1.000	1.000

Table 16 - Mann-Whitney U test results showing non-significance for all variables

4.3.1.2. Students' questionnaire results

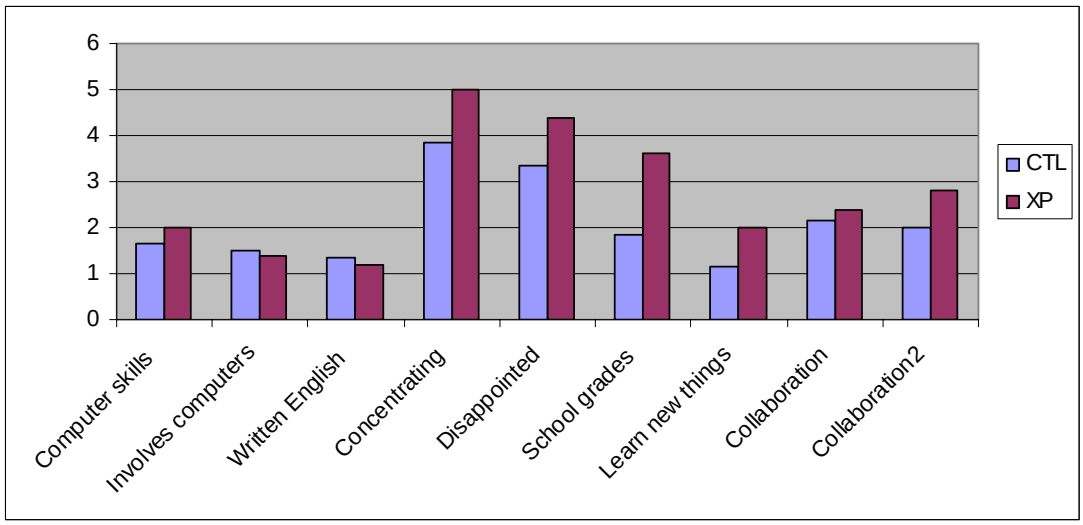
Questionnaires were sent to all participants mid-way through the pilot (on 20th September – two weeks after the start of the Eagle Racing simulation) and after the pilot (21st November). These questionnaires comprised mainly of statements about the ICDT platform, each answered by selection of a seven-point Likert-style scale:

> Agree completely / Agree / Agree a little / Neither Agree nor Disagree / Disagree a little / Disagree / Disagree completely

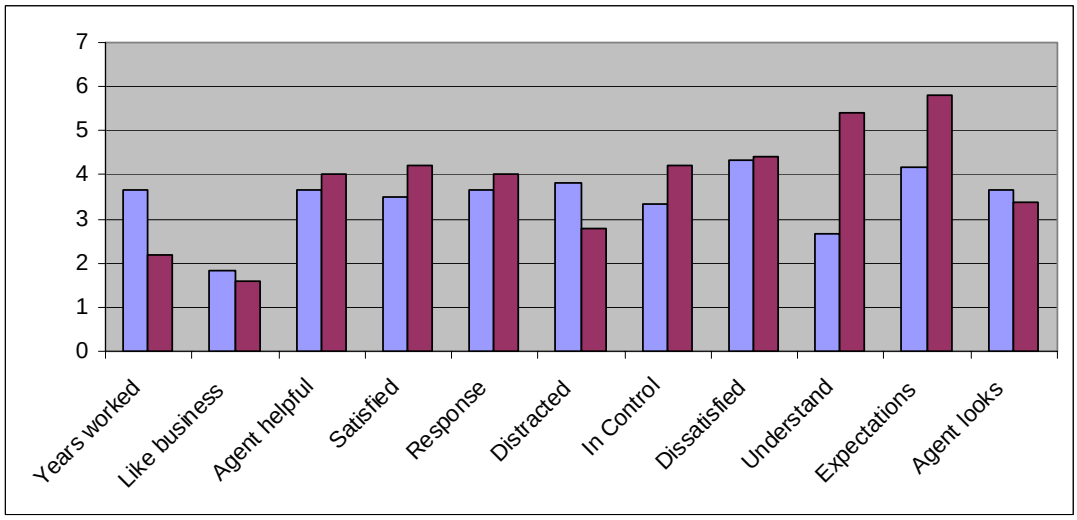
Eleven responses were received from the first questionnaire (six from the Control group and five from the Experimental). These are analysed below. The second questionnaire generated only two responses making it unfeasible to analyse quantitatively. It is thought that the business users of AtGentSchool are very busy with their regular work and once the TRIM course was over were unable to justify the time necessary to complete a second questionnaire.

The results of the first questionnaire are here split into three sections. For each section a graph is drawn showing the difference between treatment groups. Each graph is preceded by the relevant questionnaire statements in the sequence seen on the graph. (See also Appendix 3).

- Compared to my colleagues, my computer skills are (Better<-->Worse)
- I like it that my work involves computers
- I am good at understanding written English
- I have difficulty concentrating on one thing for some time
- I am often disappointed by products and services I have purchased
- At school I always received good grades
- I am quick to learn new things
- I like to collaborate whenever I can
- I like to collaborate when I can see it is in my interest



- How long have you worked in your current type of business role (years)?
- I like working in business
- Colette[10], the agent, is very helpful
- I am completely satisfied with the ICDT platform
- The ICDT platform responded quickly enough
- I am easily distracted when using the ICDT platform
- I feel in control when using the ICDT platform
- I am very dissatisfied with the ICDT platform
- I fully understand how to use the ICDT platform
- The ICDT platform fully met my expectations
- Colette (Atgentigirl), the agent, looks great



[10] Note that although the animated agent was referred to on the platform as “AtGentiGirl”, the participants had originally been introduced to “her” as “Colette”. Therefore, the questionnaire refers to “her” as “Colette, the agent”

- Colette is really friendly
- Colette is very helpful
- I think Colette likes me
- Colette is very annoying
- I understand all the headings on the ICDT platform
- I like the look of the ICDT platform
- If I make a mistake when using the ICDT platform it is easy to correct it
- Instructions on the screen about how to use the ICDT platform are really helpful
- I fully understood the description of the ICDT platform that was given at the Lidköping meeting
- I could quickly find someone appropriate to ask questions about the ICDT platform
- My questions were answered easily
- I really enjoy using the ICDT platform
- I am having to spend too much time learning how to use the ICDT platform

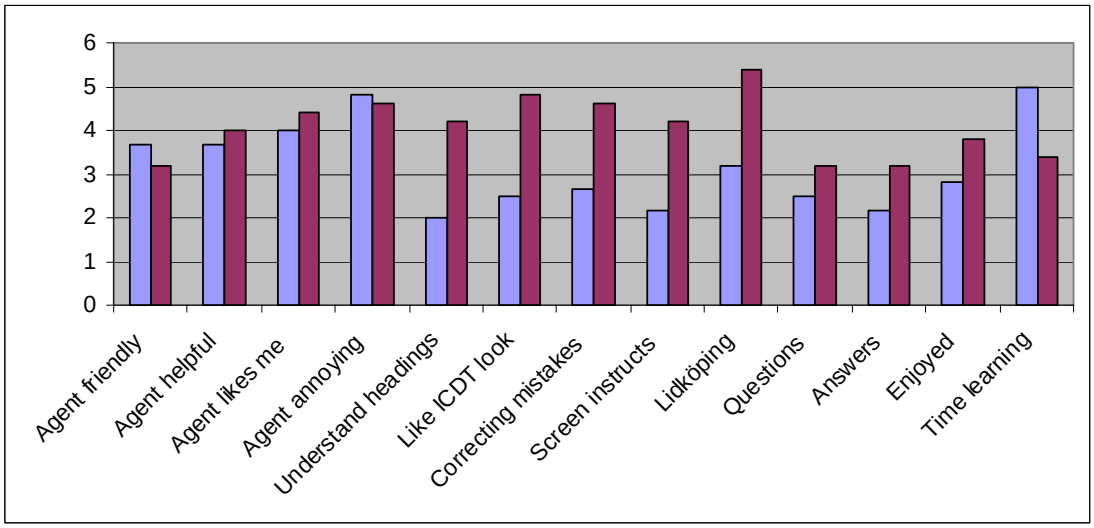


Figure 27 - Answers to AtGentNet questionnaire (this and preceding two figures)

A Mann-Whitney test was conducted to determine which results were statistically significant [11] (see Table 17). This shows four statements where the Experimental and Control groups significantly differ. In each case, the Experimental group gave a higher score of agreement than the Control. (For example, participants in the Experimental group were more in agreement that they always received good grades at school than the Control group.) The three statements that have a significant difference are:

- At school I always received good grades
- I fully understand how to use the ICDT platform
- I understand all the headings on the ICDT platform

[11] Significance referred to is at the 0.05 level. Exact significance is quoted to compensate for the small sample size (Control, n=6, Experimental, n=5). 2-tailed test is quoted as visual inspection of the data does not give a clear direction of differences

	At school I always received good grades	I fully understand how to use the ICDT platform	I understand all the headings on the ICDT platform
Mann-Whitney U	3.500	4.000	3.500
Wilcoxon W	24.500	25.000	24.500
Z	-2.180	-2.056	-2.213
Exact Sig. (2-tailed)	0.037	0.048	0.035

Table 17 - Mann-Whitney test results for AtGentNet questionnaire. To save space, only significant results are shown

4.3.1.3. Interviews

Telephone interviews were conducted with an opportunity sample of users (chosen as having shown interest in the course and thus assumed more likely to be willing to be interviewed). It was possible to conduct interviews with six students – four from the Experimental group and two from the Control group.

A semi-structured interview format was used. The starting point was the Key Indicators, with one question per indicator. The interviewers then used their judgement to encourage the interviewee to elaborate on any specific comments that may be of particular relevance. The initial questions were as follows:

- Key Indicator: Performance
  - How has the ICDT platform supported you in the ITM-concept?
- Key Indicator: Satisfaction
  - How satisfied are you with the ICDT platform?
  - Encourage them to describe the good and bad points of the system.
- Key Indicator: Learning (What have they learned?)
  - What would you say that the ICDT platform contributed to the ITM Course from the learning point of view?
- Key Indicator: Collaboration
  - How did you find the two collaboration initiatives from Albert Angehrn? (Reflections on the EIS simulation[12] (June 2007) and the Eagle Racing simulation[13].)
  - If they did not participate in the EIS-simulation then try to find out why not

Performance

Performance and satisfaction were the areas most mentioned by the interviewees. All interviewees gave ease of accessing documents as the main feature that boosted their

[12] A simulated six-month process of persuading managers in a corporation to adopt an Executive Information System

[13] A simulated set of business dilemmas in which the fictional Eagle Racing company attempts to find sponsors for their motor racing team (see Section 2.2.7)

performance. They appreciated that all information both about a seminar (location, travel details, etc.) and documents from the seminar was available in one place. Furthermore, documents are organised in a structured manner and are available immediately, regardless of the person's location. This facilitated both preparations before a seminar and follow-up study afterwards – without the need to carry home a large amount of paper. This was a very significant and much-mentioned advantage. No Experimental group-specific features were given as helpful for performance.

Satisfaction

Satisfaction

As with Performance, easy access to documents provided the greatest satisfaction. Both groups appreciated the efficiency and ease with which specific documents could be obtained using the platform. The Search facility was seen as an important feature for locating relevant people and documents.

An additional source of satisfaction was seen in the interviewees' ability to generate and maintain business and social connections with other participants. Both groups pointed to the profiles and photographs of lecturers, professors and participants as an important source of information. It allowed them to find out "who is who" and to contact them, even after a long time. Two interviewees from the Experimental group mentioned the chat window as a good way of seeing who was logged on and one mentioned liking the ability to tell who has read a document.

Two interviewees from each group liked the overall layout ("modern") and the use of "pictures".

Dissatisfaction

The predominant source of dissatisfaction was that there was too much information visible at one time. All interviewees (from both groups) mentioned this as a problem. Several interviewees thought that much of the information was intended for "administrators" and not participants like themselves.

Possibly related to this was that half the interviewees from each group considered that it was easy to "get lost" on the platform. They felt it was "hard to use", "hard to navigate", required "too many clicks" and that people would need telephone support to get started.

One interviewee (from the Control group) used alternative methods of communication where possible, such as email, "MSN" [14] and Facebook[15], as they considered these to be "much easier".

Finally, one interviewee (from the Experimental group) pointed out that they sometimes had difficulty to use a fast internet line, and that this was needed in order to use the ICDT platform.

Learning

In general, the ITM course does not intend that students will learn specific, testable, concepts, skills or knowledge. Rather, the course enables each student to improve their

[14] Windows Live Messenger™ from Microsoft®

[15] Facebook: www.facebook.com

company's ability to expert goods or services. Students learn whatever is necessary to permit this. Therefore, what is learned depends upon three things:

- How the individual decides the course may benefit their company
- What they have already learned in that area or areas
- The difference between existing learning and that necessary to achieve the benefits identified for their company

Thus, the learning appropriate per student is very individual to that person within their context. Furthermore, learning is not done for its own sake, but to facilitate other actions. It may well be that a student may not even be fully aware of what they have learned, only that they have achieved a new goal with assistance from the ITM course.

In fact, the interviewees were not forthcoming about details of specific learning that the ICDT platform had enabled. Instead, when asked about their learning they spoke mainly about the practical benefits of the platform, such as access to documents, as described under "Performance". Only two actual specific "learning" areas were volunteered (these were from Control group participants):

- Different cultures – the importance of respecting and understanding different cultures
- "Practical tools" – this was a general comment, not backed up with specifics

Collaboration

The ITM training takes place at three levels (see Figure 28). Firstly, there is a local tutor that each student meets for half a day per month. Then there is one are a small number of national seminars, each involving participants from one country, and one international seminar for all the participants. The purpose of this structure is to minimise costs for the students and course organisers by providing each aspect of the training as locally as possible, while providing access to lecturers of national and international importance. The ICDT platform acts as the background "common area", holding the disparate training elements together.

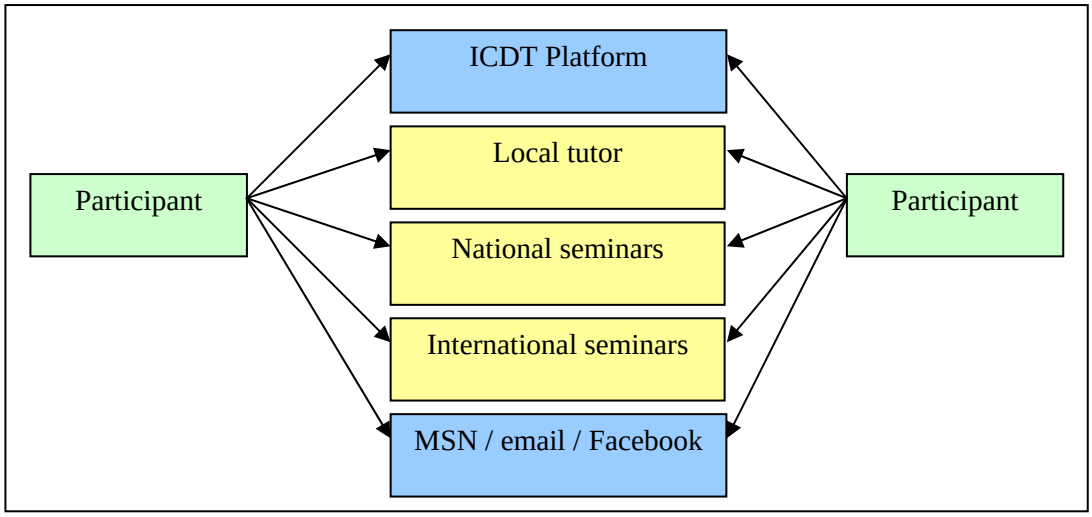


Figure 28 - Structure of the ITM training course

In relation to the ICDT platform, two forms of collaboration were mentioned by the students – collaboration with tutors and collaboration with other students.

After the seminars, students liked being able to get in touch with lecturers and other speakers easily via the platform. This allowed them to ask questions if needed. In addition, students appreciated being able to contact and collaborate with colleagues they had met at seminars. Overall, they considered it “an efficient way of taking part in a global classroom” and an interesting way of collaborate and learn.

As to specific collaborative ventures, the “Reflections on the EIS-simulation” that took place during June 2007 did not result in collaborative activity as anticipated. 11 students read Albert Angehrn’s posting (five out of 14 from the Control group and six out of 13 from the Experimental group). However, none of the students posted a reply (continued the collaboration). In the interview, most participants gave the same reason, which was that the collaboration request was issued during the summer, when they were either too busy or on holiday. Two interviewees mentioned that the collaboration request was also not sufficiently interesting.

This contrasts with the later Eagle Racing simulation, which attracted 11 participants (five out of 14 from the Control group and six out of 13 from the Experimental group). This simulation exercise was well attended, with most participants posting their thoughts about each of the three dilemmas. It was seen as “very, very good and engaging” and “a stimulating way of learning”. The materials were primarily a weekly video that participants could download and watch on their computers. Use of a video was seen as more fun and easier.

However, some concerns were raised that using video presentations may have resulted in less “learning” than a text-based method. This depends upon what it was intended for the participants to learn. In the case of Eagle Racing, one of the main purposes was to provide experience of collaborative decision making. In this sense, the use of video presentation and the ICDT platform was effective, since all the participants did collaborate to generate decisions.

Finally, one of the interviewees (from the Control group) also saw the ICDT platform as an opportunity to get in touch with platform users in other countries, “especially others not on the same course” (although no other course was accessible to them).

Administrators

A second group if users of the ICDT platform were administrators and professors who provided the training documents and information about seminars and other events. These all used the AtGentive-enhanced version of the platform (as seen by the Experimental group), and were not split into groups. They therefore do not form part of the pilot experiment.

However, it is of interest to note that the ICDT platform provides a very effective system for distributing this information. Without a system of this nature it would not be practical to run the course

4.3.2 Analysis

4.3.2.1. Summary of significant results

In the months prior to the Eagle Racing simulation, the Experimental group used two features of the system more than the Control group. These were:

- “Vi place 'homepage - knowledge'” – the main access page for reading postings

- "Vi place 'homepage - agents'" – the main access page for "interventions" – suggestions to read postings, activate the animated agent or take other actions within the system.

No significant differences were found in system usage once the Eagle Racing simulation began.

Results of the questionnaire show four questions that significantly differ between treatment groups. In each case the Experimental group agrees more with the statement than the Control group.

- At school I always received good grades
- I fully understand how to use the ICDT platform
- I understand all the headings on the ICDT platform

4.3.2.2. Analysis

Two global factors are relevant to the analysis. Firstly, the system usage by month shows greater usage during September. This appears to be due to two factors:

- The participants reported being very busy – including finding time for holidays – during the summer months
- The Eagle Racing simulation seemed to create a very large amount of motivation to use the system

Secondly, the number of participants is small, allowing a small number of people to have a large effect on the results. With these factors in mind, the results are now discussed in terms of the Key Indicators.

Performance and collaboration

During the summer (and before Eagle Racing) the main access page for reading postings and the main access page for "interventions" (suggestions to take actions within the system, such as read postings) was accessed significantly more by the Experimental group. The two may go together, in that accessing the Agents page will give recommendations for items on the Knowledge page (although the items may be accessed also from the Agents page).

These postings were particularly relevant during May to August as this is when the TRIM seminars were held and the main coursework done. Thus, the AtGentive enhancements may have been helping the Experimental group with performance in providing relevant information during this time. Students reported collaboration with lecturers and other speakers during this time. However, this tended to be through email and other channels as participants reported this as being easier to use than the platform. Attempts to create collaboration between students was not successful during this time (with no replies to the "Reflections on the EIS-simulation" posting; participants later gave being busy or on holiday as reasons).

The start of the Eagle Racing simulation coincided with the ending of the course. Thus the increase in activity cannot be put down to fetching more information. Rather, the simulation required collaboration between students. The simulation was very successful in this regard, with participants taking part each week and later reporting a liking for the exercise.

There is no significant difference between the Experimental and Control groups in their platform use during this time. This suggests that the AtGentive enhancements were not found to help the Experimental group with collaboration.

Taken together then, pre-Eagle Racing, students were busy and focussed on the ITM course. The AtGentive enhancements appear to have helped with performance but not collaboration. During Eagle Racing, the reverse was seen, with AtGentive enhancements appearing to have helped with collaboration but not performance.

This suggests that performance is supported by the AtGentive enhancements where the students' need is to obtain information, whereas collaboration between students is supported by the AtGentive enhancements where the students' need is to work on joint projects. Support for collaboration with lecturers does not appear to be affected by the AtGentive enhancements.

A meta-level factor at work here appears to be that of motivation – participants were successfully supported by the AtGentive enhancements according to their motivation. When intending to acquire information, the system supported their performance in this task. When intending to collaborate with other students, the system supported their collaboration. Thus, the AtGentive enhancements supported the intended activity, rather than promoting new activity.

It should be noted that the Experimental group scored higher on their self-assessed understanding of the platform (all agreeing with "I fully understand how to use the ICDT platform", rather than the Control group, for whom all-but-one disagreed). Similarly, the Experimental group scored higher on "I understand all the headings on the ICDT platform". Also, the Experimental group scored higher with "At school I always received good grades".

Generally, the interviews revealed that participants found the platform difficult to use. It is possible that some of the Experimental group's greater participation and collaboration may be due to a better understanding of the platform. If this were the case, better support for beginners may be indicated in future systems that support attention using perception. Interviewees reported that there was "too much information". Support for beginners could involve introducing features over time, as they get used to the system as a whole.

Attention

The participants' use if the ICDT system was uncontrolled from the evaluation perspective. Participants could use any computer at any time, may have a slow or unreliable connection, may multi-task (such as holding an unrelated telephone conversation at the same time) and may allow others to use the system (such as asking a secretary to retrieve a document). Thus it was not possible to measure attention directly.

We may, however, infer that the Agent homepage (used significantly more by the Experimental group and providing interventions" (suggestions) only to the Experimental group) is drawing the users' attention to specific postings and other system elements. Indeed, it is the theoretical effects of attention management that led, ultimately, to the AtGentive enhancements, including the Agent home page.

Satisfaction

Generally, participants stated that they were most satisfied with the fact that they had access to documents and information about lecturers and other students. The Search facility was an important feature for locating these. The predominant source of

dissatisfaction was that there was too much information visible at one time, with several interviewees assuming that much of the information was intended for “administrators”. In addition, they found it “hard to use” and “hard to navigate”. These factors were also identified by the heuristic evaluation.

However, no significant difference between the treatment groups was found. This suggests that while there are significant improvements that may be made to the platform, the problems encountered did not detract from using the AtGentive enhancements.

Learning

The purpose of the ITM course is to provide an environment where business managers can take steps to expend their company’s international trade. It is a practical, rather than academic, course (see Section 4.3.1.3). Learning that takes place is incidental (though critical) to this process. Therefore, no specific learning is facilitated. This makes any form of “testing” of participants impractical.

Furthermore, participants may not be aware of what they have learned. Nonetheless, the interview questions were included for learning. The responses underline the practical and implicit learning this course engenders. Participants spoke of learning about “different cultures” and the importance of respecting and understanding others, as well as learning “Practical tools”.

There is no evidence for any difference in learning between the treatment groups.

4.3.2.3. Conclusion

The AtGentive enhancements to the ICDT platform appear to support either, or both, performance and collaboration, in the sense that support is provided for the activity – acquiring information or collaborating with other students. However, it must be the users’ intention to perform these activities. The system was not found to promote such activity if not already the user’s intention.

4.4 Overall conclusion

The AtGentive enhancements to the ICDT platform differed notably from those of AtGentSchool. Both systems implemented three scenarios from the conceptual framework. For AtGentSchool – intended for primary school children – the activation of these scenarios by the agent was very effective. The students liked and responded to the agent, and performed better as a result.

AtGentNet was designed for busy adults who typically fitted a few minutes use of the platform into otherwise full schedules. It was anticipated that the agent-delivered scenario approach may be problematic for these users (as indicated by one of the additional experiments – see Section 5.1 – indeed, most participants did not visit the agent homepage or mention the agent in interviews or questionnaires.) Instead, the primary changes to the ICDT platform for AtGentive was the support of perception – that is, the subtle direction of attention by careful placement of on-screen elements.

While the positive results seen from AtGentNet are thus likely to be due in main to the more subtle manipulation of attention, potential problems with this approach were found. Many of the participants reported over-complication of the interface and difficulties of use. This suggests that a more subtle approach is necessary and indicates the need for more research to define the boundary between beneficial attention direction through perception and unhelpful information overload.

In terms of overall success, the data indicate that use of an animated agent for children can successfully promote Performance, Satisfaction and Collaboration, with no apparent detriment to other factors. The use of scenarios as design elements proved very effective in designing for children.

For adults, careful and subtle perceptual enhancements appear to be a better approach than animated agents. The data here indicate that performance and collaboration may be enhanced, but only where the motivation pre-exists. Perceptual-based attention support does not engender motivation, compared to the motivation provided to children by the animated agent in AtGentSchool.

5. Additional experiments on Attention for Online Learning

5.1 Acceptance of agents' instructions

This experiment investigated the feasibility of agent-provided assistance for two specific situations, (1) when a previously-interrupted task is resumed, offer to open previously-used contextual documents and (2) when a non-optimal task is begun, suggest a more suitable alternative task. A paper-based task places participants in the situations described. The concern was to maximise the balance between helpfulness and annoyance. The results are discussed in terms of timing of interruption and social effects. Overall, the agent needs to take account of the human's likely feelings towards any intervention; interventions must be both useful and perceived to be useful.

The experiment appears as a separate publication (Rudman & Zajicek, 2006). For completeness however, a longer summary, from the AtGentive perspective, is included in Appendix 8.

5.2 Animated agent's gestures verbal discrepancy

Two experimental studies investigated attention distribution between verbal and non-verbal messages conveyed by an Embodied Conversational Agents (ECA). The study adapts the Stroop task paradigm[16], widely used in cognitive psychology, to study executive attention in human-agent interaction. Analysing reaction times provides a reliable procedure to understand the effect of non-verbal communication (agent posture and facial expressions) on verbal comprehension (written text). The research focused on the following main questions.

1. Are non verbal messages conveyed by virtual bodies attended to?
2. And, if yes, do they facilitate or inhibit verbal communication?

The first study used the animated agent "Colette" (later renamed AtGentiGirl), as used in AtGentNet. In the first experiment, Emotion on the part of the agent was presented alongside an emotion word. The two emotions may agree or conflict. Participants were asked to evaluate if the word displayed on the screen was a positive or a negative by pressing one of two pre-set keys. Time and accuracy were measured. A clear interference effect was found for accuracy, but not for speed.

In order to explain this effect, a second investigation was conducted. This aimed to reproduce the results of the first study, but with a number of methodological modifications and improvements. In particular, an additional test was introduced, to address memory retention – participants were invited to recognise the list of words presented in the experiment from a set of distracters.

Results provided mixed support to the persona effect. Accuracy measurements showed a strong conflict between consistent and inconsistent conditions, confirming that non-verbal cues from the agent were processed and interfered with word naming. Analysis of

[16] The Stroop effect displays colour words (such as "Green") using non-matching coloured ink (such as "Green" printed in red ink). This effect makes naming the ink colour very difficult

the reaction times showed that negative words were systematically processed slower than positive words. The memory task showed no differences in recognition. This suggests that non-verbal messages coming from the agent have little effect on learning. See Appendix 9 for further details.

5.3 Animated agent's gestures and guidance of user's attention on screen

This study investigated the effects of a computer agent's gestures in guiding a user's attention on screen. The aim was to find out how an agent character's gestures can be used to attract and direct attention to visual interventions, when using a computer program or environment.

The experiment presented the participant with the following stimulus: first, an agent character appeared on the screen. Then, two simultaneous visual interventions were briefly shown and the agent gestured towards one of them. After this, the user was asked to remember the content of the interventions. The user's gaze was tracked with the help of an eye tracker to determine where his or her attention was focused during the tasks. The position on-screen of the agent and the direction of the intervention, were systematically varied during the tasks.

The results showed that the agent's gesture had a significant effect on how well the participants were able to remember the targeted intervention object. See Appendix 10 for further details.

5.4 AtGentNet eye-tracking study

The aim of this experiment was to collect data on the focus and shifting of the user's attention while using the AtGentNet platform. The test participants were presented with 5 simple tasks to complete using AtGentNet. These tasks were designed to simulate typical usage of the platform, such as to see the latest news items. Figure 29 shows an example "heatmap" – the main areas of visual attention are denoted by a lighter colouring.

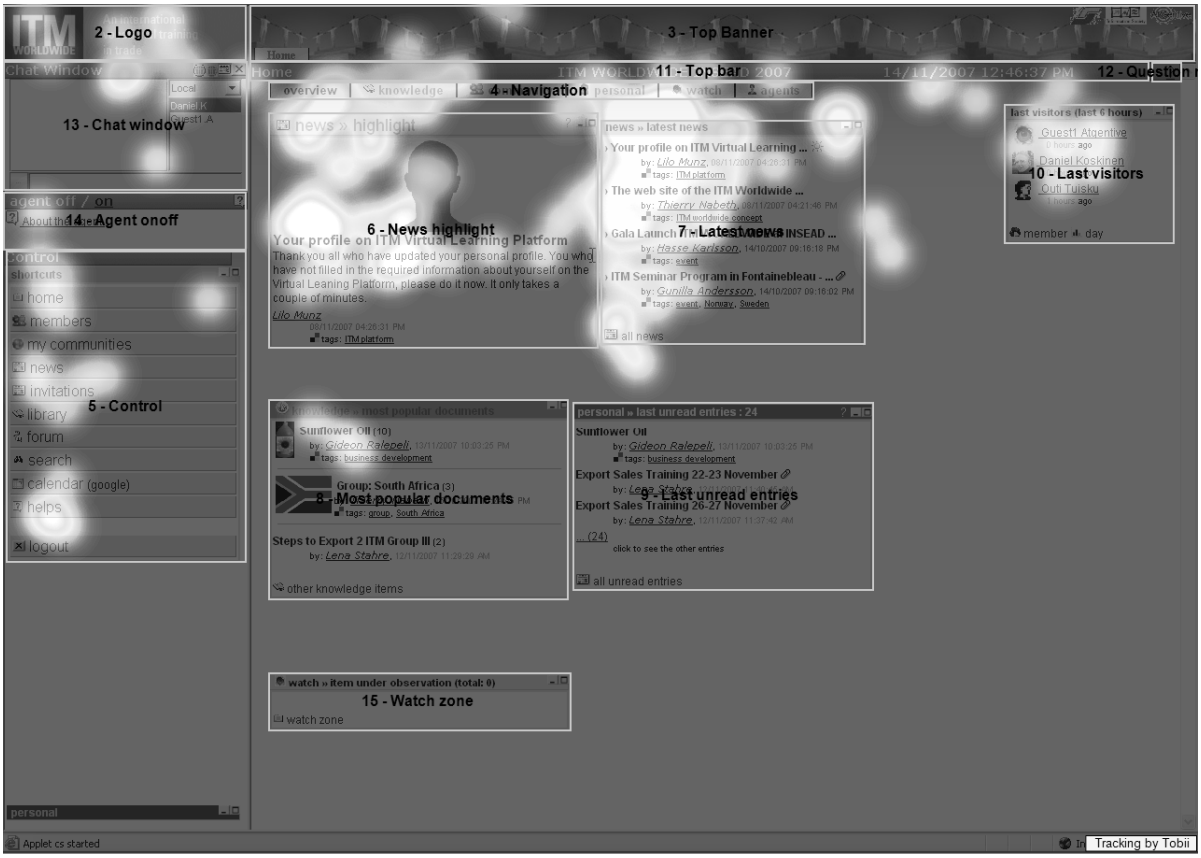


Figure 29 - Example "heatmap", showing main areas of visual attention (lighter colours)

The distribution of gaze between different elements of the screen varied between tasks. In the first task, participants had little difficulty in finding the news items, with few fixations elsewhere. On the second task of finding the help feature, none of the participants looked at the question mark link, which suggests it is hard to find and could be more prominent.

Finding a member's profile was delayed by over-attention on the images of members' faces, causing the name not to be found. The fourth task suggested that the chat window could be improved upon to better grab the user's attention. See Appendix 11 for further details.

5.5 General applicability of the conceptual framework

The purpose of this experiment is to evaluate the level of general applicability of the conceptual framework, using the "restoring context" scenario. During the project, we were able to implement the selected concepts in two different applications (AtGentNet and AtGentSchool). This is an indication of general applicability. This experiment goes further in this direction.

One of the concepts considered most interesting in the conceptual framework, was context restoration. As discussed in deliverables D1.2 – "State of the Art Report", and D1.3 – "AtGentive conceptual framework and application scenarios", the time required to restore an interrupted task is one of the highest burdens that interruptions bring to current learning and working environments. The concept of context restoration couldn't however be tested in any of the two pilot studies. This was due to the fact that the two pilots concentrated on two individual applications whilst the effects of support to context

restoration are most useful when dealing the user working on several different applications, or several devices. This section reports the experimental work in progress to evaluate the effects that support to context restoration could have in multi-application environments.

The first important result of this experiment is that it is possible, with the current technologies, to implement support to context restoration as described in the AtGentive Conceptual Framework. It is difficult to make a preliminary analysis based on the small sample currently available. However, by looking at the partial results, we can see that: task resumption times tend to be significantly smaller in the experimental group than in the control group. Users in the experimental group generally understood easily and used appropriately the tools offered by the interface, they also declared to like to use the interface. Users in the experimental group tend to work faster than users in the control group. Currently, we don't have enough data to allow us to detect differences based on age, gender, lateralization, computer experience, or any other relevant subjects' characteristics. See Appendix 12 for further details.

5.6 General applicability of the Reasoning Module

After the AtGentSchool pilot we are interested in testing how well the Reasoning Module (RM) may support user attention with applications that are not limited to those explored in the course of the project. Also we want to explore the cases in which the RM interacts with several user-level applications. By accepting events from several applications simultaneously, we assume the RM could be capable of supporting the management of attention within a good part of the tasks the user will have to perform on the computer.

A simple test platform has been implemented for experimenting with the RM in this way. Currently, the platform is being used to run an experiment. The purpose is to observe the potential effects of a service supporting users in resuming previously interrupted tasks by lowering the cognitive load that it takes to remember to do so. The intention is to verify that resuming interrupted tasks will nearly always require cognitive effort, and that displaying a reminder of the interrupted task right after the interrupting task has been completed will prove beneficial with respect to the time that it will take to resume the interrupted task and the ease with which one accomplishes the resumption. See Appendix 13 for further details.

6. Strategic Evaluation

The AtGentive project has already made a significant contribution outside of the project through publication and other dissemination activities (see deliverable D6.4 – "Assessment and Consolidation report in the perspective of further exploitation and final exploitation plan" – for a full list of publications). In addition, a number of specific post project activities are planned which will further exploit the knowledge and assets created (see also D6.4).

The purpose of the strategic evaluation is to document and evaluate the potential of the AtGentive outputs to make ongoing contributions in the outside world after completion of the AtGentive project. Three stages of strategic evaluation have been identified:

- Project Objectives – the project's goals and desired objectives, with reference to the original Description of Work
- Project Outputs – the tangible results generated by the project – concepts, knowledge and artefacts – as identified by the individual responsible project partner(s)
- Project Impacts – potential value and effects of the project outputs. This will be assessed using Key Assessors, as described in section 6.3

6.1 Project Objectives

The project's original goals and desired objectives are described in the original AtGentive Description of Work, an extract of which is reproduced here:

Objectives:

Better understand the role of attention in the effectiveness in Learning (motivation, how to be organized, etc.)
Help users (student & educators) to learn to better manage their attention (support for users with attention deficit disorder, procrastination, how to organize the time for a learner, etc.)
Reduce the cognitive load and direct focus.
Better understand the use and effectiveness of artificial characters in a learning context.
Increase motivation, and reduce attrition and dropout in open-to-use digital learning platforms and digital collaborative environments.
Increase the social activity (social attention)

In particular we address the following problems:

The difficulty for young children to concentrate for a long time (it is important therefore to identify when they have detached)
The difficulty for knowledge workers to organize their work and interaction with other (given the deadlines, the interruptions, etc.)

Technical objective:

To design an agent-enhanced collaborative learning infrastructure, informed about the learner attentional mental state, able to provide more effective interaction (less disruptive), and supporting the learner to manage his/her attention.
To design mechanisms integrating eye feedback to artificial characters technologies.
To design an attentional facet to the "brain" of an artificial agent.

Exploitation objective:

For the scientific partners, to leverage the knowledge generated in AtGentive (assessment and intelligent support of user attention via embodied characters) in publications, research projects, prototypes, education practices and cooperation with other organizations.
For the industrial partners, to consolidate (1) their technologies and platforms by incorporating "attention" related functionality, so that they enrich their offers to their customers; (2) their knowledge and experience in the educational market.
For the users' partners: to improve the richness and the effectiveness of the "services" offered to the students."

6.2 Project Outputs

By reference to existing AtGentive deliverable documents, the areas listed below have been identified as discrete project outputs that may undergo individual strategic evaluation. (Note that the list consists both of objects (e.g. the ASKME module) and knowledge (e.g. Eye-tracking experiment results).)

- Literature survey (State of the Art)
- Conceptual framework
- ASKME module
- Reasoning module
- Agents (animated characters)
- AtGentSchool
- AtGentNet
- Results of pilots
- Eye-tracking experiment results
- Publications – papers / web site
- Student experience on project

Note that the list consists both of objects (e.g. the ASKME module) and knowledge (e.g. Eye-tracking experiment results).

6.3 Key Assessors for evaluating project impacts

The main areas for evaluation are:

- How may the project output be used in its present state?
- What may the project output reasonably be adapted to achieve?

Within these main areas, the evaluation needs to quantify the effort required to adapt / utilise the output. This is done by applying the following key Assessors:

- **Implementability** – how the product may be utilised as is; how the product may be utilised in a modified form. These are from a practical perspective, rather than one of perceived desirability or usefulness.
- **Integratability** – overview of existing or likely products, systems or areas in which the output may be incorporated. This is from a practical perspective, rather than one of perceived desirability or usefulness.

- **Practicality** – expertise and effort required to utilise the output in its current form; expertise and effort required to extensively modify the output. For example, programming language – beginner / good / expert, effort – minor changes / major changes / complete rewrite necessary.
- **Limitations** – any specific known limitations to the output that are likely to be significant. For example, programming language incompatibilities, unresolved known problems. Include also Intellectual Property limitations.

6.4 Applying the Key Assessors

Each of the twelve project outputs identified in Section 6.2 is now examined with reference to the Key Assessors defined in Section 6.3 (Implementability, Integratability, Practicality and Limitations). The Assessors are preceded by an overview description of the output, and the knowledge gained in its development. Further details of each of the project outputs are to be found in deliverable D6.4, as well as other deliverables referenced below in individual descriptions.

6.4.1 Literature survey (State of the Art)

Description

The State of the Art report reviews research related to the support of attention in systems for collaboration and learning. It presents the most relevant results in attention-related research in cognitive psychology. It introduces systems that have been designed with the explicit aim of supporting some attentional processes. It provides an overview of the specific issues related to the support of attention in educational, work, and business environments. Finally, it describes the issues and technologies related to psychophysiological measurements of attention and to the integration of embodied agents in attention-aware systems. (See also deliverable D1.2 – “State of the Art Report”.)

Implementability

The State of the Art report provides a resource for future researchers – both academic and commercial - in the area of attention support, in systems for collaboration and learning. It allows publications to be easily identified and links related publications.

Integratability

The knowledge embodied in the State of the Art report may be used in future publications. In particular, we plan to further its dissemination by integrating its content, in a revised form, in an AtGentive-based book.

Practicality

The State of the Art report is written to be easily understood by anyone with sufficient expertise to benefit from its contents.

Limitations

Any State of the Art report is only complete at its date of publication. Each year papers are published that advance knowledge in the area of Attention (including papers from AtGentive itself), making the report incrementally out of date. However, until an updated version of the report is created it will remain a good starting point for researchers.

6.4.2 Conceptual framework

Description

The AtGentive conceptual framework proposes that the learning processes may be supported at several different levels (regulative, cognitive, and meta-cognitive) and that such support may be translated in a set of corresponding interventions directing the learner's attention to the appropriate foci. The process of directing attention is in itself decomposable in a set of levels (perception, deliberation, operation, meta-cognition) allowing one to support the corresponding attentional processes. The framework is based on a set of realistic scenarios that both have guided its development and have served as reference in the design, implementation, and validation phases of the project. The framework details the attention-relevant events that may guide any application in the definition of learners' current and possible-future foci of attention; it describes the reasoning that a system may be able to perform on the basis of these events; and the interventions that can be made with the users in order to support their attentional processes. (See also deliverable D1.3 – “AtGentive conceptual framework and application scenarios”.)

Implementability

The AtGentive Conceptual Framework provides a reference for future researchers – both academic and commercial – for the definition of attention support services in a wide variety of applications for collaboration and learning.

Integratability

The AtGentive conceptual framework is a theoretical framework capable of representing the wide range of attentional processes that one may want to support within learning and collaboration environment and it is not restricted to the attention support actions and modules that have been implemented in subsequent parts of the project. Experiments conducted by AUP show that the conceptual framework can be used as a reference in research and applications significantly different (both in terms of objectives, and technology employed) to those in AtGentive.

Practicality

Although some familiarity with basic concepts related to human attentional processes, as well as learning and collaboration environments, is required to fully understand the AtGentive Conceptual Framework, the description of the framework is written to be easily understood by researchers and practitioners with different backgrounds. The use of scenarios enables to connect the theoretical framework to practical situations.

Limitations

Some aspects of attention support could be better represented (i.e. modelled in more detail) in the framework, including: personalised searches, long term attention, visualisation, and adaptive resources access.

6.4.3 ASKME module

Description

The ASKME module is a Web Service that monitors users and provides monitoring information to the reasoning module. The information is based on face analysis, on mouse use, and on key presses. Face analysis includes face detection and tracking and gender classification, and various face movement measurements are derived from the tracking data. (See also deliverable D2.1 – Design specification of the Attentive agent module.)

Implementability:

Face analysis measurements and mouse and keyboard activity based arousal and activity measurements can be used in reasoning. More monitoring information could be provided, for example, by including facial expression analysis and speech recognition.

Integratability:

The ASKME module uses several open standards, including SOAP, WS-Notification, and WS-Topics, in order to support integration with other Web Services. The module can be integrated to any application that benefits of reasoning based on monitoring one ore more users and has one or more Windows systems with keyboard, a mouse, or a web camera (or similar) available. Desktop applications typically fill these requirements and learning applications with interactive agents are good example as shown in the project.

Practicality:

The ASKME module provides the monitoring information as SOAP messages through interface defined with WSDL. Use of the module requires understanding of these standards and/or tools that support the usage. A reference implementation of the simple application that uses the module has been written with Java –language that eases the actual usage.

Limitations:

The only unresolved known problem is that subscribing to one message type through pull point subscribes to all message types. Errors may happen in face analysis which causes inaccuracies to the measurement data. Apache MUSE and Intel® OpenCV library have been used in ASKME module implementation but their licenses allow non-commercial and commercial use.

6.4.4 Reasoning module

Description

The reasoning module implements the core attention-related reasoning of AtGentive. It receives information about the learner activity from the applications and the ASKME module in the form of events, and produces output in the form of interventions, suggested to the application and the user, aimed at supporting the learner in his/hers attentional choices. The behaviour of the module closely mirrors the analysis offered in the AtGentive Conceptual Framework provided. The reasoning module is designed as a multi-agent system composed of three types of agents: Event agents, Integration agents, and Intervention agents. (See also deliverables D2.1,– Design specification of the Attentive agent module, D2.2 – Design Specifications, D3.1 – Early prototype, D3.2 prototype.)

Implementability

The reasoning module prototype, as specified in D3.2, is usable by applications following the interface described in D2.2. This requires that applications (1) send events to the reasoning module, (2) handle interventions as sent by the reasoning module. Note that the application doesn't necessarily have to make use of all the events types implemented by the reasoning module. Applications that only pass a small subset of the possible events will still receive interventions from the reasoning module (obviously only related to the events that have been passed). The reasoning module behaviour can be customised by providing any application-specific rules that the reasoning module agents should apply.

Integratability:

The reasoning module software can be integrated in a very straightforward manner with task-based applications (i.e. applications that, in some form or another, describe the user activity in terms of identifiable sequences of actions).

Practicality:

The events/interventions interface of the reasoning module is defined in a standard manner so that any application developer should be able, with a reasonable effort, to generate the appropriate events for the reasoning module, and to receive the interventions. The manner in which these interventions will be used by the application may range greatly spanning from a simple presentation of the interventions to the user, to complex reasoning about the interventions and changes to the application behaviour. The customisation of the reasoning module (i.e. the creation of application-specific rules) remains a task that would be considered quite laborious but unfortunately we were unable to address this issue within the time frame of the project. The Reasoning Module has been developed using standard open source components, insuring easy reuse and evolution.

Limitations:

Complexity of reasoning module customisation (see "practicality").

6.4.5 Agents (animated characters)

Description

Two virtual characters have been created (see Figure 1), as well as a framework for creating and visualising character interventions according to their mood and the strength of their action. Prior to AtGentive, intervention scripts where directly written and executed as a list of actions composed of animations and dialogs. Now, many scripts are automatically generated from templates and are well adapted to the context of agent interventions, as provided by the reasoning module. (See deliverable D2.1 – Design specification of the Attentive agent modules.)

Implementability:

This same framework can be used without any modification in other circumstances than attention management as it allows generating several scripts from a couple of parameters and benefits to any user who needs to rapidly produce many variations of a few sample scripts. The mechanism is implemented as a client server application where the client side is a standard web browsing environment and the server side is any HTTP server supporting PHP.

Integratability:

This framework can be used in any web site where the agent communicates with the users for guiding them, giving news or being a brand ambassador for a company.

Practicality:

The technical skills required to integrate the characters into an application are the same as those required for integrating any dynamic component in web pages. The effort corresponds to any integration tasks involved in web site development.

Limitations:

The version of the Living Actor™ player used for AtGentive is 3D. It needs to be installed on local user machines and does not work properly on Vista or Mac OS X. However, a Flash version of the player is also available and would overcome this problem.

The two characters created for AtGentive project are part of the Cantoche virtual character gallery and can not be integrated for an exclusive usage in a customer application. However, this agent intervention framework does not depend on any particular virtual character and any new embodied agent can benefit from it.

Reuse of the agents using the Living Actor™ system requires a commercial agreement.

6.4.6 AtGentSchool

Description

AtGentSchool is an attention aware learning environment. Innovative learning arrangements are characterised by constructive learning tasks in a situated environment in which students work collaboratively. These environments draw largely on the regulative capacities of students allowing high control over their own learning process. Many students are unable to successfully sustain in these environments, due to a lack of self-regulative learning skills. Research findings indicate that scaffolding of the learning process with an emphasis on self-regulative learning skills may support these learners. In the learning environment AtGentSchool, the learning is scaffolded by an attention aware virtual agent. Dynamic and adaptive scaffolding are made feasible through the attention management system.

Implementability:

The AtGentSchool platform is a prototype. It is usable in both laboratory and real school settings for specific research projects to learn more about the relation between attentions management and scaffolding of the learning process. The AtGentSchool system is not ready to be sold as a product due to the little practical experience we have with this new prototype. As prototype it will be tested to fine tune the scaffolding to more diverse settings and conditions in schools. This will lead towards a real integrated use within the e-learning platform Ontdeknet and other e-learning applications.

Practicality:

Ontdeknet are able to enhance the existing reasoning module for future purposes and could support implementation within other e-learning platforms.

Limitations:

AtGentive has shown that the implementation of this attention aware agent within an e-learning platform requires both engineers and pedagogical expertise to work closely together.

There was a significant overhead in creating the intervention rules for AtGentSchool. This will impact on the scalability of the approach.

Reuse of AtGentSchool requires a commercial agreement.

6.4.7 AtGentNet

Description

AtGentNet is a platform aimed at supporting the online interaction of groups of people engaged in an offline training programme in which they can only meet physically during short periods of time (a few days every several weeks). In particular, AtGentNet aims to help this group stay “in touch” while they are physically dispersed, and to contribute to helping them know more about each other, stimulate their interaction and knowledge exchange about the programme, and keep them motivated.

Implementability:

The AtGentNet system is currently in use and fully implemented.

Practicality:

The system is Web based and therefore accessible from all over the world.

Limitations:

There are a few design limitations which were identified during Heuristic Evaluation which are due to be corrected in the subsequent version of AtGentNet. In addition, further work is indicated in assessing the trade-off between perception and information overload.

Note: AtGentNet is not open source.

6.4.8 Results of pilots

Description

The results of pilots report experiences gained from the two deployments of AtGentive-enabled systems. In AtGentSchool, an existing commercial e-learning platform was extended with attention management features and deployed in Czech schools. In AtGentNet, en existing e-learning portal was augmented with attention management features and deployed for virtual training of business managers. The both pilots used a combination of questionnaires, interviews, and transaction logs to record the user feedback and system usage.

See also deliverable D2.3 – “Specification of the approach and mechanisms for validating and improving the user interaction”, deliverable D5.1 – “Specification of the implementation of the pilots”, and Sections 4.2 and 4.3 of this report.

Implementability

The pilot results can be implemented in general as “know-how” for organising a pilot for any educational software. This can communicate how a pilot works what may be important for both software providers and users.

AtGentSchool may serve as a case study, to be shown to schools reluctant to join any research project and fearing that the school / student life will be adversely affected. The pilot itself showed that none of those was true. Results also showed how differently children respond to software depending on how much they were used to working with educational software before. Deliverable 5.2 includes a description of the preparatory phase: the number of meetings to be carried out (including their content) before the actual pilot starts and also lists all the small obstacles coming along with a pilot execution within an everyday school life. In addition, the level of understanding of the software by teachers has an influence on pilot results.

The experiences gained from the pilots also provide academic and commercial researchers with knowledge about the challenges in piloting complex educational software systems remotely in an international, multi-cultural setting.

Integratability:

If the results were to be put in a form of manual or “a case study” they could be integrated into any document.

Practicality:

The results create a very practical document for research projects and pilots in this area. The report on the results of pilots is written to be understandable by anyone familiar with the problem domain. Very often projects include only a few lines on pilots (length, time, involvement of teachers / students) and only very general guidance on how to carry it out and the critical do’s and don’ts.

Limitations:

Some of the knowledge gained from the pilot experiments may be culture - or country - specific. However, the pilot results are potentially beneficial to all developers of educational software at a general level.

6.4.9 Physiological experiment results

Description

Four separate experiments are described – see Section 5 of this report.

Implementability

The results of the experiment offer us valuable information on the effects that gesture- and expression-based cues of an embodied agent have on a user’s attention and, by extension, learning performance. This information can be used to enhance the usability and role of agents in various applications and environment that involve the guidance of attention.

Integratability:

Although the experiment was conducted with the Cantoche agent, the results can potentially be extended to other similar humanlike embodied agent with similar capabilities to gesture.

Practicality:

The results of the experiments provide a solid basis for further experiments in this area. On their own, they provide an interesting point of view into understanding human-agent interaction.

Limitations:

The situations investigated only represent a part of the agent’s capabilities for attention guidance. Furthermore, visual attention will also be affected by things as interface elements, usability of the system, the user’s interests and level of skill.

6.4.10 Publications – papers / web site

Description

Appendix 7 contains a full list of AtGentive-related publications, as at this report’s date.

6.4.11 Student experience on project

Description

UTA: Five PhD students from the University of Tampere, Finland, have been working on AtGentive; their areas of study are physiological and affective computing, emotions and attention guidance, computer vision in human-computer interaction, eye-tracking and gaze-assisted interfaces, and interactive information visualization. One of the PhD students successfully defended his thesis during the project and another defence is scheduled for December 2007. Two of the dissertations under preparation are directly connected to the themes of the AtGentive project.

6.5 Strategic Evaluation – Meeting the Project Objectives

AtGentive sought to meet the project objectives using a step-by-step approach based upon practical action. Beginning with the Conceptual Framework, the components of attention support were analysed and a support approach proposed. This led to the adoption of two strategies:

- The use of scenarios as design elements
- The support for perception as a component of attention for adults

These strategies led to systems designed for example to help users to learn to better manage their attention, through meta-cognitive support in AtGentSchool, and to support users in organising their time, through directing their focus to relevant postings in AtGentNet.

For AtGentSchool, support offered by the animated agent addressed the difficulty for young children to concentrate throughout a lesson without direct teacher intervention, by offering direct motivational support. It has been found that artificial characters are more appropriate to children's, rather than adult learning. For adults, the perceptual support for attention provided by AtGentNet was more relevant than an animated agent. For AtGentNet, support for social activity was particularly relevant, and improved collaboration was observed in the pilot.

Learning was not seen as directly affected for either pilot system due to the AtGentive enhancements. However, as noted in the Summative evaluation, learning was very difficult to measure within the constrains of the pilot project. No detriment to learning was found with the application of AtGentive, and it is suggested that a much larger study will be necessary if learning gains are to visible.

From a technical perspective, the ASKME and Reasoning Modules have demonstrated in AtGentSchool the ability to assess the learner's attentional mental state and convert this into attentional support (interventions), delivered by an animated agent in an effective manner. Experiments have been conducted to assess the opportunity of integrating eye feedback into this system.

In terms of academic exploitation, considerable success has been achieved in analysing the effect of attention support and of the embodied agent, using such methods as the Stroop effect and eye-tracking. These results will contribute significantly (as detailed in Appendix 7) to the body of work underpinning the understanding of the use of attention support and embodied agents.

The industrial partners have incorporated attention-related functionality into their platforms, increasing their experience in the educational market, in due course, benefitting their customers.

6.6 Strategic Evaluation - Conclusion

The project approach of attacking the problem from several perspectives corresponding to the strengths and expertise of individual participants has proved very fruitful. AtGentive has met its principle objectives to successfully design and run two pilot studies which advance the support and understanding of attention with regard to educational software. While the support is situated in particular applications, knowledge gained may have implications in a wider context. This work has contributed across several disciplines in the research community – teaching and learning, collaborative systems, human-computer interface, and intelligent agents to mention but a few – and offers a basis for continuing research in this new and expanding area.

7. References

Angehrn, A. A. (2004). *Designing Effective Virtual Communities Environments: The ICDT Platform* (CALT Report 10-2004): INSEAD CALT.

Beck, K. (1999). *Extreme programming explained : embrace change*. Harlow Addison-Wesley.

Bruner, J. S. (1983). *Child's talk: learning to use language*. Oxford: Oxford University Press.

Chalmers, A. F. (1994). *What is this thing called Science?* (2nd ed.). Milton Keynes: Open University Press.

Czerwinski, M., Horvitz, E., & Wilhite, S. (2004). *A diary study of task switching and interruptions.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems, Vienna, Austria.

DeVellis, R. F. (2003). *Scale Development, Theory and Applications* (2nd ed.). London: Sage Publications:.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

ISO. (1998). Ergonomic requirements for office work with visual display terminals (VDTs) (Vol. Part 11: Guidance on usability): International Organization for Standardization.

Jokela, T., Aikio, K.-P., & Jounila, I. (2005). *Satisfied and Dissatisfied at the Same Time - A Preliminary Two-Factor Theory of User Satisfaction.* Paper presented at the 11th International Conference on Human-Computer Interaction, Las Vegas, Nevada.

Jones, A., Scanlon, E., Tosunoglu, C., Morris, E., Ross, S., Butcher, P., & Greenberg, J. (1999). Contexts for evaluating educational software. *Interacting with Computers, 11*(5), 499-516.

Kantner, L., & Rosenbaum, S. (1997). *Usability Studies of WWW Sites: Heuristic Evaluation vs Laboratory Testing.* Paper presented at the SIGDOC, Salt Lake City, UT.

Nielsen, J. (1993). *Usability engineering*. Boston AP Professional, .

Nielsen, J. (1994, Apr 24-28). *Enhancing the Explanatory Power of Usability Heuristics.* Paper presented at the CHI, Boston.

Nielsen, J. (2006). *Ten Usability Heuristics*. useit.com. Available: http://www.useit.com/papers/heuristic/heuristic_list.html [2006, Mar 20].

Nielsen, J., & Mack, R. L. (Eds.). (1994). *Usability inspection methods*. New York: Wiley.

Rolf, M., & Jakob, N. (1990). Improving a human-computer dialogue. *Commun. ACM, 33*(3), 338-348.

Rudman, P., & Zajicek, M. (2006). *Autonomous agent as helper - Helpful or Annoying?* Paper presented at the IAT 2006 - IEEE/WIC/ACM International Conference on Intelligent Agent Technology, Hong Kong.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem-solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89-100.

8. Appendixes

8.1 Appendix 1 - Usability heuristics

8.1.1 Established heuristics [17]

Heuristics	Description
Aesthetic and minimalist design	Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.
Consistency and standards	Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.
Error prevention	Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.
Flexibility and efficiency of use	Accelerators -- unseen by the novice user -- may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.
Help and documentation	Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.
Help users recognise, diagnose, & recover from errors	Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.
Match between system and the real world	The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.
Recognition rather than recall	Minimise the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.
User control and freedom	Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.
Visibility of system status	The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.

[17] {Nielsen, 2006 #125}

8.1.2 Additional AtGentive heuristics

Indicators	
Key Indicator One: **Attention Distraction is minimised**	***Attention distraction*** The user should not be interrupted in their task, unless the interruption assists that task significantly or is justified by the importance of the interruption. Where appropriate, interruptions should be delayed until the user is less busy. Any animated agent should not be unduly distracting ***Success in attracting attention*** Where the system attracts the user's attention, it should do so in a manner that will not be accidentally overlooked or misinterpreted
Key Indicator Two: **Performance** (Effectiveness and Efficiency)	***Task is performed well*** Interventions should not cause a task to be performed less well overall. Where the intervention is intended to improve the performance of a task, it should do so
Key Indicator Three: **Satisfaction**	***Overall satisfaction*** All suggestions / interventions made by the system should appear to the user to have at least some effective purpose. The user should not consider any suggestion to be "pointless" or "stupid" ***Positive image of the animated character*** The user's immediate reaction to seeing any animated character should be at least neutral and preferably positive. The user should anticipate that the character's appearance will make their task easier, not more difficult. The user should not have negative feelings about the animated character (threatened, humiliated, etc.) ***User control and freedom*** This is an extension to the "standard" heuristic. The user should feel in control of the AtGentive interventions. The user should not be worried that they will be interrupted at any moment, or that they are likely to miss something important
Key Indicator Four: **Learning** (Learning experience is supported)	***Improvement of the learning experience*** Interventions should not cause the learning experience to be degraded. Where the intervention is intended to improve the learning experience, it should do so
Key Indicator Five: **Collaboration** Collaboration is supported	***Improvement of collaboration*** Interventions should not discourage collaboration. Where the intervention is intended to improve collaboration, it should do so.

8.2 Appendix 2 – AtGentSchool - Abstract concepts and questions for Questionnaires

Questions are labelled for use as follows:

- [First day] = at the first use of the final system
- [Bi-Weekly] = once every two weeks (at weeks 2, 4 and 6)

Topics	Questions
Demographics	
For teachers	
Age	**Age group** (18-25, 26-35, 36-45, 46-55, 56+)
Gender	**Gender** (Male, Female)
Experience of teaching in general	**How long have you worked as a teacher (in years)?**
Experience of computer-based teaching	**How long have you taught classes where children use computers (in years)?**
General attitude towards teaching	**I like teaching** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
General attitude towards students	**Most of my students are cooperative** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely) **Most of my students are hard-working** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely) **Most of my students are difficult to control** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely) **How often do you need to ask your pupils to pay more attention?** **It is easy to capture the pupils' attention throughout a class** **(**Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely) **Most of my students are quick to learn** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
Personal computer-skills	**Compared to my colleagues, my computer skills are** **(**Much better, Better, Similar, Less good, Much less good)
English ability	**I am good at understanding written English** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
Lived / travelled outside Europe?	**Have you travelled outside Europe?** (Yes, No)
Knowledge of New Zealand	**I know nothing about New Zealand Agree completely** (Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)

Atgentive (goals & interferences)	**I need to know more about AtGentive and its goals** **(**Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely) **I worry that the AtGentive pilot will interfere with my teaching** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
Not yet measured:	*General attitude towards computers* *Age range taught* *Qualifications* *Use of other learning management systems?*
Attention	
For teachers	
Assessment of the direction of students' attention (task/error recovery/teacher/other (useful) / other (non-useful) [Bi-Weekly]	**Compared to an ordinary computer class, how did students spend their time:** **Working on the computer** (Much more, More, A little more, The same amount, A little less, Less, Much Less) **Working on the task but not on the computer** (Much more, More, A little more, The same amount, A little less, Less, Much Less) **Speaking to you** (Much more, More, A little more, The same amount, A little less, Less, Much Less) **Not doing anything useful** (Much more, More, A little more, The same amount, A little less, Less, Much Less)
Ease of (teacher's) diverting students' attention to AtGentSchool [Asked in general, not per pair] [Bi-Weekly]	**It was easy to direct students' attention to AtGentSchool** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
Ease of (teacher's) diverting students' attention away from AtGentSchool [Asked in general, not per pair] [Bi-Weekly]	**It was easy to direct students' attention away from AtGentSchool to something else** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
Problems observed in software's direction of attention [Bi-Weekly]	**I noticed times when Honza tried to direct the students' attention and failed** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
For students	
Helpfulness of the agent when it said what they should do (Individual question per intervention?) [Bi-Weekly]	**Honza, the agent, helped me a lot** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Satisfaction	
For teachers	
General (own) satisfaction with AtGentSchool [First day] [Bi-Weekly]	**I am completely satisfied with AtGentSchool** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)

General (own) dissatisfaction with AtGentSchool [First day] [Bi-Weekly]	**I am very dissatisfied with AtGentSchool** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
Observed students' comments on AtGentSchool [Bi-Weekly]	**Please give any comments that the students made about AtGentSchool** (Text box)
General speed of system response [First day] [Bi-Weekly]	**AtGentSchool responded quickly enough** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
Assessment of students' engagement with tasks [Bi-Weekly]	**Students were really engaged in using AtGentSchool** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
(Own) feeling of control [First day]	**I feel in control when using AtGentSchool** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
(Own) feeling of control [Bi-Weekly]	**When students are using AtGentSchool, I feel in control of the lesson** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely) **During the lesson, I have sufficient control over AtGentSchool** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
(Own) level of understanding of software [First day] [Bi-Weekly]	**I fully understand how to use AtGentSchool** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
Meeting of (own) pre-existing expectations [First day][Bi-Weekly]	**AtGentSchool fully met my expectations** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
Questions for teachers as expert evaluators [First day]	**Students will be able to understand AtGentSchool easily** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely) **AtGentSchool will be easy for students to use** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely) **AtGentSchool will be very useful teaching purposes** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely) **AtGentSchool will be good for supporting project work** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely) **I understand AtGentSchool very well now** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
For students	
General satisfaction with AtGentSchool [Bi-Weekly]	**AtGentSchool does just what I want** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
General dissatisfaction with AtGentSchool [Bi-Weekly]	**AtGentSchool does not do anything that I want** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Liking of what the agent looks like [Bi-Weekly]	**Honza looks great** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Friendliness of the agent [Bi-Weekly]	**Honza is really friendly** (Agree completely, Agree, Undecided, Disagree, Disagree completely)

Helpfulness of the agent [Bi-Weekly]	**Honza is very helpful** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Annoyingness of the agent [Bi-Weekly]	**Honza is very annoying** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
How much they think the agent likes them [Bi-Weekly]	**I think Honza likes me a lot** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
General speed of system response [Bi-Weekly]	**When I use the keyboard or mouse, AtGentSchool does something in response straight away** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Understandability of the language used [Bi-Weekly]	**I understand everything AtGentSchool tells me I should do** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Liking of what AtGentSchool looks like generally [Bi-Weekly]	**I like the look of AtGentSchool** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Ability to understand the system generally [Bi-Weekly]	**I know what I'm doing when I use AtGentSchool** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Feeling of control when using AtGentSchool [Bi-Weekly]	**I feel in control of AtGentSchool** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Helpfulness of software response when they made an error	[This concept is not questioned to avoid inferring that the children made errors]
Usefulness of on-screen instructions / help (excluding agent) [Bi-Weekly]	**The instructions on the screen are really helpful** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Usefulness of paper/verbal instructions/other [Will there be any?] [Bi-Weekly]	**I understood what the teacher told me to do** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Time taken to ask teacher questions [Bi-Weekly]	**I could quickly get hold of the teacher to ask questions** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Ability of teacher to answer software-related questions [Bi-Weekly]	**The teacher knew all about AtGentSchool** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Meeting of pre-existing expectations	[This concept is not questioned as "expectations" implies an understanding of the situation in which the software is provided. It was decided to infer this from questions about their experience of computer use]
How much they enjoyed the lesson [Bi-Weekly]	**I really enjoyed the lesson** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
[Bi-Weekly - First time only]	**What do you use computers for?** (Tick boxes: Games, Web browsing, Mail, Instant Messaging, Programming, Homework, Something else – what? [a text box])
Performance (Effectiveness and Efficiency)	
For teachers	
Quality of work per student pair	**The quality of work was the best work they have ever done** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree,

[Bi-Weekly]	Disagree completely)
Proportion of teacher's time helping with software use [Bi-Weekly]	**Please estimate what proportion of the lesson did you spend:** **Helping students use the software** (0-100%) **Helping students with the learning tasks** (0-100%) **Other** (0-100%)
Frequently observed / persistent difficulties [Bi-Weekly]	**Please describe any specific difficulties that occurred more than once** (40x40 text box)
For students	
Time spent using vs. learning software [Bi-Weekly]	How much time did you spend learning how to use AtGentSchool, compared to actually using it?
Learning	
For teachers	
Assessment of students' meeting of learning goals [Bi-Weekly]	**We had ambitious goals for this lesson** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely) **The students met these expectations very well** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
Learning breakdowns observed [Bi-Weekly]	**Please describe any specific instances where student(s) failed to learn what was expected** (Text box)
Learning breakthroughs observed [Bi-Weekly]	**Please describe any specific instances where student(s) succeeded in learning by overcoming previous difficulties** (Text box)
For students	
Description of what they think they have learned (informal, not test) [Bi-Weekly]	**What are the most interesting things you have learned about New Zealand?** (Text box)
Collaboration	
For teachers	
Comparison of collaboration in non-computer class [Bi-Weekly]	**The collaboration was very good** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
For students	
Who typed on the computer most often (proportion) [Bi-Weekly]	**We shared equally the typing on the computer** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Who thought what to type most often (proportion) [Bi-Weekly]	**We shared equally deciding what to type on the computer** (Agree completely, Agree, Undecided, Disagree, Disagree completely) **We shared equally deciding how to do the task** (Agree completely, Agree, Undecided, Disagree, Disagree completely)

8.3 Appendix 3 – AtGentNet - Abstract concepts and questions for Questionnaires

Topics	Questions
Demographics	
Age	**Age group** (18-25, 26-35, 36-45, 46-55, 56+)
Gender	**Gender** (Male, Female)
Personal computer skills	**Compared to my colleagues, my computer skills are** **(**Much better, Better, Similar, Less good, Much less good)
Use of other learning management systems	**Please list any other computer systems you have used for distance learning** (e.g., Moodle, Drupal, ...)
General attitude towards computers	**I like it that my work involves computers** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
English ability	**I am good at understanding written English** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
Attention	**I have difficulty concentrating on one thing for some time** **(**Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
Satisfaction	**I am often disappointed by products and services I have purchased** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
Performance	**At school I always received good grades** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
Learning	**I am quick to learn new things** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
Collaboration	**I like to collaborate whenever I can** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely) **I like to collaborate when I can see it is in my interest** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely) **What do you use computers for?** (Tick boxes: Word processing, Presentations, Spreadsheets, Web browsing, Mail, Instant Messaging, Skype, Programming, Games, Virtual worlds (Second Life, There, etc.), … **Something else – what?** [a text box])
Qualifications	**Please list any formal business qualifications** [Text box]
Experience of business in general	**How long have you worked in your current type of business role (in years)?** [Text box]
General attitude towards business	**I like working in business** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)

Attention	
Helpfulness of the agent	**Colette, the agent, is very helpful** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Satisfaction	
General satisfaction with the ICDT platform	**I am completely satisfied with the ICDT platform** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
General dissatisfaction with the ICDT platform	**I am very dissatisfied with the ICDT platform** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
General speed of system response	**The ICDT platform responded quickly enough** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
Assessment of engagement with tasks	**I am easily distracted when using the ICDT platform** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
Feeling of control	**I feel in control when using the ICDT platform** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
Level of understanding of software	**I fully understand how to use the ICDT platform** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
Meeting of pre-existing expectations	**The ICDT platform fully met my expectations** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)
Liking of what the agent looks like	**Colette, the agent, looks great** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Friendliness of the agent	**Colette is really friendly** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Helpfulness of the agent	**Colette is very helpful** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Annoyingness of the agent	**Colette is very annoying** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
How much they think the agent likes them	**I think Colette likes me** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Understandability of the language used	**I understand all the headings on the ICDT platform** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Liking of what the ICDT platform looks like generally	**I like the look of the ICDT platform** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Helpfulness of software response when they made an error	**If I make a mistake when using the ICDT platform it is easy to correct it** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Usefulness of on-screen instructions / help (excluding agent)	**Instructions on the screen about how to use the ICDT platform are really helpful** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Usefulness of paper/verbal	**I fully understood the description of the ICDT platform that was given at the Lidköping meeting**

instructions/other	(Agree completely, Agree, Undecided, Disagree, Disagree completely)
Time taken to ask questions	**I could quickly find someone appropriate to ask questions about the ICDT platform** (Not applicable, Agree completely, Agree, Undecided, Disagree, Disagree completely) **My questions were answered easily** (Not applicable, Agree completely, Agree, Undecided, Disagree, Disagree completely) **Please list who, if anyone, you asked about how to use the ICDT platform** [Text box]
How much users enjoy using the software	**I really enjoy using the ICDT platform** (Agree completely, Agree, Undecided, Disagree, Disagree completely)
Performance (Effectiveness and Efficiency)	
Time spent using vs. learning software	**How much time did you spend learning how to use the ICDT platform, compared to actually using it?** [Text box]
Learning	
Description of what they think they have learned (informal, not a test)	**What are the most useful things you have learned?** [Text box]
Collaboration	
General collaboration	**I collaborated with other students as part of my learning** (Agree completely, Agree, Agree a little, Undecided, Disagree a little, Disagree, Disagree completely)

8.4 Appendix 4 – AtGentSchool – Results of the Heuristic evaluation

8.4.1 General comments

1	A significant amount of text has not been translated into English. This includes some of the agent's speech.

8.4.2 Established heuristics (Nielsen, 2006)

Visibility: Information provided about system state

The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.

1	There is a delay of 7-22 seconds before the smileys have an effect, by which time the agent's actions appear unconnected to use of the smiley.
2	Sometimes the Mind map is read-only. It seems to be when selected from "Workspace / Assignments / Assignments: Countries / Concept Map". When selected from the My Ontdeknet page it seems to be amendable. However, this may not be the real cause.

Familiarity: Match between system and the real world

The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.

1	Before logging in, one can select to read diaries from the login page. To return to the login page the user must click on "portal"; this is not a common term for a login / home page.
2	"My Expert / My experts / My expert / Projects"; What is the blue, purple and green screen for? Is this a mind map the expert created? The key is in Dutch so it's difficult to guess.

Freedom: User control and freedom of choice

Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.

Consistency: The system should be consistent and follow standards

Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.

1	There are pictures of a "talking fish" on the home page, and other pages are decorated with a marine theme. However, the agent is now a boy, rather than a fish, so it's not clear what relevance the marine references are.
2	Before logging in, the login page has a purple box with four "cogs" showing pictures of a "talking fish". This is inconsistent with the use of a boy as agent / assistant.
3	Before logging in, the login page has a purple box with four "cogs" showing pictures of a "talking fish". Clicking once on this box causes the fish to "talk", but no sound or text is shown, causing the user to think that they have missed something.
4	Before logging in, the login page has a purple box with four "cogs" showing pictures of a "talking fish".

	Clicking once on this box causes the fish to "talk", clicking again takes the user to a diary entry. There is no apparent purpose to requiring two separate clicks for navigation here.
5	Before logging in, the login page has a purple box with four "cogs" showing pictures of a "talking fish". Having four separate pictures implies four selectable functions, but all four "cogs" seem to do the same thing.
6	Before logging in, one can select to read diaries from the login page. This produces a set of four tabs (Portal / About Me / Diary / Forum). Underneath the Portal tab there is a barely visible tab. Clicking on this tab (with difficulty) tells the user to login. This tab should be properly visible.
7	Before logging in, one can select to read diaries from the login page. To return to the login page the user must click on "portal" at the top-right of the screen. One of the tabs is also labelled "Portal", but clicking on it does something different.
8	"Introduce yourself" allows insertion of a media file. The button for this is termed "Edit". This suggests changing existing data. However, there is no data as yet. A better term would be "Create" or "Insert".
9	In the Media Library, there are buttons "Edit" / "Delete" / "New"; while "New" looks different to the others, it is not apparent which buttons are active and which are not.
10	The terms "Concept map" and "Mind map" are used interchangeably. However, they do not mean the same thing. ("Concept map" is an interlinked web of concepts created by an expert; "Cognitive map" is one created by a learner; "Mind map" is a proprietary term used by Tony Buzan with one key concept and multiple subsidiaries. AtGentSchool appears to be using a Mind map.)
11	If I select the Mind map from the main screen, it appears under the "headings" of "Workspace / Assignments / Assignments: Countries / Concept Map". If I then click on "Assignments" (as in "Workspace / Assignments") I get a list of assignments, including "Country", which I need to select in order to see the "Concept Map" tab again. This is not at all clear. It seems at this point (before I select "Countries") that there are two possible screens available at the "Workspace / Assignments" level.
12	Generally, if the home screen jumps the user to a sub-sub-sub-level within the software the user will not know where they are, how to get there again (apart from repeating the "jump") or what the surrounding levels are for. It also removes the context from the level they are at.
13	"My Ontdeknet / My Expert" jumps to a completely different place to the main "My expert" tab.
14	"My Ontdeknet / My Expert" jumps to a place that seemingly cannot be reached using the menus (unlike the other jumps).
15	"My Ontdeknet / My Expert / Diary / List" lists the diary entries. It is not possible to select an item on the list to view.
16	"My Ontdeknet / My Expert / Signup" shows a page entitled "Sign on", with text "Are you sure you want to sign on" and a button "Sign off". What does this mean? It is not the same as logging off. Did I sign on, up or off? From what?
17	"My Workspace / Assignments / Assignments / Export To ???" exports the assignment text to a word processor. This does not fit with the remainder of the functionality, e.g. nothing else appears to use the word processor; it is not clear how to import the text back to the system after use.
18	"My Workspace / Assignments / Assignments: The decision / Assessment"; I already submitted the assignment to the teacher from a different screen. This is a second place where I can submit it – also, since it's already submitted, it should say so, rather than let me submit it again (as it did once I submitted it a second time from this new place).

Error prevention: The user interface should prevent mistakes

Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.

1	After submitting assignments to the teacher, there is no "undo".

Recognition: The use of the system should be based on recognition rather than recall

Minimise the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.

1	There is no apparent information on the function of the check boxes in the mind map.

Flexibility: The use of the system should be efficient and flexible

Accelerators -- unseen by the novice user -- may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.

Aesthetics: Aesthetics and minimal design should be considered in the user interface

Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.

Error messages: The system should help users to recognise, diagnose, and recover from errors

Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.

1	If a login fails the message “Add multi students” is displayed.

Help: The system should provide sufficient help and documentation

Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

1	The Help system is not in English (so cannot test).
2	It’s really not that clear what one should be doing after you get to “Discover”. In particular, there are a number of functions which are not mentioned on the home page, or by the agent, such as “Portfolio” and “Discussion”. What are these for? Should I use them?
3	“My Workspace / Assignments / Assignment / Move to Portfolio”; Why would I want to move an assignment to Move to the “Portfolio”?

8.4.3 Additional AtGentive heuristics

Key Indicator One: Attention

Distraction is minimised

The user should not be interrupted in their task, unless the interruption assists that task significantly or is justified by the importance of the interruption. Where appropriate, interruptions should be delayed until the user is less busy. Any animated agent should not be unduly distracting

1	The agent tells me “A good question is an open question…” but I am on the Workspace / Assignments / Assessment page, which has nothing to do with asking questions. (Maybe I should be on the asking questions page, but that’s not clear).

Success in attracting attention

Where the system attracts the user’s attention, it should do so in a manner that will not be accidentally overlooked or misinterpreted

1	The agent’s speech bubble disappears rather quickly. Maybe the first speech text after a while should stay on the screen longer to allow the user to focus on it. It would be better with audio as well.

Key Indicator Two: Performance (Effectiveness and Efficiency)

Task is performed well

Interventions should not cause a task to be performed less well overall. Where the intervention is intended to improve the performance of a task, it should do so

Key Indicator Three: Satisfaction

Overall satisfaction

All suggestions / interventions made by the system should appear to the user to have at least some effective purpose. The user should not consider any suggestion to be “pointless” or “stupid”

Positive image of the animated character

The user’s immediate reaction to seeing any animated character should be at least neutral and preferably positive. The user should anticipate that the character’s appearance will make their task easier, not more difficult. The user should not have negative feelings about the animated character (threatened, humiliated, etc.)

1	The character often repeats a phrase. For example “Your expert does not know yet what you want to learn…”
2	Sometimes, the character’s head becomes much larger prior to speech text appearing. This seems a kind of threatening action.

User control and freedom

This is an extension to the “standard” heuristic. The user should feel in control of the AtGentive interventions. The user should not be worried that they will be interrupted at any moment, or that they are likely to miss something important

Key Indicator Four: Learning

Learning experience is supported

Interventions should not cause the learning experience to be degraded. Where the intervention is intended to improve the learning experience, it should do so

Key Indicator Five: Collaboration

Collaboration is supported

Interventions should not discourage collaboration. Where the intervention is intended to improve collaboration, it should do so

8.4.4 Other

Apparent bugs

1	It is possible to use parts of the system without logging in
2	The system continually sends and receives data over the internet. This is a problem where bandwidth is limited or restricted.
3	Sometimes a page gives an apparently blank page (white space), or mostly blank but a few millimetres of the page showing at the top. In actual fact, the page has loaded into a very small scroll area just below the tabs. This happens on the test computer enough to be a serious problem if repeated with real users.
4	When the agent talks about doing something “here”, it is not always standing in the correct place, or moves to the correct place after stating that something should be done “here”.
5	“My Workspace / Assignments / Assignments / Export To ???” – The word after “Export to” appears below and partly obscured by the button, and cannot be read
6	After submitting the assignments and selecting “The Decision” from the home page, it shows a list of questions (e.g. “What are the advantages of the country you have researched?). However, the answer boxes are not active so I cannot answer the questions. If there is a legitimate reason for this it should be stated (although I note that the browser consistently gives “error on page”).

8.5 Appendix 5 – AtGentNet – Results of the Heuristic evaluation

8.5.1 General comments

1	HTML generated by the system is not standard-compliant. The W3C Markup Validation Service reports 11 deviations from HTML 4.01 Transitional.
2	System expects a very wide window (1100 pixels) to show content without scrolling. The required minimum window width should be lower to improve accessibility.
3	Some of the views display text “pretty display” at the bottom. It is actually a link that will change the view to a “prettier display” with boxes. However, there is no link to return to the previous view.
4	It is difficult to get overview of the system contents. It takes a lot navigation to find all content, and some of it may very well remain hidden for new users. There should be an overview of the complete information content to help the new users to form a mental image of the system.

8.5.2 Established heuristics (Nielsen, 2006)

Visibility: Information provided about system state

The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.

1	Selected a posting and clicked on “to newspaper”. It is not clear if this means everyone sees it now in the newspaper or just me
2	“Newspaper” items don’t have to be postings, but I can add a posting to the newspaper, and it looks just like the other news, so it’s not clear which newspaper items are also postings and which are not
3	My Community: this takes the user out of the space they are in to a higher level where they can choose a community. This is not at all clear. It just looks like they have switched to a completely different web site
4	Re-created the member that had been deleted using “add”: received the message “An account with this name already exists. This account will be added to the users list of this community, but no new account will be created and the password will not be changed” (note the misspelling). However, on selecting “Save” received the message “…is now a new Member of STC V4”

Familiarity: Match between system and the real world

The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.

1	“Cybrary” is not a word in general use in the business community. It is not listed on dictionary.com
2	“helps” is a non-standard use of English language, and would usually be “help”
3	“XML feed” is not a term in general use in the business community
4	Community/Members/Tools click/Community (bestiario / bestario): “bestiario” and “bestario” are not in dictionary.com
5	Community/Members/Tools click/Community (bestario): some text appears in grey for a short while as the display is built up. This text is not in English
6	Community/Members/Tools click/Community (bestario): what is this displaying and why?
7	Agents: “Info” on a proposed intervention displays a page of technical information that the business

	user may not understand (e.g. “MOOD_STRENGTH”)
8	“Portlet” is not a word in general use in the business community. It is not listed on dictionary.com
9	Action log: The graphs offer an “XML graph description”. “XML” is not a word in general use in the business community
10	Delete a member / search box: Under “Click on your name in the list, then click on the "OK" button”, the other button is labelled “Annuler”; if one does not highlight an item then the error is not in English; the window title is “about: blank”
11	Clicking on the AtGentive logo goes to the internal wiki. This is not intended for the public
12	The title of the “agent off / on” box is misleading. The current status “agent off” is indicated with white text, while the inactive “on” is in red. A more common practice is to indicate the active state with a vibrant colour
13	The Search Module has two search modes: “Full Text” and “Document”. The “Full Text” searches text entered into system, and the “Document” option searches inside a specified document. The expression “Full Text Search” usually refers to an operation where the document content is searched as well.

Freedom: User control and freedom of choice

Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.

1	Home page: when moving boxes around the page with left-click+drag, sometimes the cursor gets “stuck” to the box for no apparent reason, even when the mouse is released; it’s then not clear how to disconnect the cursor from the box.
2	Home page: after moving boxes around the page, refreshing the page moves them all back to their default location. For example, the “news >> latest news” box is by default on top of the “last visitors” box if the window is not wide enough. It will snap back to the default position when the view is refreshed. If the user is allowed to move the boxes, the new positions should be saved
3	Deleted a community member – was not expecting to be able to do this. No “undo”

Consistency: The system should be consistent and follow standards

Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.

1	Home page: clicking on an “unread entry” replaces the home page with that entry, whereas clicking on a news item changes the contents of the small “news highlight” box (which can also look as though nothing happened)
2	Home page: most boxes may be moved as required, but the “last visitors” box may only be moved vertically
3	Personal page: most boxes may be moved as required, but the “bookmarks” box is fixed (and appears on top of the “news” box)
4	Home page: there is a line of buttons just below the line with the date, and one button (“Home”) above the line with the date. However, the “home” button looks like a tab (but acts like a button)
5	Menu: “My communities” appears to go too a very different interface
6	Chat: can change chat to Global and then click on blue icon for large chat window, but this opens as Local, with the tiny window still set to Global
7	“XML pretty display” (e.g. Knowledge / Forum): this shows the same information using a different interface. This could cause confusion. Also, selecting “Categories” seems to use a third interface (if

	the subdirectories were expanded by default it would look more similar)
8	Opening another Chat window into main view will show the user twice in the Chat participant lists. However, closing one of the chat windows will log the user out from the both chats. If the purpose of a new Chat window is to provide another view to chat, then the user should listed only once, and if the purpose is to duplicate the participation, then the logout operation does not do the right thing.
9	The "Home/community" view has a misleading name. It is really more about statistics of system use than view of the community.
10	Some of the views have a question mark on the right end of the window bar. Sometimes it displays a textual help message, and other times it tries to activate the embodied agent. This behaviour should be consistent.

Error prevention: The user interface should prevent mistakes

Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.

1	The time is shown using the wrong time zone (i.e. one hour ahead). It also only updates when the screen is refreshed. Showing the wrong time may create errors
2	Home page: the date and time is shown using very small font in cream-on-white and is difficult to read accurately
3	Community/Members/Tools click/Community (bestario): the names and pictures seem to run away from the cursor, making it difficult to click on the one intended
4	The title of Home view acts as a current path in the system hierarchy, but it is visually similar to other view titles.
5	The "Forum Exchanges (byEnhanced)" view has a confusing menu bar. The first item is an icon with a check mark, and it is actually a button. The next item is a static text "Select and"? Perhaps the intended use is to select items in the view and then apply a command from the drop-down menu to them. This is far too difficult for users to figure out.

Recognition: The use of the system should be based on recognition rather than recall

Minimise the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.

1	Three icons in the top bar of the Chat window are not self-explanatory and do not have tooltips or any other documentation. The first icon presumably opens a virtual meeting space, the second one opens a larger Chat window into Home view, and the third one displays a list of pending chat messages. They all should have recognizable icons.
2	The "Show options/Hide options" button has the same icon as the "Play sound" and the "Show icon" buttons. However, the functionality is quite different – the first button toggles the button row, but the other two insert sounds and emoticons to the chat stream. The buttons should have distinct icons.

Flexibility: The use of the system should be efficient and flexible

Accelerators -- unseen by the novice user -- may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.

1	My communities: there is nothing to suggest that clicking on the picture is the entrance to the chatrooms
2	The Search Module does not have a default action for return/enter key press. A common default in this kind of a situation is to initiate the search.
3	Selecting one of the proposed interventions does not take to the actual message. Instead, the user sees a “Proposed intervention” box where is a link to the message (“check this resource”) and choices “validate” and “discart” (should be “discard?”).

Aesthetics: Aesthetics and minimal design should be considered in the user interface

Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.

1	Home page: in the “personal >> my last” the hyperlinks for “... (22)” and “all unread entries” appear to do the same thing.
2	The “Local/Global” and “Local/Global/User” popup menus are confusing. Why there are two of them, and how do they relate to each other?
3	A lot of menu items in control menu are duplicated in the sub-menu of Home page (members and Home/community/members, my communities and Home/community/my communities, logout and Home/community/logout etc.), but not all of them. This is potentially confusing for the new users.
4	The “agents >> pending interventions” box has a lot of static text: “intervention There is a lot of activity around the message “<message>”. You should maybe read it.” Extracting the relevant information is slow because of the repetition.

Error messages: The system should help users to recognise, diagnose, and recover from errors

Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.

1	When a search does not return any documents the error “ No documents found.” is given (the same strange characters appear on all headings when a document is found)

Help: The system should provide sufficient help and documentation

Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

1	The rollover for the EC logo at the top of the screen states “sponsored by the European commission”. Does this mean ITM is sponsored by the European commission, or AtGentive? To me it implies ITC. If this is not the case then it is misleading

8.5.3 Additional AtGentive heuristics

Key Indicator One: Attention

Distraction is minimised

The user should not be interrupted in their task, unless the interruption assists that task significantly or is justified by the importance of the interruption. Where appropriate, interruptions should be delayed until the user is less busy. Any animated agent should not be unduly distracting

Success in attracting attention

Where the system attracts the user's attention, it should do so in a manner that will not be accidentally overlooked or misinterpreted

1	Watch: The Watch page is mostly blank space. If one scrolls down there is the actual watch list
2	When someone starts a chat, it is very easy to not notice
3	The agent does not do anything unless the user specifically switches it on

Key Indicator Two: Performance (Effectiveness and Efficiency)

Task is performed well

Interventions should not cause a task to be performed less well overall. Where the intervention is intended to improve the performance of a task, it should do so

Key Indicator Three: Satisfaction

Overall satisfaction

All suggestions / interventions made by the system should appear to the user to have at least some effective purpose. The user should not consider any suggestion to be "pointless" or "stupid"

Positive image of the animated character

The user's immediate reaction to seeing any animated character should be at least neutral and preferably positive. The user should anticipate that the character's appearance will make their task easier, not more difficult. The user should not have negative feelings about the animated character (threatened, humiliated, etc.)

User control and freedom

This is an extension to the "standard" heuristic. The user should feel in control of the AtGentive interventions. The user should not be worried that they will be interrupted at any moment, or that they are likely to miss something important

1	Some of the agent's speech text disappears before there's been enough time to read it

Key Indicator Four: Learning

Learning experience is supported

Interventions should not cause the learning experience to be degraded. Where the intervention is intended to improve the learning experience, it should do so

Key Indicator Five: Collaboration

Collaboration is supported

Interventions should not discourage collaboration. Where the intervention is intended to improve collaboration, it should do so

8.5.4 Other

Apparent bugs

1	Selected Search/Document/"N" then "News concept similar to newspaper concept" and typed in "newspaper" as the search string. No documents were found. Clicked on "back" and did the same thing but selected "newspaper" from the previous search term list. This time the document was found
2	Agent: The proposed intervention did not change after I used the "check this resource" and "Validate"
3	Profile/Activities/Logged into: Person has logged in twice; graph has Y-axis labelled as 0-1-1-2-2-3 and shows a bar from 0 to about 2.1
4	Profile/Activities/Logged into: Heading states "Click on the points to see the created documents" but clicking on a bar shows the login times
5	Action log / Type of actions (and others) / XML graph description (also "XML" at top of page): Gives the error "XML Parsing Error: not well-formed"
6	Action log / Total creation by member (and others) / XML graph description (also "XML" at top of page): Gives the error "This XML file does not appear to have any style information associated with it. The document tree is shown below."
7	Forum: The X and "show commands" toggle the command line view. However, one must click on these twice before anything happens
8	Personal (bottom of left hand frame): There is a "logout" option at the bottom, but it just appears as a grey-blue bar (possibly it's designed for a bigger screen)
9	Chat: When the big chat window is opened one appears as being on the platform twice
10	Chat: When the big chat window is opened the text colours change; it would be best if my text was one colour and other people's another (or one each)
11	The "Full Text" search does not include document titles in the search

8.6 Appendix 6 – AtGentSchool – Pre- and post-test questions

The 17 pre-test questions (in Czech and English) are as follows:

	Pre-test questions	**(in English)**
1	Neoblíbenějším sportem na Novém Zélandu je rugby	The favourite sport in New Zealand is rugby (true)
2	Na Novém Zélandu jsou tři hlavní ostrovy	There are three main islands in New Zealand (false, there are two)
3	Na Novém Zélandu se mluví německy	People speak German in New Zealand (false)
4	Královna Nového Zélandu je i královna Anglie	The Queen of England is also Queen of New Zealand (true)
5	Kiwi je zélandský ještěr	Kiwi is a lizard on NZ (false)
6	Nejbližší zemí k Novému Zélandu je Austrálie	The closest country to New Zealand is Australia (true)
7	Hlavním městem Nového Zélandu je Wellington	The capital of New Zealand is Wellington (true)
8	Když je na Novém Zélandu zima, u nás je léto	When it's winter in Czech Republic, it's winter in New Zealand as well (false)
9	Původní obyvatelé Zélandu jsou Maorové	The original inhabitants of New Zealand are the Maori people (true)
10	Na Zélandu jsou fjordy jako v Norsku	There are fjords on NZ as in Norway (true)
11	Na Zélandě nejsou žádné sopky	There are no volcanoes on New Zealand (false)
12	Maorové se naznak pozdravu otírají nosem	The Maori people give kisses with their noses (true)
13	Nový Zéland je spojen s Austrálií tunelem	New Zealand is connected to Australia by a tunnel (false)
14	Na Novém Zélandu není skoro žádná příroda	There is almost no nature on New Zealand (false)
15	Kolem Nového Zélandu nikdy neplují velryby a delfíni	There are no whales or dolphins around New Zealand (false)
16	Pták kiwi jí jenom kiwi plody	The Kiwi eats only kiwi (false)
17	Novozélanďani slaví upálení JanaHusa	People commemorate the execution of Jan Hus in NZ (true)

The 15 post-test questions (in Czech and English) are as follows:

	post-test questions	**(in English)**
1	Novy Zeland ma dva hlavni ostrovy (true)	New Zealand has two main islands
2	Moa je velke zvíre podobne tigrovi (false)	Moa is a big animal resembling a tiger
3	První obyvatele Zelandu prisli z Gronska (false)	The first settlers were people of Greenland
4	Nejvyssi horou Zelandu je Mount Cook (true)	Mt. Cook is the highest mountain on New Zealand
5	Kiwi je narodnim ptakem Zelandu (true)	Kiwi-national bird of New Zealand
6	Ze Severniho ostrova na Jizni se dostanete autem (false)	You can go by car from South to West Island
7	Nejvetsim skudcem Noveho Zelandu je possum (true)	The biggest pest of New Zealand is possum
8	Puvodni obyvatele Zelandu byli Anglicane (false)	Original inhabitants were English
9	Vic obyvatelu zije na Jiznim ostrove (false)	Most people live on South Island
10	Maorove maji svuj vlastni jazyk (true)	Maori have their own language
11	Hlavnim mestem Zelandu je Auckland (false)	The capital of New Zealand is Auckland
12	Nejblizsi kontinent k Zelandu je Australie (true)	The closest continent is Australia
13	Expert cestoval po Zelande hlavne na kole (false)	The expert travelled around New Zealand by bike
14	Nejvyssi stavba na Jizni polokouli je postavena na Novem Zelandu (true)	The highest building in the Southern hemisphere is built in New Zealand
15	Hora Mt Taranaki je zaroven sopkou (true)	Mt. Taranaki is also a volcano

8.7 Appendix 7 – Summary of Academic Dissemination Activities

8.7.1 White papers / conference submissions / publications:

Angehrn, Albert A.; Mittal, Pradeep Kumar; Roda, C., & Nabeth, T.: Using Artificial Agents to Stimulate Participation in Virtual Communities. IADIS CELDA (Cognition and Exploratory Learning in the Digital Age) conference, Porto, Portugal, 14 - 16 December 2005

Clauzel D., Roda D., Ach L., and Morel B.: Attention Based, Naive Strategies, for Guiding Intelligent Virtual Agents. Proceedings 7th International Conference on Intelligent Virtual Agents (Poster section). 17th - 19th September 2007, Paris France

Clauzel, D., Roda, C., Stojanov, G.: Tracking Task Context to Support Resumption. Proceedings of the HCI 2006 workshop on computer assisted recording, pre-processing, and analysis of user interaction data. London, UK. 12.9.2006, pp.43-54.

Clauzel, D., Roda, C., Stojanov, G., Mind-prosthesis metaphor for design of human-computer interfaces that support better attention management. Proceedings AAAI 2006 Fall Symposium on "Interaction and Emergent Phenomena in Societies of Agents", Arlington, Virginia, May 26-29, 2007.

Laukkanen, J., Roda, C., & Molenaar, I.: Modelling Tasks: a Requirements Analysis Based on Attention Support Services. Proceedings of the Workshop on Contextualized Attention Metadata: personalized access to digital resources CAMA 2007 at the ACM IEEE Joint Conference on Digital Libraries June 17-23, 2007 - Vancouver, British Columbia, Canada

Maisonneuve N. (2007); Application of a simple visual attention model to the communication overload problem; Attention Management in Ubiquitous Computing Environments (AMUCE 2007), Innsbruck

Molenaar, I., P.J.C. Sleegers, C.A.M van Boxtel: The Effects of the Constructional Nature of Task on the Learning Process and Learning Outcomes; Explanation of the experimental setting. Toogdag 2006 Amsterdam.

Molenaar, I., Sleegers, P.J.C., van Botel, C.A.M: Scaffolding metacognition in collaborative learning with an virtual agent. Proposal for the SIG Metacognition 2008

Molenaar, I., Roda C.: Attention management for dynamic and adaptive scaffolding. Pragmatics & Cognition (Technology & Cognition series), 2008 (Expected). UNDER REVIEW

Nabeth T.; User Profiling for Attention Support for School and Work; Book Chapter in Mireille Hildebrandt and Serge Gutwirth Editors (2008), Profiling the European Citizen; Springer

Nabeth T., Karlsson H., Angehrn A. A., Maisonneuve N.; A Social Network Platform for Vocational Learning in the ITM Worldwide Network; IST Africa 2008, 14 - 16 May 2008, Windhoek, Namibia; Submitted.

Roda, C.: Supporting Attention with Dynamic User Models (extended abstract). Proceedings Interactivist Summer Institute 2007, Paris

Roda, C., Nabeth, T.: The AtGentive project: Attentive Agents for Collaborative Learners. First European Conference on Technology Enhanced Learning EC-TEL'06 (2006), Crete, Greece.

Roda, C., Nabeth, T.: The Role of Attention in the Design of Learning Management Systems. IADIS International Conference CELDA (Cognition and Exploratory Learning in Digital Age) (2005) Lisbon, Portugal, pp. 148 - 155.

Roda, C., Nabeth, T.: Attention Management in Organizations: Four Levels of Support in Information Systems Book title to be defined, A. Bounfour, Editor. Routledge (Advanced research series in management) 2008 (expected).

Roda, C., Nabeth, T.: Supporting Attention in Learning Environments: Attention Support Services, and Information Management. Proceedings Second European Conference on Technology Enhanced Learning (EC-TEL 2007). 17-20 September 2007, Crete, Greece

Roda, C., Nabeth, T.: Attention Management in Virtual Community Environments. Proceedings Journée de recherche de l'AIM (Association Information et Management) "Innovation et Systémes d'Information" ; October 6 2006

Roda, C., Zajicek, M.: Attention Management in Ubiquitous Computing Environments (Introduction to the AMUCE 2007 Workshop) Ubicomp 9th International Conference on Ubiquitous Computing. 2007. Innsbruck, Austria

Roda, C., Zajicek, M.: Towards Supporting Attention in Ubiquitous Computing. Proceedings of Community Computing workshop (CommCom2007) at International Symposium on Ubiquitous Computing Systems UCS 2007 - Akihabara, Tokyo, Japan (Nov. 25-28, 2007)

Rudman, P. and Zajicek, M.: Agile Evaluation for Attention-aware Ubiquitous Computing Environments. Proceedings Ubicomp 2007 9th International Conference on Ubiquitous Computing. 2007. Innsbruck, Austria

Rudman, P., Zajicek, M.: Artificial Agents and the Art of Persuasion, IAT 2006 - IEEE/WIC/ACM International Workshop on Communication between Human and Artificial Agents, 2006 Hong Kong

Rudman, P., Zajicek, M.: Autonomous Agent As Helper – Helpful or Annoying?, IAT 2006 - IEEE/WIC/ACM International Conference on Intelligent Agent Technology, Hong Kong

Surakka, V., Vanhala, T.: Recognition of heart rate patterns during voluntary facial muscle activations. Submitted for evaluation to appear in Esposito, A., Keller, E., Marinaro, M., and Bratanic, M. (Eds.) NATO Advanced Study Institute on The Fundamentals of Verbal and Non-Verbal Communication and the Biometrical Issue.

8.7.2 Conferences (conferences, presentations, posters):

- Cost B27 meeting, Skopje, Macedonia
- 2006 IEEE/WIC/ACM International Conference on Intelligent Agent Technology
- HCI 2006 - Engage, London, UK
- IADIS International Conference CELDA (Cognition and Exploratory Learning in Digital Age)
- Workshop on Attention at HCI 2007 : Workshop on Attention
- Workshop Pm 2006: Meeting For Technology-Enhanced Learning Projects From Ist Call 4, Luxembourg
- AAAI 2006 Fall Symposium on "Interaction and Emergent Phenomena in Societies of Agents", Arlington, Virginia
- IAT 2006 - IEEE/WIC/ACM International Conference on Intelligent Agent Technology, Hong Kong
- 7th international conference on Intelligent Virtual Agent, September 2007, Paris, France
- Second International Conference on Affective Computing and Intelligent Interaction (ACII2007), Lisbon, Portugal
- Annual IEEE International Conference on Tools with Artificial Intelligence (ICTAI07), Patras, Greece
- AIED2007, Los Angeles
- Presentation of Atgentive at the networking session on "Learning and Cognition in Humans and Machines", IST 2006.
- Presentation at UBICOMP 07 “Application of a simple visual attention model to the communication overload problem”
- Presentation at the university of Amsterdam: “Adaptive scaffolding of Self Regulated learning in an innovated learning arrangement”, June 21, 2007
- Presentation at the Open University of Heerlen: ”Ontdeknet and an Embodied Agent”, June 26, 2007,
- [4] July 12th, 2007, Scaffolding learning with an embodied agent supported with an attention management system, Aied

- Poster - Molenaar, I.: Scaffolding in Innovative Learning with an Embodied Agent supported by an Attention Management System. AIED2007, Los Angeles
- Poster for Online Educa Berlin Exhibition 2006
- Poster for 7th International Conference on Intelligent Virtual Agents (Poster section)

8.7.3 Non-academic level presentations

- CELN members' meeting held in December 2005 and July 2006 (51 schools altogether)
- EU Information Days, Best Practices Section, Prague on 5th and 13th December 2005, Information Agency for Supporting Czech Subjects in EU programs
- Presentation of AtGentive in the networking session on "Learning and Cognition in Humans and Machines" at IST 2006
- Meeting of headmasters of the region of Liberec, November 15, 2007
- A presentation for the Dutch research school of education including the AtGentive developments, Toogdag, VU Amsterdam, October 10, 2006
- a presentation for schooltrainers that implement Ontdeknet in schools including the AtGentive developments „New Developments in educational environment", Zeist, December 12, 2006
- A presentation to design students as an introduction to an assignment around the development of agents „Designing a new agent", Hogeschool Rotterdam, December 21, 2006
- Learning with an expert and an embodied agent, Organisation of Education research, department ICT development, January, 26, 2007
- Presentations for future education activities within the Dutch Schools, including AtGentive agents. Breedband Symposium, Dordrecht. March 1, 2006
- Ontdeknet and agent technologies, for an assembly of Publishers in the Netherlands May 14th, 2007
- Ontdeknet and agent technologies, for the ICT coordinators of collaborating school boards. October 5, 2007
- Demonstration to 20 schools on how an agent can support lessons on computers., Utrecht Lunetten NL, APS-IT Diensten; June, 2nd 2007;
- Presentation of the way lessons can be create and supported by an intelligent agent (AtGentive), Utrecht NL, Surffoundation, June 18th 2007; .
- ArboNed; shown how a web-based application can be supported by an intelligent agent. Utrecht NL, July 31st 2007;
- Rabobank Nederland; shown how a web-based application can be supported by an intelligent agent. Utrecht NL, August 31st 2007
- CELN Infocourse MEDTEL, Project presentation, Prague, 5th June 2007-10-29

8.7.4 Others

- Undergraduate course Advanced User Interfaces will participate in an AtGentive-related experiment.
- Press conference + press release, CELN, 22.3. 2006
- Press conference + press release, CELN, 17. 1. 2007

8.7.5 Web presence

- http://www.atgentive.com/
- http://www.celn.cz/
- http://www.calt.insead.edu/
- http://www.calt.insead.edu/LivingLab/AtGentive/Wiki/
- http://cms.brookes.ac.uk/computing/research/advancedinterfaces/Projects/atgentive.htm
- http://www.ac.aup.fr/
- www.cantoche.com/
- http://www.cs.uta.fi/hci/ESC/research.html

8.7.6 Exhibitions

- **Schola Nova 2006** (13th International Specialized Fair Under the auspices of Ministry of Education, Youth and Sport of the Czech Republic), Prague, Czech Republic
- **Invex 2006** , Brno, Czech Republic

8.8 Appendix 8 – Acceptance of agents' instructions (OBU)

Extended summary

Paul Rudman and Mary Zajicek (OBU)

Introduction and Scenarios

This investigation (Rudman & Zajicek, 2006) looked at the feasibility in practice of agent-provided assistance for two of the AtGentSchool scenarios from those described in deliverable D1.3 – "AtGentive conceptual framework and application scenarios".

The first scenario investigated here relates to one consequence of task complexity and / or multitasking is an increased difficulty in the selection of the most appropriate information or task to attend to in the available time. For example, given a limited amount of time available to perform a task, and two pending tasks of similar urgency but different durations, if one of the two tasks can be completed within the available time and the other one cannot, it is often more profitable to attend the task that can be completed within the available time rather than the other one. These types of time-allocation evaluations are often disregarded in complex multitasking environment. This is particularly noticeable in learning environment or in stressful situations. In the former case, students may not even be able to evaluate the length of time necessary to complete a task and instructors may play an important role suggesting the best activity to be performed in the available time. This may be summarised as "Support to limited time resources allocation". (See example in Figure 30.)

Support to limited time resources allocation

The student starts reading the text for a new lecture. The system recognises that a relevant exercise task was previously interrupted (or that the exercise was previously suggested by the application). The agent also evaluates that the exercise task could be completed within the time available to the student whilst reading the text for the new lecture requires longer than the time available to the student. The system suggests working at the exercise.

Figure 30 - Example of "Support to limited time resources allocation"

The second scenario investigated relates to the concept of a task context. As reported in D1.3 (and also (Czerwinski, Horvitz, & Wilhite, 2004) we found that the definition of task is very much a subjective one and, in order to satisfy the need of different users, in different environments, it is necessary to maintain such a definition as general as possible. For the same reason, AtGentive began with a simple definition of task context as including: (1) all the application windows necessary for completing the task, and (2) the task hierarchy for which the task is either the root or an internal node. This led to the scenario "Restore historical context" as described in Figure 31.

Restore historical context

The system will keep track of the sequence in which the user opens documents (Documents). For every document, a 'list' will be held of the documents that were selected immediately both before and afterwards (I will refer to each of these as a "contextual document" or "C-document").

When a user selects a document the system will look at the last time they opened the same document and offer the user the n (number to be determined) C-documents which (s)he had previously selected immediately before and after the original document.

To reduce the cost of interruption, the user will be offered the additional documents (C-documents) only immediately upon selection of a document. While the user may select one of the proffered C-documents (which will each open in an additional new window), no action need be taken by the user if they so choose.

Figure 31 - Restore historical context

The Investigation

The investigation placed participants in situations where each of these two interventions occurred. Afterwards, they filled in a questionnaire and took part in a short interview, both to elicit their feelings and opinions about the interventions. The purpose of the investigation was to look for potential user-related problems with these specific interventions, so that such problems can be circumvented or minimised as far as possible in any future agent implementation.

The investigation was conducted using low fidelity prototyping tools. The intention was to ensure that any problems found by the "users" were not created by the software interface used by the investigation, rather than the task situation itself. Therefore participants were given a "pen and paper" task, during which they would be put in each of the two situations under investigation.

The domain of herbs and their purported medicinal values was chosen, as this comprises a large amount of well documented and inter-related information, and has been used previously by one of the researchers. A paper-based task places participants in the situations described. The concern is to maximise the balance between helpfulness and annoyance.

Conclusion

In both interventions trialled in this investigation, the source of negative feelings was similar. Where a participant's viewpoint was in some way called into question, without there appearing to be sufficient reason, the result was negative feelings, such as annoyance and frustration.

Results from "Support to limited time resources allocation" show that suggesting to a person that they switch tasks, having just begun a task, will be difficult to achieve consistently without a negative emotional response. Timing is critical in that the suggestion is much more acceptable before the person feels they are committed to that task. However, this may be difficult to achieve in practice. An alternative strategy would be to maximise the believability of the person/agent making the suggestion as this appears to influence the person (related factors, such as trust and likeability may also be relevant).

Results from "Restore historical context" show that timing was not a major issue (although the sooner the information is given the more useful it is). What is important is the possibility that the person may have changed the manner in which they intend to approach the task,

thereby rendering the contextual information out of date. Offering this information may generate a negative emotional response, possibly based on the person not wanting to have the new approach undermined, rather than the information simply being unhelpful.

It is clear, then, that in any human-agent interaction the agent needs to take account of the human's likely feelings towards any intervention. Simply giving information that "should" be helpful, in terms of task efficiency, speed, etc., is not sufficient. This study suggests that a software agent, at least in these situations, should be likable and offer advice that is timely and believable. Above all, it needs to take into account the possibility that the user may know best and not make suggestions that may be to the contrary without backing them up.

8.9 Appendix 9 – Animated agent's gestures verbal discrepancy (OBU)

Revisiting the persona effect: Attentional biases in the interaction with Embodied Conversational Agents

Antonella De Angeli (for OBU)

The content of this appendix is too large to fit in this document, and may be found in Attachment A.

See Deliverable D4-4 Final Evaluation Report - Attachment A.doc

8.10 Appendix 10 – Animated agent's gestures and guidance of user's attention on screen (UTA)

The effects of a computer agent's gestures in guiding a user's attention on screen

Daniel Koskinen, Kari-Jouko Räihä, Harri Siirtola, Veikko Surakka, Kimmo Vuorinen (UTA)

The content of this appendix is too large to fit in this document, and may be found in Attachment B.

See Deliverable D4-4 Final Evaluation Report - Attachment B.doc

8.11 Appendix 11 – AtGentNet eye-tracking study (UTA)

Daniel Koskinen and Outi Tuisku (UTA)

Introduction

The aim of this experiment was to collect data on the focus and shifting of the user's attention while using the AtGentNet platform. The test participants were presented with 5 simple tasks to complete using AtGentNet. These tasks were designed to simulate typical usage of the platform. For test purposes, the ITM Community was chosen due to the high amount of activity within the community.

Test procedure

Participants

Five people took part in the test, three female and two male, between the ages of 24 and 27. All are students at the University of Tampere in their 4^{th} – 7^{th} year.

Participant	Age	Gender
1	24	Female
2	24	Female
3	25	Male
4	27	Male
5	25	Female

Table 18

Procedure

The participants were first given a predefined user name and password to log on to the ITM community platform. They were then briefed about the use of eye-tracking during the test and the eye-tracker (a Tobii 1750 device with 17" screen at 1280x1024 resolution) was calibrated.

Participants were then presented with several tasks. The instructions were presented separately on strips of paper, one at time.

1. You wish to see the latest news items.
2. Find the help feature and read the help available for the current page.
3. Find the profile of Paul Rudman.
4. Edit the section titled "Areas of interest" in your own member profile.
5. Log out of the system.

Results

Task 1: Finding the news

The participants spent between 8 to 55 seconds on the home page, with an average of 26 seconds. None of them had any problems finding the latest news.

Distribution of gaze

On average, the two news item boxes (Figure 1, the two large concentrations of fixations in the upper middle part of the screen) were the first to attract the participants' attention. All of the participants looked at the News highlight box, although the Latest News box collected the most fixations (40 %) altogether. This was mainly due to a very high fixation count in the case of one participant.

The control area on the left attracted 10 % of all fixations, the rest of the fixations were fairly evenly distributed between the ITM logo, Top navigation, Last visitors, and the Chat window. The 'most popular documents' box and the 'Agent on/off' box contained fixations in the case of only one participant. The 'Last unread entries' box attracted no fixations at all.

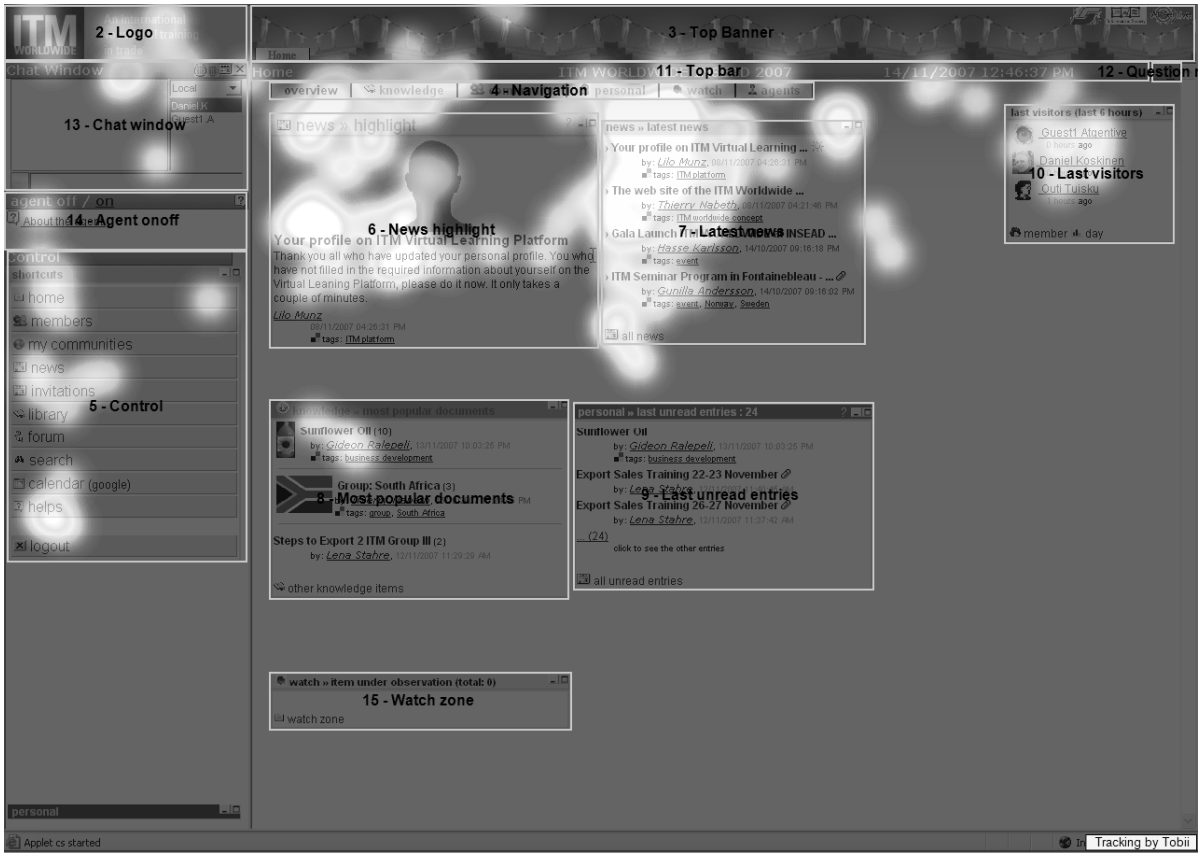


Figure 32 *First view of the home page.*

Task 2: Finding the help feature

In subsequent viewing of the home page (Figure 2) the participants' task was to find the help feature of AtGentNet. All participants sought help via the 'Helps' link in the Control section. In general they were very confused and could not find appropriate help relating to the page they were originally viewing (i.e. home page).

Distribution of gaze

The largest concentrations of fixations are on the navigational areas of the screen: 22 % on the left control box and 10 % on the top navigation bar. Another 10 % of fixations are located on the information bar above the top navigation. It is likely that most of these fixations were recorded due to the proximity of the navigation. There is also a concentration of fixations (10 %) on the 'last visitors' box on the far right of the screen. Despite the nature of the task – looking for the help feature – there were no fixations at all on the question mark link (far right of the screen, below the banner).

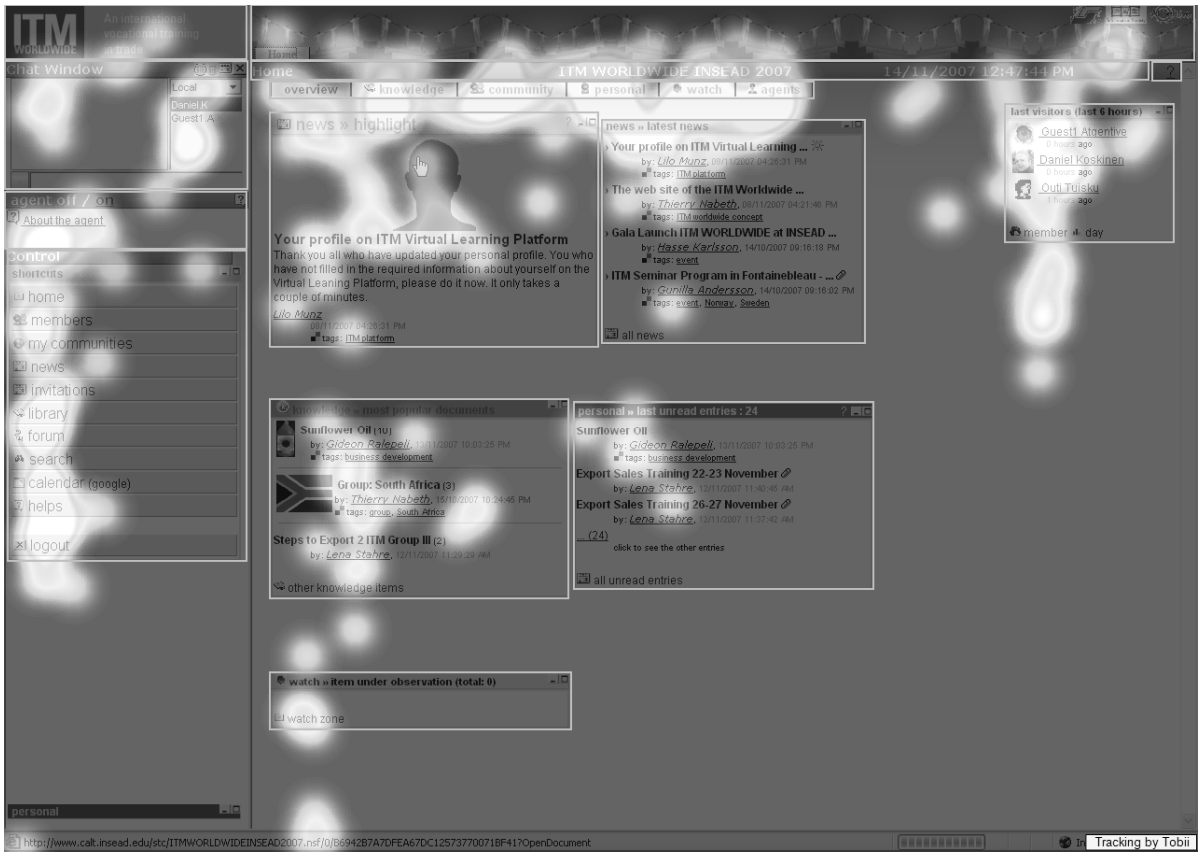


Figure 33 *Second view of the home page.*

Task 3: Finding the member profile of 'Paul Rudman'

Participants 1,2,3 and 5 navigated directly to the 'Members' page after being handed the task. We had chosen a member profile which was not visible on the main Members page on purpose. This was to force the user to search for the profile by other means. On average the four participants spent 54 seconds looking at the Members page before deciding what to do. The shortest time spent looking at the page was 16 seconds, but the participant in question (Participant 3) came back for another 50 seconds after glancing at the 'Search', 'Community' and 'Personal' pages. Participants 2 and 3 proceeded to use the generic Search function to find requested profile, whereas participants 1 and 5 eventually clicked on the 'Members (alphabetical)' link on the right of the screen. Participant 4 did not go the Members page at all but instead went directly to the Search page.

Distribution of gaze

Almost half (49 %) of all fixations were on the central grid containing member profile boxes. 14 % of fixations were on the box to the left listing the last visitors on the site. 7 % of fixations were on the filter box on the right.

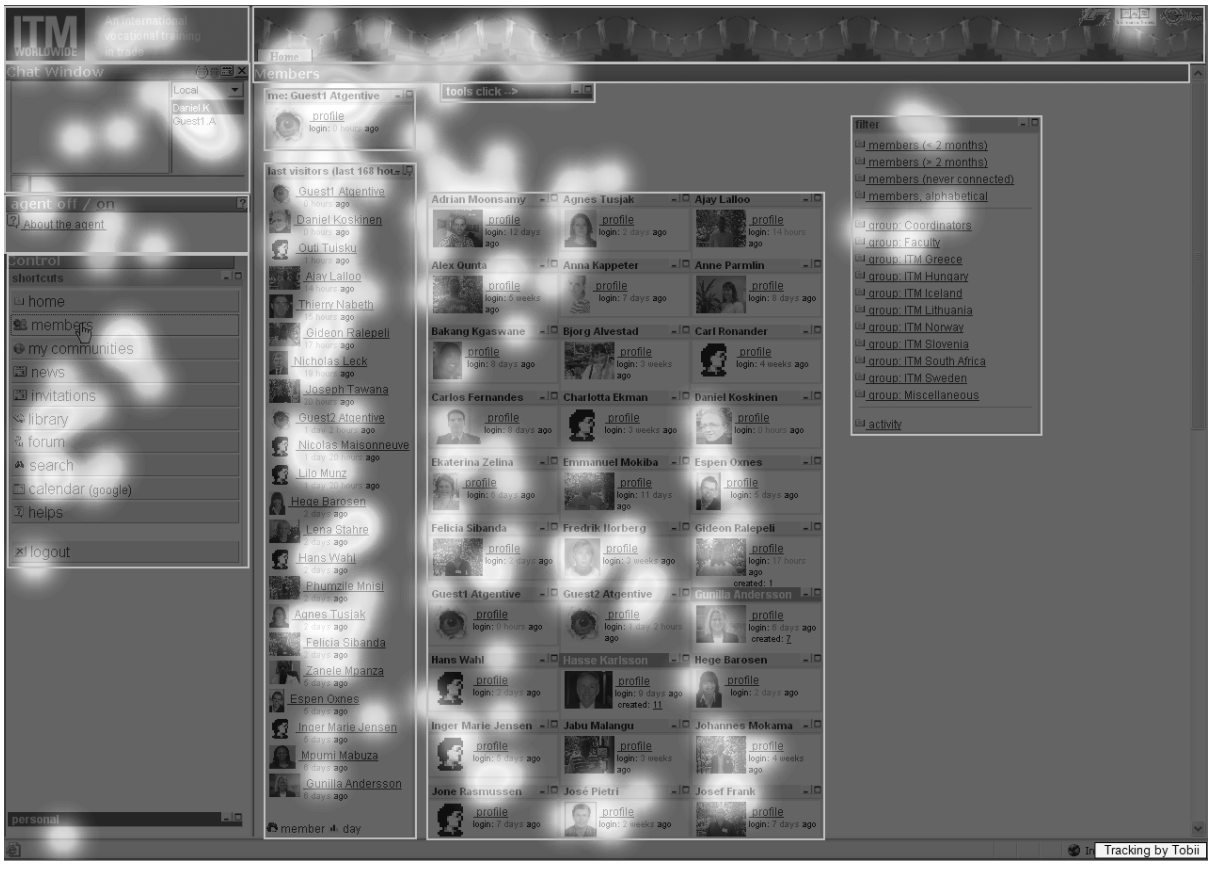


Figure 34 *First view of the Members page*

Task 4. Editing the user profile

The main purpose of this task was to give the participants something to do while a test operator in the adjacent room attempted to attract the participant's attention via the built-in chat window in AtGentNet.

None of the test participants reacted to the chat stimulus, although one (Participant 1) said she had noticed it when interviewed afterwards. However, only participants 3 and 5 had any fixations at all on the chat window.

Distribution of gaze

96 % of all fixations were on the edit form (figure 4), with only the chat window and left control box gathering more than 1 % each.

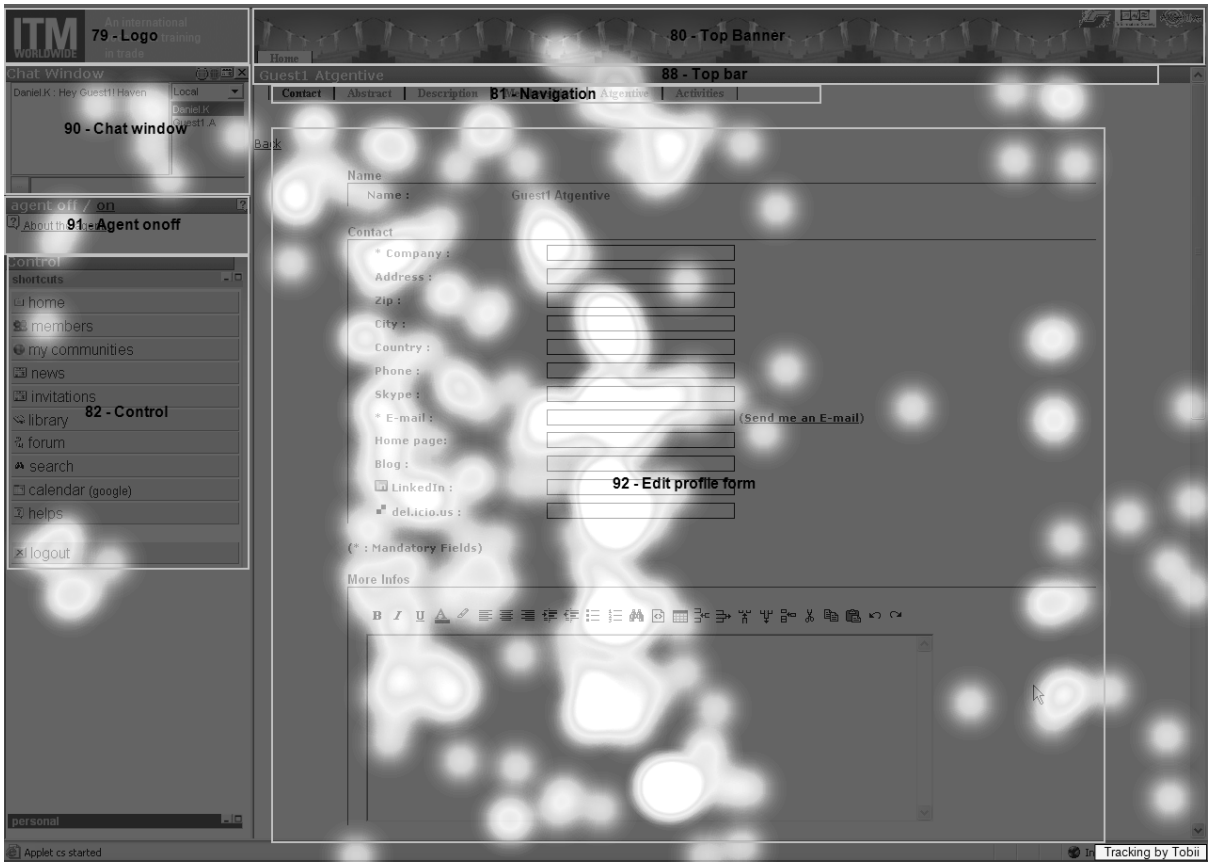


Figure 35 *Editing the user profile*

Discussion and suggestions

The distribution of gaze between different elements of the screen is likely to vary depending on the task. In the first task the participants were asked to look for the latest news, thus it is no surprise that over 60 % of fixations were on news-related items. The fact that there were not many fixations elsewhere apart from the navigation shows that the participants had little difficulty in finding the news items.

On the second view of the home page participants spent a lot of time looking at the navigation links, trying to find the help function of the current page. None of the participants

looked at the question mark link, which activates the contextual agent. This suggests it is hard to find and could be more prominent.

On the Members page participants spent a great deal of time looking at people's profile faces, and did not pay much attention to other elements on the screen. Since human faces tend to draw attention, participants assumed they would find the requested profile among them. They looked for other means of searching only after deciding that 'Paul Rudman' was not among the avatar-equipped profiles on the main Members page.

The results of the fourth task suggest the chat window could be improved upon to better grab the user's attention.

8.12 Appendix 12 – Evaluation of the level of general applicability of the conceptual framework: restoring context (AUP)

By AUP

Introduction

One important aspect of the evaluation of the conceptual framework is to establish its general applicability. The fact that, during the project, we were able to implement the selected concepts in two different applications (AtGentNet and AtGentSchool) already can be taken as a demonstration of such generality. In this section we discuss a further effort in this direction.

One of the concepts that we considered most interesting in the conceptual framework, was context restoration. As discussed in deliverables 1.2 and 1.3, the time required to restore an interrupted task is one of the highest burdens that interruptions bring to current learning and working environments. The concept of context restoration couldn't however be tested in any of the two pilot studies. This was due to the fact that the two pilots concentrated on two individual applications whilst the effects of support to context restoration are most useful when dealing the user working on several different applications, or several devices.

This section reports the experimental work we are carrying out in order to evaluate the effects that support to context restoration could have in multi-application environments. To this end we are running an experiment that compares the effects of interruptions on users who are completing a fixed set of tasks using either a classic graphical interface (control group) or using an interface that supports context restoration (experimental group). The interface supporting context restoration is a multi-desktop interface that allows users to retrieve the context of interrupted tasks intact (as they left it when interrupted) on task resumption. Our objective to evaluate whether through such interface, support to context restore will enhance both performance and user satisfaction. The experiment is being run at the time of this writing we are therefore currently unable to report any results, we detail however the experiment design, the hypothesis that we will be testing, as well as some preliminary observations on the data so far collected.

Measuring the benefits of support to context restoration

Our experimental hypothesis is that if users had the possibility to easily restore the work context of interrupted tasks, then the negative impact of interruptions on their activity would be significantly mitigated.

In order to verify this hypothesis we have designed and experiment in which users are asked to complete three tasks: copying a list of numbers in a spreadsheet, translating some words with the help of a dictionary, answering simple questions on a short movie. Users are asked to complete the tasks quickly but making sure that they make no mistakes. They are asked to start working on one specific task and they are interrupted at predefined times corresponding to specific states of advancement in their work, e.g. after they have translated 6 words, or after they have answered 2 questions about the movie. Each interruption redirects users to a different task, for example, as the user is working at inputting numbers, one interruption will require that he/she moves on to work at the translation task. Users know that they have to swap tasks when requested to do so. Each time an interruption redirects a user to a task that had been previously interrupted we record the time passed between the acknowledgement of

the interruption and the task context restoration. The experiment is being run in two environments, the control group performs the tasks in a classic window environment on GNU/Linux, the experimental group performs the task in a very similar environment but augmented with context restore facilities (this environment is briefly described in section 2). We ask subjects to fill a pre-text questionnaire (this is mainly used to assign subjects to either the control or experimental group), and a post-test questionnaire in which they are asked several questions that we will use to assess possible differences in the personal perception of interruption in the two groups. All subjects receive instructions and a practice session where they rehearse with both the computer environment and the tasks that they will have to perform during the experiment. The activity of the users in the two environments is both logged and videoed.

The hypotheses that we intend to verify are the following:

1 Personal perception

1.1 The users on the enhanced interface perceive work being more pleasant than the users on the traditional one.

1.2 The user feels more productive on the enhanced interface than on the traditional one.

2 Effect on work

2.1 The user completes the required tasks faster on the enhanced interface than on the traditional one.

2.2 The user completes the required tasks with a better quality (less errors) on the enhanced interface than on the traditional one.

2.3 The user resumes faster a suspended task on the enhanced interface than on the traditional one.

3 Effect on interface management

3.1 The user is doing fewer interface management actions on the enhanced interface than on the traditional one.

3.2 The user is doing fewer mistakes in interface management actions on the enhanced interface than on the traditional one.

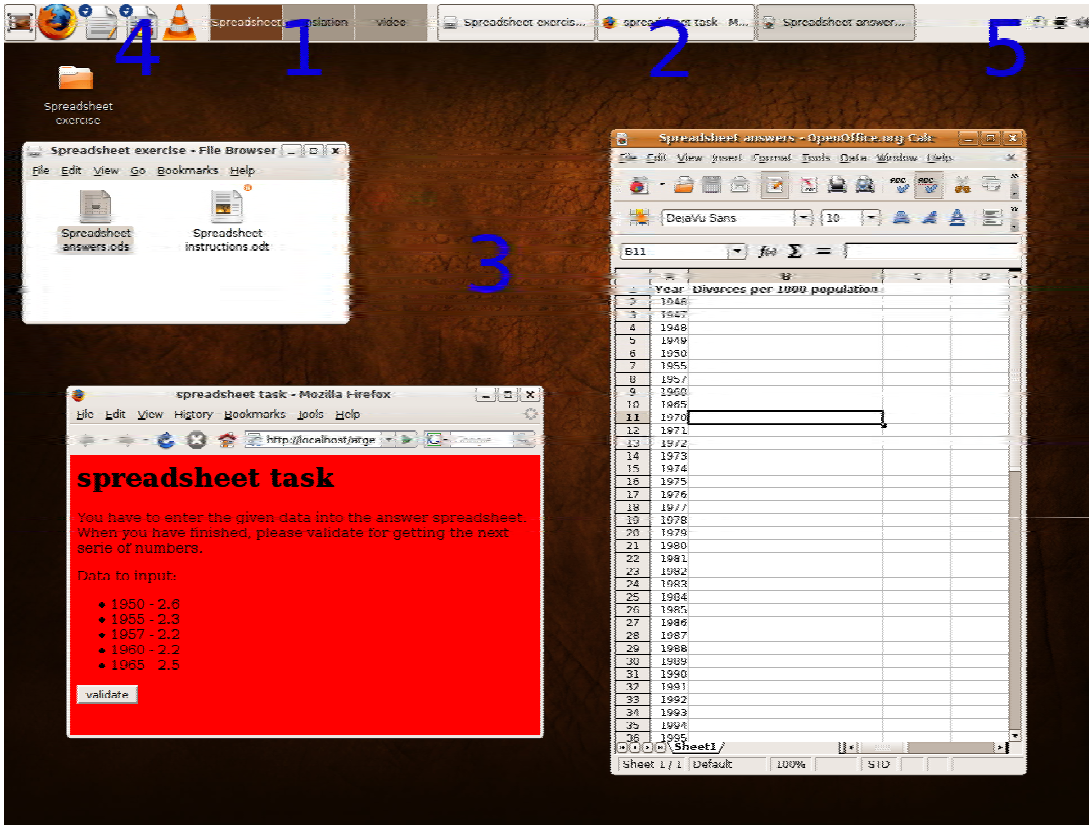


Figure 1: The experimental setup

The computing environment for the experiment

For the experimental environment we used only unmodified software for the applications and we customized through configuration the desktop environment. This was aimed at ensuring that our experiment could be easily reproduced, and at making it possible for the experimental environment to be used as a regular working environment on any computer. We rely exclusively on free software (mainly through the Ubuntu GNU/Linux distribution).

Figure 1 shows our experimental setup. This is a conventional computer interface except that, for the purpose of the experiment, we have eliminated items that are not directly relevant to the user's activity. We use the compiz-fusion implementation of virtual desktops, which allows users to organize their work on several virtual desktops, and to easily swap between desktops. For the purpose of the experiment each task is associated to a virtual desktop and acts as a workspace; a workspace being the collection of all the elements necessary for performing a task: documents, windows, applications, etc. This representation allows us to provide context restoration simply by allowing the user to access the desktop associated to the task being restored. By navigating between the workspaces, the user can browse his tasks, run and suspend them, and get a better insight on the current work. Our hypothesis is that this approach allows the user to "forget" about a task without fearing to loose something in the retrieving process.

The experimental setup is composed of five main elements. They are noted in blue on

1. **Workspace management**: The workspace management widget represents the existing tasks on the computer. The navigation between the workspaces is done by clicking on the target (keyboard shortcuts may also be available). The active workspace is highlighted.

2. **Windows management**: The windows management widget represents the collection of windows existing in the current workspace, regardless of the application they belong to. Navigation between windows is performed by clicking on the target (keyboard shortcuts are also available). The active window is highlighted.
3. **Task resources** (workspace): main working area. It contains open windows, resources, documents, running applications, etc.
4. **Application launcher**: The icons of this area allow users to start an application.
5. **System support**: This area is system related. Here the user can control various non-work elements, like network connection and sound volume.

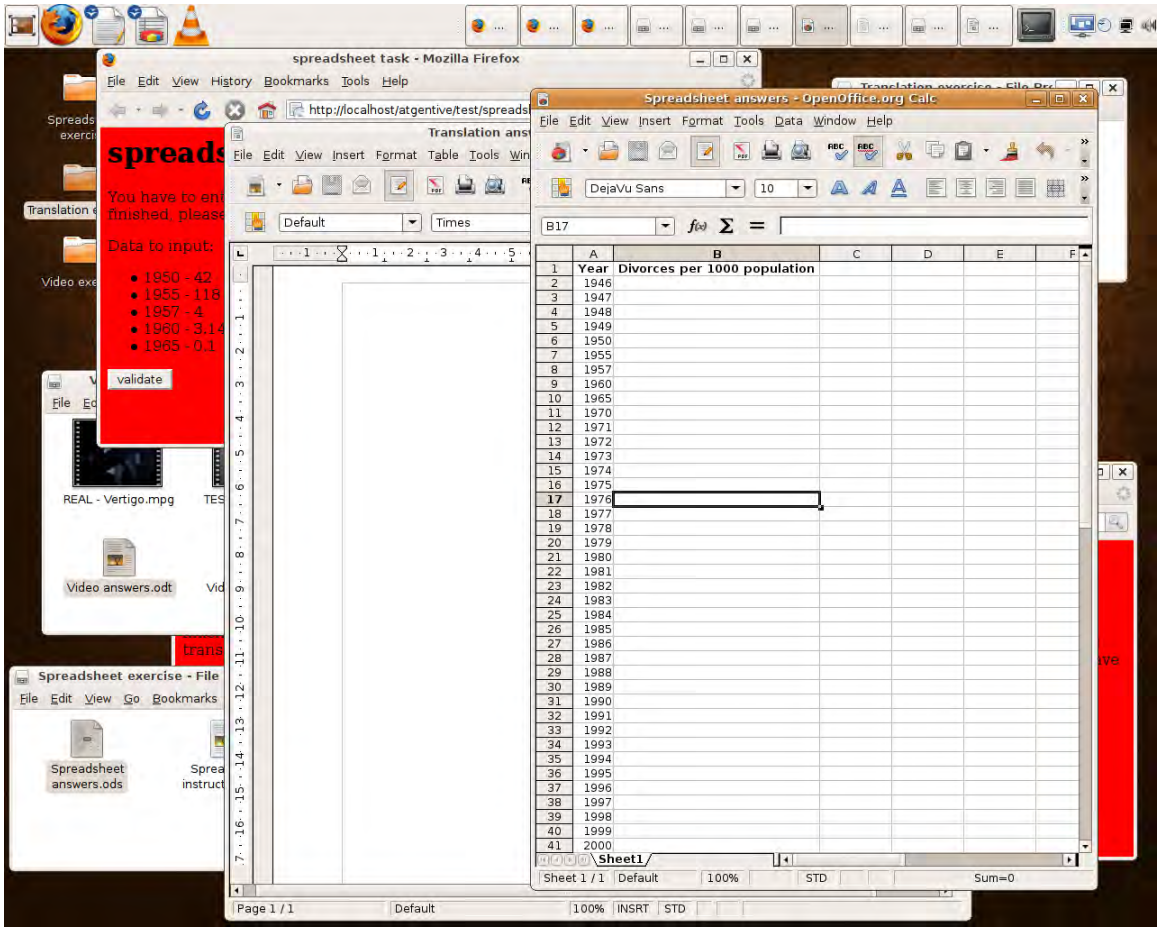


Figure 2: The control environment

Figure 2 shows the environment for the control group. In this case the tasks are the same but they cannot be separated in different workspaces (the environment is a classic window system) therefore application windows from different tasks all appear on the same workspace.

Preliminary observations

The first important result of this experiment is that it is possible, with the current technologies, to implement support to context restoration as described in the AtGentive Conceptual Framework.

The experiment being in progress, it is difficult to make a preliminary analysis based on the small sample we currently have. But by just looking at the partial results, we can see that: task resumption times tend to be significantly smaller in the experimental group than in the control group. Users in the experimental group generally understood easily and used appropriately the tools offered by the interface, they also declared to like to use the interface. Users in the experimental group tend to work faster than users in the control group. Currently, we don't have enough data to allow us to detect differences based on age, gender, lateralization, computer experience, or any other relevant subjects' characteristics.

8.13 Appendix 13 – Evaluation of the level of general applicability of the Reasoning Module (AUP)

By AUP

After the AtGentSchool pilot we are interested in testing how well the Reasoning Module (RM) may support user attention with applications that are not limited to those explored in the course of the project. Also we want to explore the cases in which the RM interacts with several user-level applications. By accepting events from several applications simultaneously, we assume the RM could be capable of supporting the management of attention within a good part of the tasks the user will have to perform on the computer.

For experimenting with the RM in this way, we have implemented a simple *Test Platform* that represents a functional desktop environment. The applications available on the test platform, which could be of various complexity and sophistication, can be started from a start menu and are displayed in windows allowing interaction that is similar to standard desktop environments. Figure 1 is a screenshot of the test platform with a notepad application window visible on the left hand-side. On the right hand-side a second window allows the user to interact with the RM (this part of the interface is discussed later).

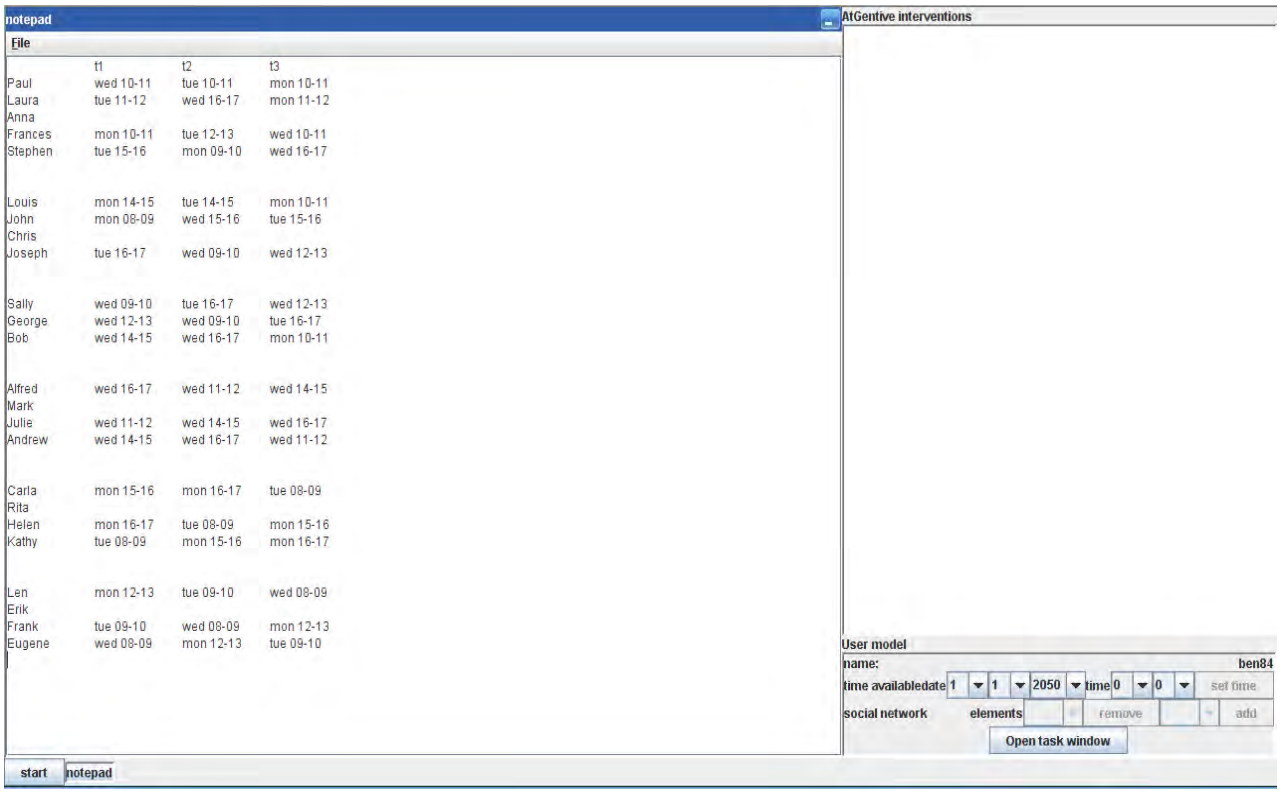


Figure 1. The desktop is divided in two. The user has a view space for applications on the left and a sidebar for interacting with the RM on the right.

The test platform supports applications that conform to one of two simple interfaces depending on whether they are AtGentive enabled or not (the first mentioned being a special case of the latter). Most notably, AtGentive enabled applications differ from normal

applications in that they provide a task model for the user's activities. For the RM to be able to provide assistance among tasks in several applications, the RM must gather a unified view to the task space of the user. For this purpose the test platform combines the task models of several different applications into one[18].

One interesting concept we would also like to evaluate in this context of several AtGentive enabled applications is the interaction with the RM using a sidebar provided for that purpose. This sidebar is visible in figure 2 on the right hand side. The sidebar displays the interventions arriving from the RM and allows the user to interact with the RM and the tasks by letting the user edit the user model and the properties of tasks in the user model, as well as allowing the user to directly *open* those tasks (or those suggested by interventions arriving from the RM) in their associated applications, as shown in figure 2.

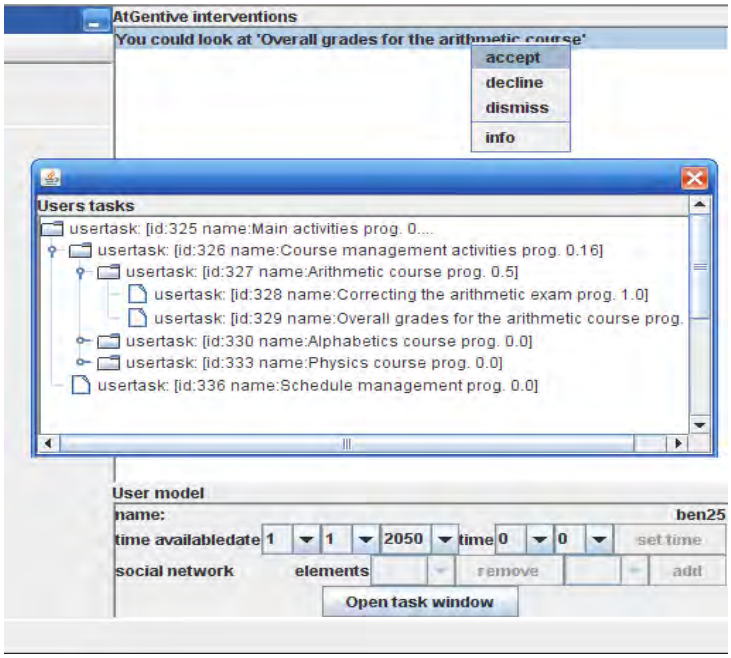


Figure 2. User has right clicked on an intervention to open the task from the context menu. The user also has the task window open with a representation of his task model.

Our main assumption is that the RM could be used to assist the user with his attentional choices among tasks in several different applications. Also we assume that providing additional easy to use, intuitive utilities (the sidebar) for working in the task context (and making task switching fast) as opposed to always interacting with several applications and application windows, will make it easier and more efficient for the users to take use of the services provided by the RM.

Currently, we are using the platform to run an experiment to observe the potential effects of a service supporting users in resuming previously interrupted tasks by lowering the cognitive

[18] A more realistic implementation would be to change the current RM's event model allowing different applications to connect to the RM, which would integrate the task models, as opposed to using an intermediary actor (here the sidebar *module* does this so the RM is not aware it is working in combination with several applications).

load that it takes to remember to do so. In the experiment we want to verify that resuming interrupted tasks will nearly always require cognitive effort, and that displaying a reminder of the *interrupted* task right after the *interrupting* task has been completed will prove beneficial with respect to the time that it will take to resume the interrupted task and the ease with which one accomplishes the resumption.

In the experiment the user will have two main activities. First, in our imaginary *course management system* the users will need to grade their students on exams and on their overall performance in three subjects. The evaluation is based on a set of data describing student's answers to test questions and some instructions on how to evaluate the students' answers. The tasks, whilst mechanical in nature, will require a considerable amount of cognitive effort. In the exam correction activities for example the user will not have a straight example answers to compare the students answers to but he will need to refer to a simple formula given in the instructions to verify if the answers are valid. At pre-specified moments in those tasks, the user will additionally be interrupted and asked to work on a secondary task. In the secondary task the user is asked to build the schedule for a conference. The interruption is a declaration of some requirements for updating the schedule. The user will need to access the view for conference planning and make the changes to the schedule. This secondary task also requires a considerable amount of cognitive load.

After users have completed the *interrupting* task (conference planning) we expect them to return to the previous task (participants will have been asked to complete the primary tasks in order and as soon as possible) and we evaluate the time it takes to resume the interrupted task and also how they accomplish that (e.g. how many steps it takes to get back to the expected task, how long it takes to take the first step, etc.). The different tasks will be part of the same application so resuming the task will always require exactly the same steps (4 steps) whilst the user could make wrong choices along the way. Evaluating the manner in which participants resume the task is therefore an interesting issue because arriving to the interrupted task fast could simply mean the user did so by chance after a quick random search. Resuming the task in a less deterministic fashion would then signal that resuming the task was difficult for the user, even if performed fast.

We will employ a within subjects design in the experiment and the participants will be supported by the RM (in the form of interventions suggesting the user to resume the previously interrupted task) only half the times they are interrupted. Support will also be divided among tasks in some symmetrical fashion and the participants will be divided into two groups to minimize any learning effect. Whilst the experimental setup is fairly simple (users are expected to work on one primary task at a time which quite likely makes resuming the correct task somewhat easier), informal test runs have been encouraging. Signalling efficient use of the service participants have both reacted actively to the interventions when resuming the interrupted task by looking at the sidebar, clearly in search for guidance on where to go, and also, used that information with success by then proceeding swiftly and with ease to the task that they were resuming.

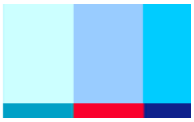
A significant positive effect either in the time it takes to resume the interrupted tasks or in the way those tasks are resumed would effectively show that the service is useful. A speedup in the resumption is something we in fact expect to observe (unless viewing the interventions takes relatively too much time for the small task space, reminding the user of the suspended task should at least not have a negative effect on resumption time). It is much less unclear if we should also expect that not receiving support from the RM will also more often result in the user taking the d-tour to the intended task, behaviour that, in real working conditions, could be seen to often increasing the risk of getting starting more tasks, further increasing multitasking with the cost of adding complexity to the management of attention. Hence, our second goal is to verify this effect of more frequent derailment during the resumption when the user has not been reminded about the task he needs to resume and, without a good concept of his goals, will end up where he ought not be.

8.14 Appendix 14 – Additional pedagogical Analysis (AUP)

The content of this appendix is too large to fit in this document, and may be found in Attachment C.

See Deliverable D4-4 Final Evaluation Report - Attachment C.doc






Revisiting the persona effect:
Attentional biases in the interaction with Embodied Conversational Agents

AtGentive Del 4.4: AtGentive Final Evaluation Report
Appendix A

Dr. Antonella De Angeli

Consultancy Oxford Brookes University SM/2304

Table of Contents

1. Introduction

The objective of Atgentive is to investigate the use of artificial agents for supporting the management of attention in e-learning. One of the possibilities explored in the project is the use of Embodied Agents to coach learners in the management of their attention. This objective is motivated by a large corpus of research on Pedagogical Agents, PA's (Lester et al., 1997 , Moreno et al., 2001), and Embodied Conversational Agents, ECA's[1] (Cassell et al., 2000). Following the *persona effect*, these lines of research claim that the *embodiment* facilitates interaction (Lester et al., 1997 , Van Mulken et al., 1998). The facilitation effect is explained in terms of communication richness. It assumes that the use of non-verbal messages (such as facial expressions or deictic gestures) provides important and natural cues which enlarge the interaction bandwidth without increasing cognitive workload.

Despite the widespread use of PA's in e-learning settings, their success in attracting, orienting and maintaining the learner attention is still controversial (Dehn and van Mulken, 2000, Clark and Choi, 2005). Critics have long ago made a case against ECA's based on the possibility of distracting the user (Shneiderman, 1997). This claim is grounded on the limited information-processing capability of human-beings. Attention works a as a focused spotlight: devoting attention to a target implies detracting it from somewhere else. Although critics have become less and less common in the general ECA literature, there are no definitive answers, as yet, on the effect of animated agents on the learner performance, and the debate about attention is still open.

This paper contributes to this debate by presenting a research methodology and results of two experimental studies to investigate attention distribution between verbal and non-verbal messages conveyed by an ECA. The study adapts the *Stroop task* paradigm, widely used in cognitive psychology, to study executive attention in human-agent interaction. Analysing reaction times on a word comprehension task, this experimental paradigm provides a reliable procedure to understand the effect of non-verbal communication (agent posture and facial expressions) on verbal comprehension (written text). The research focused on the following main questions.

1. Are non verbal messages conveyed by virtual bodies attended to?

2. And, if yes, do they facilitate or inhibit verbal communication?

The report begins by defining attention and reviewing related study focussing on ECA's. This section also describes the stroop task paradigm and its application to a word processing task in human-agent interaction. Next, we present an in depth evaluation of the non verbal repertoire of Colette, the embodied agent designed for the Atgentive project by Cantoche. Experiment 1 and Experiment 2 are reported respectively in presented in section 4 and 5. We conclude addressing the relevance of our findings on the on-going debate on the reliability of the persona effect and proposing suggestions for future work.

[1] Although the two terms are often used interchangeably, PA's and ECA's are different in terms of application context and, often, technological sophistication level. PA's focus on learning, whereas ECA's focus on communication. In this paper, we refer to ECA's as the most general instantiation of anthropomorphic interface design, by which embodied agents are used to mediate the interaction. We refer to PA's when the results are only specific to the learning context.

Key contributions of the report include: 1) a multi-staged conceptualisation of attention to disambiguate controversial findings in the ECA literature; 2) a method for the study of *executive attention* in human-agent interaction; 2) the evaluation of Colette's non-verbal communication; 3) experimental results showing that embodied agents do indeed capture user attention, and can disrupt performance under certain circumstances; 4) an experimental validation of the persona effect.

2. Attention

Attention serves as a set of mechanisms which regulate cognitive processes and feelings. Recent advances in neuroimaging techniques have supported the existence of different cognitive networks relating to specific aspects of attention (Posner and Rothbart, 2007). Three different networks were identified which supports different types of tasks: alerting, orienting and executive attention.

- *Alerting* is the achievement and maintenance of a state of arousal, or sensitivity to incoming stimuli.
- *Orienting* involves the selection of information from a source of incoming stimuli.
- *Executive attention* involves mechanisms capable to monitor and resolve conflicts among incoming stimuli (physical objects and events, thoughts, and feelings).

These three tasks can be conceived as separate steps which lead to information processing. Alerting stays at the button level: it refers to arousal (the subject is ready to receive information). Orienting and executive attention are involved at different stages of the selection of information. Cognitive processes happening at the level of executive attention regulate the contents of working memory (Engle, 2002). Executive attention is the ability to maintain or suppress information in working memory, focussing to relevant parts of the perceptual field, while ignoring tasks irrelevant stimuli. Hence, executive attention is involved in a variety of higher-cognitive tasks underlying intelligence, such as reading and listening, learning, and self-regulation of positive and negative affects.

This distinction of attention as separate networks devoted to specific tasks is important and may help to clarify some of the contradictory results reported in the HCI literature.

2.1 Attention & ECA

In recent years, increasing effort has been devoted to the study of the distribution of the user attention to different elements of the computer interface during task execution (Roda and Thomas, 2006). ECA's are special interface elements, as their anthropomorphic appearance can induce social attributions and biases. The human face is an extraordinary stimulus. Research in psychology has demonstrated an extremely efficient detection of facial expressions, with a particular relevance to threat and fear (Hansen and Hansen, 1988). There is evidence that affective facial expressions are automatically processed and can interfere with other tasks (Stenberg et al., 1998). If the emotion conveyed by a face does not match the emotional valence of a verbal message, understanding is delayed or even impeded. Significant for the design of ECA's is the finding that emotional expressions in a face can be perceived outside the focus of attention and tend to guide focal attention to the location of the

face (Eastwood et al., 2001). Evidence of the importance of consistency between verbal messages and facial animations in human-agent interaction is reported by (Berry et al., 2005). Inconsistency strongly decreased memory for verbal information.

These findings are of utmost importance for the design of ECA's, as the status of the art in graphical rendering of emotions and their synchronisation with timing and content of the verbal message cannot still guarantee a perfect match between the two information channels. In case of conflict, ignoring the agents' facial expression and focussing on the verbal message may be complex. Hence, we may expect that agent interaction in suboptimal conditions can hamper communication.

Although attention is a common dependent variable in many evaluations of ECA's, conclusive evidence is still missing. Several factors can be held responsible for such a lack of agreement (Clark and Choi, 2005, Dehn and van Mulken, 2000, Gulz, 2004). Firstly, empirical research on ECA's suffers from a generalised lack of methodological rigour, affecting operational definitions of core constructs, and, as a consequence, methods and procedures for measuring them. Secondly, most evidence consists of results from single ecological studies, whereas scientific generalisation would require a series of experimental studies. Generalisation in PA research is further complicated by the large variance introduced by testing different learning environments, users, and agent instantiations. Finally, these evaluations tend to address a large set of dependent variables at once, including performance indicators (e.g., learning outcomes), cognitive processes (e.g., attention allocation, memory, problem-solving), motivational and attitudinal measures (e.g., willingness to use, satisfaction). Although ecological studies have potentials in addressing social and motivational variables, they lack the control required by the study of cognitive processes.

2.1.1 Operational shortcomings

At the heart of the persona effect lays the assumption that non verbal cues conveyed by an embodied agent have the potential to guide the user attention towards important elements of the task at hand. Analysing this effect within the conceptualisation of attention as a multi-staged process implies that embodied agents have the capability to *increase alertness*, *support attention orientation*, and that, in doing so, they do not add any demands to the *executive control of attention* (e.g., there is no distraction induced by the increase in stimulation). We believe that this multilayered framework of analysis can help interpret some of the inconsistencies in the empirical research on the role of attention in agent-human interaction.

In human-agent research, attention has been defined in many different ways. For instance, it has been associated to *time* spent performing a primary task, such as playing cards (Takeuchi and Naito, 1995) or filling in a questionnaire (Sproull et al., 1996) while interacting with a virtual face as compared to a control condition (i.e., no face). Although both studies reported longer response time in the face condition, their authors interpreted this effect in opposite ways. Takeuchi and Naito (1995) associated the longer time to distraction. They claimed that the user attention was detracted from the primary task because it focused on interpreting the facial expressions of their virtual opponent. Conversely, Sproull and colleagues (1996) explained the longer time taken by users when answering psychological tests in the face condition as a measure of attentiveness. They concluded that people were paying more attention to the primary task, when the questionnaire was presented by a character than when it was presented by textual display. This conclusion was derived by the association of performance time with increased arousal, which according to the authors fostered self-reflection, thus slowing down the activity. An alternative interpretation could be that both studies implicitly dealt with executive

attention (distribution of attention between the primary task and the agent). Following this line of reasoning, their findings seem to indicate a certain degree of interference between the face and the primary task.

Other studies have investigated attention indirectly via post-test recall of visually attended items (Hongpaisanwiwat and Lewis, 2003) or message content (Berry et al., 2005). Results were controversial. Hongpaisanwiwat & Lewis (2003) reported no differences in memory for items which were pointed at by a non-anthropomorphic agent (a finger), an anthropomorphic agent (a human-like puppet), versus a control condition (no visual pointers). Conversely, (Berry et al., 2005) reported worst performance when the user interacted with an emotion-less virtual face and a virtual face displaying emotions which were inconsistent with the verbal message, versus a number of control conditions (emotion less human picture, voice only and text only condition). However, when the user interacted with an agent which displayed consistent emotions, the performance improved up to the level of the control conditions. Overall, these studies seem to indicate some interference of the agent with the task at hand, particularly when the agent non-verbal behaviour conflicts with the verbal information.

Attention has also been associated to anxiety (Rickenberg and Reeves, 2000). The idea here is that anxiety is a measure of arousal, and hence it indirectly addresses attention. An experimental study demonstrated that participants reported being more anxious (aroused) when interacting with a monitoring agent (an agent which was explicitly paying attention to the users' behaviour), than when interacting with an idle agent (which appeared to be preoccupied by other activities), or with a control condition (no agent). The type of agent was also found to have a significant effect on performance, as participants in the monitoring condition were less accurate in an information retrieval task than participants in the idle condition. These results can be interpreted as a sign that participants paid attention to the agent and that the agent presence affected their psychological status and performance.

Orientation of attention has also been explored overtly by tracking the user eye gaze (Prendinger et al., 2007, Witkowski et al., 2001). Both studies indicated that the agent attracted the user attention, often detracting it from other interface elements. Interestingly, no important differences were found in the agent's capability to orient attention as compared to a text only and a voice only condition.

To conclude, the literature seems to support the idea that embodied conversational agents have some influence on the first level of the attention process: they seem to be capable of increasing the state of alertness of the user. However, their effect on orientation does not seem to be more effective than other pointing devices. The most controversial point remains their effect on executive attention, which is the main topic of our research.

2.1.2 Measurement shortcomings

Attention in ECA's research has been measured by an array of techniques, which can be clustered in three general categories: subjective evaluations, performance measures, physiological monitoring. All of these techniques have well-known potentials and shortcomings which will be presented below.

A large number of studies rely on self-reports (Sproull et al., 1996, Koda and Maes, 1996). When asked, participants have reported that embodied agents attract attention (; van Mulken et al., 1998) and that they do not distract from the task at hand more than other interface features (van Mulken et al., 1998). Questionnaires and interviews

are a very convenient way to address the problem, yet their results may be misleading. Indeed, research on meta-cognition (people's ability to evaluate their own mental process and performance) has demonstrated that self-reports are often unreliable {REISBERG, 1985 #266} {Nisbett, 1977 #267}. People do not know when they are being distracted and are prone to judgement biases when asked to indicate what distracts them. Introspection leads people to use a set of expectations on what influence performance, rather than to perform an objective assessment of their performance. This cognitive bias may be due to the fact that people have little or no access to higher level cognitive processes {Nisbett, 1977 #267}, and in particular may not know when these processes are disrupted by a source of distraction. Alternatively, people may have access to higher-level cognitive processes, but they may be incapable of evaluating properly their outcomes.

Performance-based measures of attention in ECA interaction mainly addressed execution time and errors in a primary task. This procedure has the benefit to provide objective data of the user behaviour. It implicitly follows the dual-task paradigm, commonly used in cognitive psychology to study executive attention, but due to the uncontrolled experimental settings of many ECA evaluation studies, as compared to the typical laboratory settings where this experimental paradigm was originally developed, these measures can seriously be affected by a number of confounding factors, which do not regards attention. Other performance based approach focussed on the analysis of recall. Although attention is a necessary pre-requisite for memory, the opposite is not necessarily true.

Because of these shortcomings, eye tracking techniques has been recently used to evaluate users' attention to ECAs (Prendinger et al., 2007, Witkowski et al., 2001). Yet, research in psychology suggests that this approach may not be the expected panacea to achieve an accurate measurement of attention distribution in ECA interaction. Indeed it has been demonstrated that attention orientation may be accompanied by eye movements, but it can also happen covertly because attention precedes an eye saccade {Rayner, 1998 #244}. Another limitation of eye tracking is that it can only address the level of orientation and does not provide information on the outcome of executive attention processes.

In this project, we aim to contribute to the study of attention in agent interaction by proposing a simple reaction time experiment to investigate not only if participants do look at the agent, but also if the information attended to is elaborated by the user. In our study, we explicitly addressed executive attention within the tradition of the *Stroop task* paradigm. This implied creating a conflict between two information sources and analysing if this conflict created interference in information processing (longer reaction time and increased number of errors). In the experiment we focussed on the effect of non-verbal communication (agent posture and facial expressions) on verbal comprehension (written text), by comparing congruent situations (the two channels provide redundant information) with incongruent conditions (the two channels provide conflicting information).

2.2 The stroop task

Executive control of attention is often studied by tasks which involve conflicts, such as various versions of the stroop task (see (MacLeod, 1991) for a review). In the original formulation of this task, subjects were asked to name the colour of ink in which a word was presented. The word itself could be a colour name, which was either printed in the same colour ink (congruent condition) or in a different colour ink (incongruent condition). Reaction times and error rates were compared against a neutral condition, in which a letter string was presented in coloured ink. In the

incongruent condition (when the ink colour and the word meaning disagree), strong interference was found. The term interference in the stroop task literature is used to denote the difference in reaction time between congruent and incongruent conditions. In the incongruent condition longer reaction time and more errors occurred (people tended to read aloud the name of the word rather than to name the colour). A less reliable, but often observed facilitation effect was also found when the ink colour and word agree.

Performance on the Stroop task relies on executive attention to maintain in memory the primary goal of naming the colour, while suppressing the stronger tendency to read aloud the word (Engle, 2002). The Stroop effect has been explained in terms of automaticity of the response and relative speed of processing. Automatic processes are involuntary; they are induced by a stimulus and run their course to completion once started without conscious intervention. To block an automatic stimulus requires effort and has side effects on the processing of other stimuli, as demonstrated by the increased reaction time in the incongruent condition.

The Stroop task paradigm has been applied to a variety of stimuli combinations. These stimuli include picture-word combinations, and face-word combinations. In the picture-word combination, the word is known to be the strongest stimulus. Indeed, incongruent words printed inside pictures strongly interfered with picture naming, whereas incongruent pictures had only a marginal effect on word reading.

The combination of faces and words is interestingly as it associates two stimuli which are automatically processed. Stenberg and his colleagues (Stenberg et al., 1998) applied the Stroop task paradigm to analyse the effect of faces on a word processing task. They tested compound stimuli, consisting of words superimposed on pictures of affective faces. Participants were given the task of evaluating the affective valence of the word, while disregarding the information coming from the face. Results of three experiments demonstrated an effect of facial expression on word evaluation. Negative words shown within a negative face were facilitated as compared to positive words; vice versa, positive words were facilitated when shown within a positive face as compared to negative words. The author concludes that affective facial expressions are automatically processed and can interfere with other task. We base our study on this assumption and we aim to verify if this effect is maintained even when the face is not human and the word is not superimposed on the stimulus but rather displayed close to the ECA, as in the traditional speech bubble interaction.

3. Evaluation of Colette's non-verbal communication

The first phase of our research involved an in depth evaluation of the expressiveness of Colette's non-verbal communication. This was necessary in order to select a set of non-ambiguous stimuli to be tested in the stroop experiment. Colette provides a total of 47 standard body animations and 13 facial expressions which can be controlled by a viewer application (Figure 1). Selecting one item of these lists or a combination of them, results in Colette performing the action. The basic emotions conveyed by the agent are happiness (positive), sadness (negative), anger (negative), and surprise (neutral).

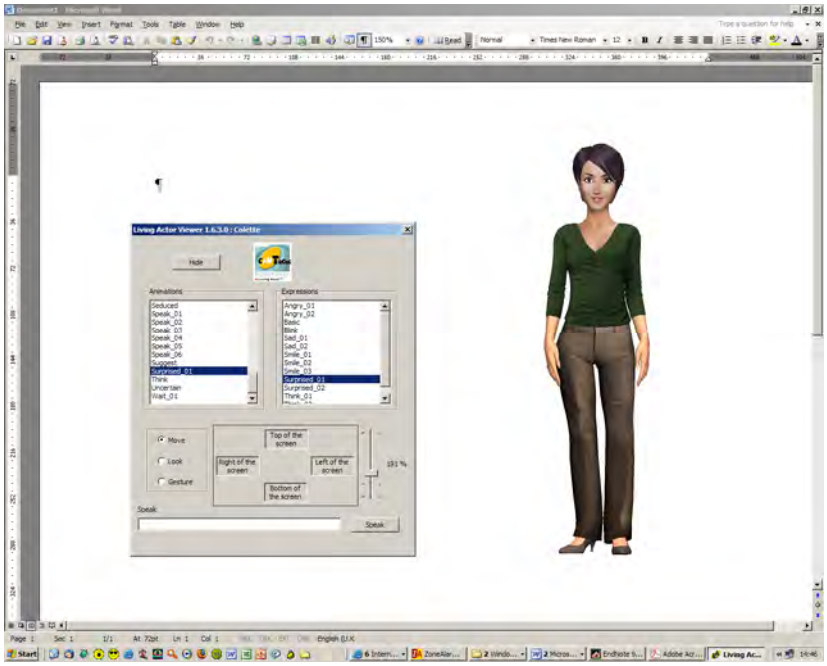


Figure 1. Colette and the animation viewer

The evaluation was performed in two main phases. Phase 1 selected a subset of emotions to be tested by an on-line survey in phase 2.

3.1 Emotion selection

3.1.1 Method

Phase 1 consisted in an expert based-evaluation, supported by well established knowledge on signal characteristics of facial expressions and their link to emotion appraisal (Ekman and Friesen, 1975), and knowledge on affective semantic of body posture (Argyle, 1988). Based on this knowledge, the author systematically analysed the list of animations and expressions of Colette, testing them individually and in every possible combinations.

The expert based evaluation allowed identifying a smaller subset of combinations (N=30) which were tested with 6 people (3 males and 3 females, covering a broad age range, from 25 to 63). The viewer was operated by the experimenter and was hidden to the participant by a cardboard screen taped to the computer screen. To avoid interferences which could be generated by the label of the animation (automatically displayed in a speech bubble), the part of the screen above Colette's head was hidden by a dark tape. Under these conditions, participants were shown the set of 30 animations and asked to indicate if they represented a positive, a negative or a neutral emotion. They were also presented all the individual facial expressions and selected body animations and asked to name them. At the end of the trial, participants were invited to comment and motivate their scoring for all the animations which they had incorrectly labelled. This procedure allowed selecting the 12 stimuli tested in the on-line study, and evincing a number of limitations of Colette's non verbal repertoire.

3.1.2 Results

The analysis evinced some problems related to the usability of the viewer application and a number of specific observations about the emotional expressiveness of Colette. A summary of the major findings are summarised in Table 1.

1	Usability (lack of functionality)	The viewer does not allow any control on the speech bubble displaying the label of the animation activated. The viewer provides an option to add new words but this is active only when the speech-related animations are selected. Hence, there is no way to have a 'sad' Colette saying 'I am sad', for example.
2	Usability (lack of control)	There is no timing control on the combination of animation and expression. They are performed in sequence rather than simultaneously, as it would be required for a clearer emotional communication. Animations are displayed by default with the last selected facial expression. If no facial expression was previously selected the *basic* face is displayed, but this basic face is not really neutral but smiling.
4	Seduction (body postures)	All of the participants in this evaluation made explicit mention to the fact that Colette seemed to have been designed with the main aim to seduce the user with her movements, rather than interacting with them. Two out of 3 males found it amusing, but all the females (N=3) found it annoying.
4	Expressiveness (body postures)	Most of the animations in Colette's repertoire were designed to display actions related to communication. A few of them, however, explicitly addressed emotions, i.e., sadness, anger, happiness, and surprise. *Anger* was not clearly identified and it was labelled as 'teasing' (N=2) 'arguing' (N=2), and 'playing' (N=1). *Happiness* was never properly identified. The problem is that the animation happy-01 (showing the agent crossing its arms around its body) is in shark contrast to the stereotypical representation of this emotion. At least in western countries, happiness is associated to openness and reaching out to others. When exposed to this animation, people commented 'she is offended' or 'she is flirting'. Happy-02 is also very peculiar (the agent raises her hands palm up towards its head while slightly flexing its torso towards the user). No person was capable to recognise this movement as a manifestation of happiness and most of them concluded 'she is surprised'. *Sadness* appeared to be very clear (especially when displayed in combination with the proper facial emotion). However, 2 people said that the agent was 'tired' or that it was 'passing away'. *Surprise* was systematically perceived as being scared. Because of this general mismatch between intended and perceived emotional valence of body animation, we decided to test in phase 2 also animations which were designed to convey communication actions (e.g., congratulation and greetings).
3	Expressiveness (facial emotions)	The facial emotions of Colette were difficult to discriminate. Overall, these facial expressions were mainly rendered by eyes movement. The animation of the lower part of the face (cheeks and mouth) was much less effective. Overall, most of the animations were perceived as positive. This is due to the fact that Colette's tends to be smiling in many emotions. Also her big bright eyes add to this general positive feeling.

		Another major problem in detecting negative emotions in Colette's facial expressions was due to the fact that, when no body animations were selected, Colette's kept performing some small movements which were perceived by all users as aimed at seducing them. Thus, they were perceived as communicating a positive feeling towards the observer, which often mismatched with what was meant to be a negative face. The positive bias was noted by all users and the only emotion which was systematically perceived as negative was SAD-02. *Anger* was very controversial, because of the mismatch between the frowned eyes and the mouth which was systematically perceived as smiling (Figure 2). A male participant, shown the expression labelled as ANGRY-02, commented 'Wow she is angry now, she would be my perfect girlfriend!' No participant perceived the emotion labelled *ANGRY-01* as negative, because in that case the eye movement is much reduced. The expression labelled as *basic* was systematically perceived as happy due to the fact that Colette appear to be smiling. *Blink* was also very controversial, as the agents just closed its eyes. People commented 'she has gone to sleep' or 'she is playing hide and seek'. The emotions conveying s*adness* were easily recognised: all users understood the meaning of SAD-02 and 4 out of 6 that of SAD-01. The others thought she was 'annoyed', or 'thinking'. *Happiness* was always recognised by all users. They also appeared to be capable to recognise the intended increase in emotional intensity, between HAPPY_01, happy_02 and HAPPY_03. Only one participant perceived the meaning of *surprise* correctly. *Think* was labelled as 'anger' or 'concentration'.

Table 1. Summary of the main usability and communication problems

Figure 2 illustrates the two face expressions describing anger in Colette (ANGER-01 on the left, followed by ANGER-02) and compares them with a picture of a real face expressing the same emotion (Seyedarabi et al., 2006). It clearly emerges that Colette fails to convey the full complexity of gestural cues related to this emotion. In particular, both the eyes and the mouth appear to be smiling, rather than expressing the tension denoting anger.

Figure 2. Anger in Colette and in real life

3.1.3 Conclusion

This evaluation evinced several limitations in the expressiveness of Colette's non-verbal behaviour which, on the average, appeared to be quite loosely linked to prototypic representations of emotions. The study clearly suggested the need for

combining both facial emotions and body movements to convey clearer emotions, and to include also animations designed to convey communication action (agree, argue, congratulate, greet). Twelve combinations were identified (6 positives, 6 negatives) which were tested in the on-line survey (Table 2). Note that surprise is tested as a negative emotion, due to the negative comments collected during the test. Most participants perceived the animation and facial expression related to surprise as expressing fear.

Emotional valence	Animation	Expressions
Negative	Decline_01	Angry_02
Negative	Argue_03	Angry_02
Negative	Surprised_01	Angry_02
Negative	Sad_02	Sad_02
Negative	Angry_02	Angry_02
Negative	Surprised_01	Surprised_02
Positive	Greet_01	Smile_03
Positive	happy_01	Smile_03
Positive	Congattulate_02	Smile_03
Positive	Argue_02	Smile_03
Positive	Happy_02	Smile_03
Positive	Congratulate_01	Smile_03

Table 2. The 12 animations selected for evaluation

3.2 Emotion evaluation

To test the reliability of the 12 combinations of body animations and facial expressions, an on-line survey was designed. Short movies were recorded using *Camtasia* to capture Colette's animations produced by the viewer application. All animations were recorded starting from a rest phase, then selecting the facial emotion and, finally, the body animation. The screen capture area was selected in such a way as to cut out the speech bubble accompanying the animation.

All videos lasted an average of 6 seconds and were uploaded on YouTube. The survey was designed using surveygizmo (www.surveygizmo.com). An introductory page briefly explained the purpose of the study. Participants who agreed to participate in the study were then shown 12 pages, each of them displaying a video player and a semantic differential item to evaluate the emotional valence of the animation (Figure 3). Participants evaluation was modulated on 7 points (1 = very negative; 7 = very positive). Participants activated the video by clicking on the image and could replay it as many time as they wanted. Once they had rated the animation, they could go to the next page. The order of the animations was randomised across subjects.

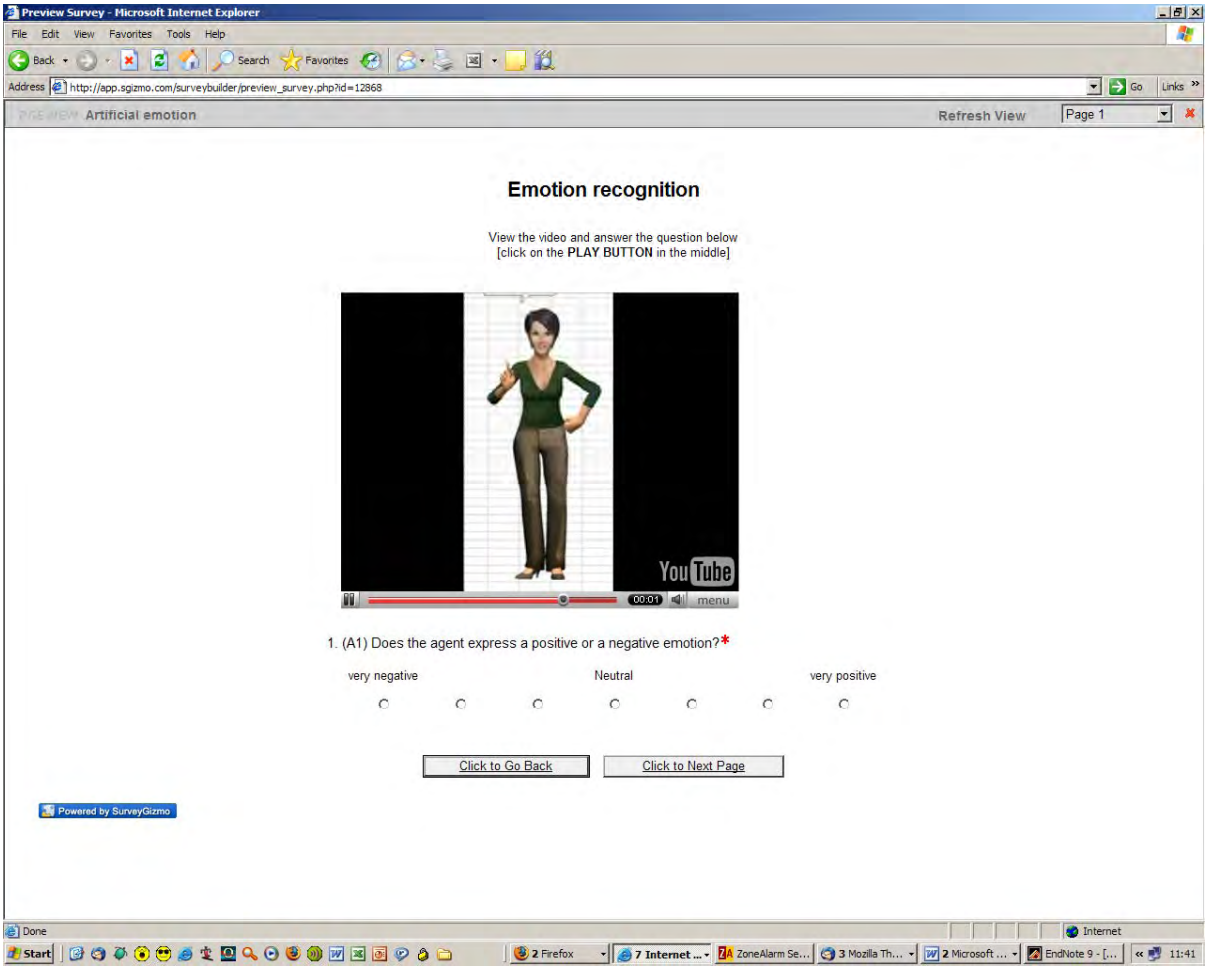


Figure 3. Questionnaire Layout

3.2.1 Method

The survey was open for 2 weeks in August 2007. Participants were recruited by e-mail invitations to friends and acquaintances, following the procedure of convenience sampling and snowball. In the e-mail, participants were invited to complete the study and to forward the invitation to others.

3.2.2 Results

The survey was accessed by 240 people. Of these, 57 people did not progress over the introductory page and 100 completed the evaluation for all 12 images. The statistics reported in this report are based on the participants who completed the survey.

Basic descriptive statistics of emotional evaluations for the 12 animations are reported in Table 3. The table is sorted by mean value (the lowest the more negative). Overall, it appears that, despite all the effort in the selection of the stimuli, most of these animations still did not convey the intended meaning. Looking at the standard deviation, a measure of the spread of a set of data, it emerges also that often participants disagreed on the evaluation of the emotional valence conveyed by the agent.

Target	Emotion	Animation	Expression	N	Range	Minimum	Maximum	Mean	Std. Deviation
Animation6	Negative	Sad_02	Sad_02	100	5	1	6	2.23	.973
Animation8	Negative	Angry_02	Angry_02	100	5	1	6	2.59	1.102
Animation4	Negative	Surprised_01	Angry_02	100	6	1	7	3.31	1.468
Animation7	Negative	Surprised_01	Surprised_02	100	6	1	7	3.63	1.376
Animation5	Positive	happy_01	Smile_03	100	6	1	7	4.20	1.231
Animation3	Negative	Argue_03	Angry_02	100	6	1	7	4.24	1.199
Animation1	Negative	Decline_01	Angry_02	100	6	1	7	4.49	1.514
Animation10	Positive	Argue_02	Smile_03	100	6	1	7	4.64	1.069
Animation11	Positive	Happy_02	Smile_03	100	4	3	7	4.79	.891
Animation12	Positive	Congratulate_01	Smile_03	100	4	3	7	5.63	.895
Animation2	Positive	Greet_01	Smile_03	100	5	2	7	5.82	.947
Animation9	Positive	Congattulate_02	Smile_03	100	6	1	7	6.23	1.053

Table 3. Descriptive Statistics for the 12 animations

The analysis of the mean values suggests that the perceived strength of the emotion was often weak. With the exception of Animation 6 and 8 (negative emotions) and Animation 2, 12 and 9 (positive emotions), the others were ambiguous. Figure 4 best illustrates this tendency by plotting the mean values of the 12 animations in a histogram. The y axis reports the number of cases which fits in each evaluation category (x axis). If we focus on the middle point of the evaluation scale (4), we see that most of the sample concentrates between the value of 4 (neutral) and 5 (slightly positive). This slight bias towards a positive evaluation also emerges by looking at the frequency of cases falling at the extremes of the scale (1-2; 6-7).

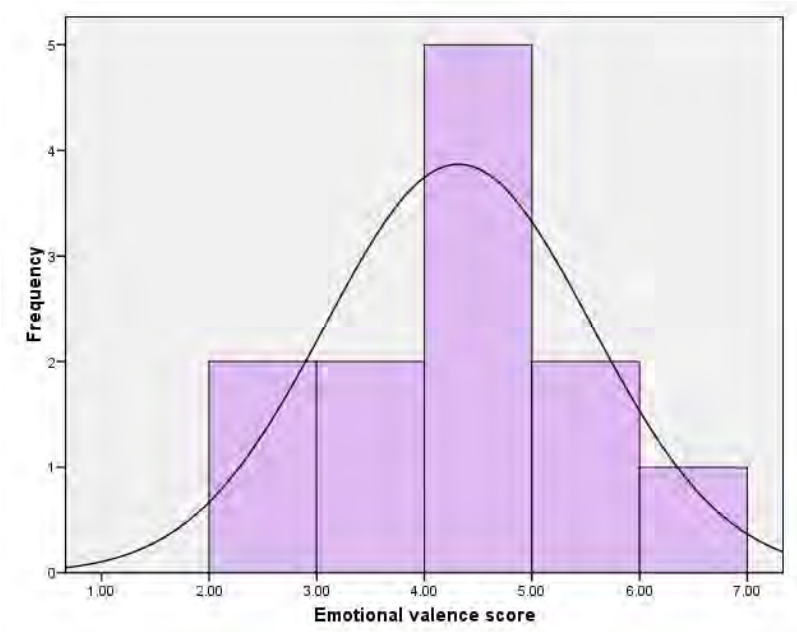


Figure 4. Histogram representation of mean evaluation score of the 12 animations

A one sample t-test was run on all animation scores to test if their mean values statistically differed by 4 (the middle point of the scale). To increase the reliability of the test, the probability level was set to 0.001. Results are summarised in Table 4. The animations which did not differ from 4 (i.e., did convey a neutral emotion) are reported in bold and labelled with an *. It is interesting to notice that one of them

(Animation 5) includes the animation labelled as Happy-01, which had been already identified as problematic in phase 1. It was tested in the survey due to the limited number of alternatives available to choose from.

	One-Sample Test Test Value = 4					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference Lower	Upper
Animation1 *	3.236	99	.002	.490	.19	.79
Animation2	19.221	99	.000	1.820	1.63	2.01
Animation3 *	2.002	99	.048	.240	.00	.48
Animation4	-4.700	99	.000	-.690	-.98	-.40
Animation5 *	1.625	99	.107	.200	-.04	.44
Animation6	-18.193	99	.000	-1.770	-1.96	-1.58
Animation7 *	-2.690	99	.008	-.370	-.64	-.10
Animation8	-12.797	99	.000	-1.410	-1.63	-1.19
Animation9	21.184	99	.000	2.230	2.02	2.44
Animation10	5.989	99	.000	.640	.43	.85
Animation11	8.867	99	.000	.790	.61	.97
Animation12	18.211	99	.000	1.630	1.45	1.81

Table 4. One-sample t-test results (test value =4)

3.2.3 Conclusion

This survey study provided quantitative support to many of the findings evinced in the first phase of the evaluation. In particular, it indicated that the communication clarity of Colette's non verbal language is far from perfect. However, the study allowed identifying four stimuli (2 positive and 2 negative emotions) to be tested in the stroop experiment. These stimuli are highlighted in Table 3 by a grey background. It has to be noted that although 3 animations classified as suitable in the positive dimensions (high average), Animation 12 was not included in the experiment, as we could not find a third negative emotion which clearly conveyed its intended meaning. Furthermore, Animation 12 is quite similar to animation 9, and hence it was discarded.

3.3 Discussion

This in depth evaluation of the non verbal behaviour of Colette confirms our previous remarks on the difficulty of designing effective non verbal communication for virtual agents. It clearly emerged that often the designer stereotypes did not match the user ones. Although we believe that Colette's gestures could be strongly improved by the rigorous application of theories and models of emotions, as well as by the strict application of a user-centred design approach, we must admit that Colette is a prototypic exemplar of the technological level of the embodied agents currently on the market. The many limitations of current technology urge the systematic analysis of possible semantic conflicts between verbal and non verbal messages in human-agent interaction. This state of the art supports the value of our research approach based on the stroop task paradigm over a simple eye-tracking experiment: it is not only important to understand if a user look at the agent, but also if the information is elaborated.

4. Experiment 1

This experiment examined the effect of Colette's animations displaying positive or negative emotions in a word evaluation task. The study was designed to understand if non-verbal cues provided by agents were attended to and processed. The stimuli consisted of short videos of Collette gesturing and displaying a word. The user was asked to react as quickly as possible to the word indicating if it expressed a positive or a negative meaning.

Following the stroop task paradigm, three experimental conditions were tested. In the *congruent condition* non verbal cues and textual message conveyed the same emotion, in the *incongruent condition* non verbal cues and textual message conveyed opposite emotions, in the *neutral condition* Colette did not display any emotion.

Hypothesis 1: Persona Effect. Following the persona effect, it was predicted that participants would pay attention to the non verbal behaviour of Colette. Consequently, negative gestures should facilitate negative judgement relative to positive ones, positive gestures should facilitate positive judgement relative to negative ones (Stenberg et al., 1998). Therefore we predicted the following set of results, with regards to both reaction times and errors.

Congruent condition < neutral condition < incongruent condition.

Hypothesis 2: Positive Valence Advantage. We also predicted latency differences between positive and negative words. This difference, known as Positive Valence Advantage PVA, is due to the fact that negative words tend to evoke more extended processing, leading to slower categorisation (Stenberg et al., 1998).

4.1 Method

4.1.1 Participants

Twenty-one[2] people participated in the experiment. They were member of staff or postgraduate students at the Manchester Business School of the University of Manchester. Nine participants were native English speakers; 12 were proficient in English but not native.

4.1.2 Materials

The animations of non-verbal messages were selected following the procedure described in section 3. Two videos displaying negative emotions (sadness and anger) and two videos displaying positive emotions (greetings and congratulation with a happy face) were used in the experiment. The neutral videos were recorded by combining a basic face and the animations labelled as SPEAK_01 and SPEAK_02, in the viewer. The videos were recorded using a blank Excel spreadsheet as background. They were displayed on a black screen and the words were written in white.

All videos lasted 5 seconds, and occupied an area of 480x360 pixels. They started with Colette in a rest position, then the facial animation was selected and finally the body movement. All video ended with Colette in the rest position. After 2.5 seconds a

[2] One participant was excluded from the analysis as she appeared to have misinterpreted the task, and only selected the Z key.

word was displayed on the screen. The word could randomly appear in one of four positions, at each corner of the video (Figure 5). Both the word and the video (showing Colette in the rest position) remained visible on the screen until the participant pressed a key.

Figure 5. Experimental layout

The verbal stimuli (positive and negative words N=94) were selected from the list of emotional words developed and tested by (Larsen et al., 2006). These words are completely balanced on length, frequency of use and orthographic neighbourhood size, thus overcoming an important threat to internal validity of word recognition experiments due to variations of lexical characteristics.

Three lists were constructed by randomly pairing words with videos, with the constraint that each video-word combination was represented by 14 items. For each participant one of the three lists was randomly selected at the time of the experiment.

4.1.3 Procedure

The experiment was implemented on DirectRT and displayed on a Dell Latitude D600 laptop (screen size 14''; screen resolution 1400 x 1050 pixels). Participants were tested individually. The experiment was introduced by written and verbal instruction, and the experimenter left the room as soon as she ensured that the participant had understood the task. Participants were told to evaluate if the word displayed on the screen was a positive or a negative by pressing one of two pre-set keys. The animations were to be ignored and participants were invited to act as quickly as possible, while maintaining high accuracy.

The first 10 trials acted as a practice sequence. They were followed by 84 experimental trials.

4.1.4 Design

Word valence (2: positive versus negative) and animation valence (3: positive, negative and neutral) were manipulated in a within-subjects design. The order of the trials was randomised between participants.

4.2 Results

On the average, some 8% of the trials resulted in errors. The mean error rates for each condition are shown in Table 5. It is clear that errors were not randomly distributed but rather triggered by specific factors. As expected, a strong effect of word-animation consistency emerged (χ^2=15.97 p < .001). Participants exposed to positive animations were more likely to commit an error when evaluating a negative word rather than a positive one. Vice versa, participants exposed to negative words were more likely to commit errors when evaluating a positive word than a negative one.

A general tendency for negative words to generate more errors than positive ones was also evident. The gap between the negative and the positive words emerged both in the neutral comparison and by looking at the consistent conditions.

		Word	
		Positive	Negative
Animation	Positive	3%	6%
	Negative	18%	9%
	Neutral	4%	8%

Table 5. Mean error rates in the 6 experimental conditions

Analysing the distribution of errors across individual subjects, it emerged that some people were more prone to interferences (the individual error rates ranged from 0 to 45%). The difference between native and non-native speakers was significant (χ^2=15.97 p < .001). Non-native speakers tended to commit on average 4% more errors than native speakers.

Before the correct reaction times (RT) were analysed, the outliers in each cell were removed, using a simple recursive procedure {Van Selst, 1994 #268}. The mean and standard deviation used for determining the cutoff for the first selection were computed using the entire sample. All RTs differing more than 4 standard deviations (SDs) from the mean were considered outliers and were removed from the sample. The mean and the SD of the resulting distributions were then computed and the procedure was repeated until no outliers remained. Following this procedure, a total of 2.8% of the trials were deleted.

The remaining RT data were then averaged across participants and experimental conditions and entered as dependent variables into a 2-way repeated measures ANOVA, with word (2) and animation (3) as factors. There was a significant effect of word valence $F_{(1,18)}$ = 4.3 p = .05. Positive words were processed faster than negative ones (mean difference = 44 msec). The main effect of animation and the interaction word*animation were not significant (F < 1).

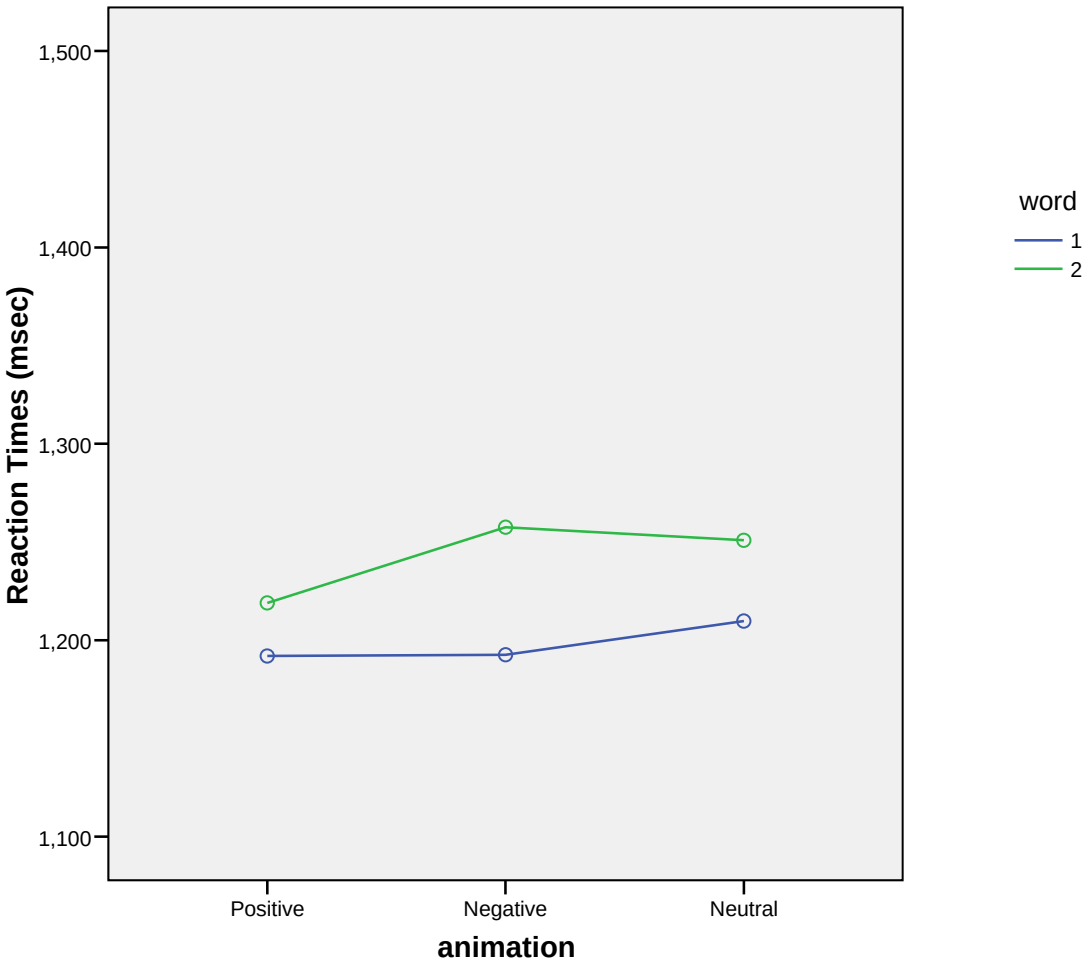


Figure 6. Mean reaction times as a function of animation and word (1=positive; 2= negative)

4.3 Discussion

Overall, the experiment provided mixed support to the persona effect. The analysis of the errors proved a clear conflict between consistent and inconsistent conditions, confirming that non-verbal cues from the agent were processed and interfered with word naming. In particular, the evaluation of negative words was strongly affected by the concurrent presentation of a positive emotion. On the average, the amount of errors was sizable (8%), significantly higher that the error percentage (3%) reported in (Stenberg et al., 1998), where participants were exposed to a combination of word and pictures. The difference may be due to the stronger effectiveness of videos in conveying emotions and capturing participants' attention or to the mixture of native and non-native English speakers tested in our sample.

The analysis of the reaction times showed a more complex framework. The only reliable effect was the positive valence advantage, explaining how negative words are systematically processed slower than positive words.

These findings raise an important question: why might we have found a clear interference effect only for a test measuring accuracy and not for a test measuring speed? Test sensitivity does not seem to be the answer. Indeed, the stroop task literature suggests that reaction times are more sensitive to detect smaller attentional conflicts than errors (MacLeod, 1991). Alternative explanations include possible

experimental artefacts induced by the type of videos used in the study, and the experimental task. Despite great care in the preparation of these stimuli, the 6 animations were not equivalent. Some of the main problems which may have affected our results are commented below.

1. *Animation evolution*. An important limitation of the experiment was that all videos started and ended in a rest position, displaying a slightly smiling face. After the display of negative emotions, the rest position may have created some cognitive conflicts, which may have exerted an influence in the word evaluation task, if the subject had not reacted yet to the word.

2. *Synchronisation issue*. Experiment 1 tested words appearing during the video (interval 2,500 msec), as we believed that this solution would have created stronger conflicts. Yet, due to a different time evolution of the 6 animations, at this time interval the emotional message was more or less strong. In both animations reflecting positive emotions the word appeared at the apex of the gestural communication. This point happened to consist of a salient harm movement (waving and OK gesture), which may have been used by the user as a clue that the word was going to be displayed. This explanation can account for the quickest processing time of negative words when displayed concurrently to positive emotion, than to neutral and negative one.

3. *Control condition*. The neutral condition was not completely emotionless, because Colette is strongly biased towards positive emotions, both in the posture and in facial expressions (section 3).

In order to investigate the reliability of the findings of experiment 1, an improved experimental design was tested in Experiment 2.

Experiment 1 also suggested an interesting difference between the performance of native and non-native English speakers, suggesting that individual differences may be an important factor in attention distribution.

5. Experiment 2

This Experiment aimed to reproduce Experiment 1 with a number of methodological modifications and improvements. Based on the difference evinced in Experiment 1 between native and non-native English speakers, we decided to concentrate exclusively on people who, despite being fluent in English, have learned it as a second language. This sample is important as it reflects a large proportion of educated European students and professionals, who are daily confronted with Internet resources in English.

The same hypotheses stated in Section 4 were tested (persona effect and positive valence advantage). An additional test was introduced, to address memory retention (Stenberg et al., 1998). At the end of the RT task, participants were invited to recognise the list of words presented in the experiment, from a set of distracters. Assuming that words encountered in inconsistent conditions were more deeply processed, to counteract the effect of the disturbing stimulus, we expected that they should have been more easily recognised (*Inconsistency advantage*).

5.1 Method

5.1.1 Participants

Twenty-two people participated in the experiment. They were postgraduate students at the University of Manchester. All of them were proficient in English, but not native speakers.

5.1.2 Materials

New videos were recorded to show a progression of emotion. They all started with the facial expression showing the lowest level of a specific emotion (e.g., angry 1), and then progressed to a more marked facial expressions (e.g., angry 2), immediately followed by the corresponding body animation (e.g., angry) 2. All videos lasted 3 seconds. The final clip, which remained visible on the computer screen until the user pressed a key, showed the agent in the apex of the emotion. The words were displayed 100 msec before the video was completed. Only four videos (two positive and two negative emotions) were recorded as in Experiment 2, no video was displayed in the control condition.

The word lists were also substantially revised. Six lists were created, completely balanced on average word valence, length and frequency of use. These lists were paired two by two, according to the procedure proposed by (Larsen et al., 2006). These authors developed two lists of positive and negative words. In each list, individual main words (either positive or negative) were matched to their opposite, balanced by length, orthographic variation, and frequency of use. Main words and their opposite were counterbalanced in our study, so that each list contained the same number of direct stimuli (main words) and opposite ones. Once again, length and frequency of use was kept constant. The six lists were then assigned to animation conditions, with the criteria that matched list could not be proposed in the same animation condition, and that each video could not include the same list twice.

A new list of 72 words was prepared to be tested in the memory items. It included 36 words tested in the first part of the experiment and 36 distracter items. Distracters were selected by the lists developed by (Larsen et al., 2006), with the constraint that

they could not have been previously used in the word evaluation task. Half of these words had a strong positive valence, the other half a negative one.

5.1.3 Procedure

The main procedure reflected that applied in Experiment 1. Participants were tested individually in a small room, on a Dell Latitude D600 laptop (screen size 14''; screen resolution 1400 x 1050 pixels). Each participants was randomly assigned 1 of the 6 list developed for the study. Participants were told to evaluate if the word displayed on the screen was a positive or a negative by pressing one of two pre-set keys (assignment of responses to keys was randomized). The animations were to be ignored and participants were invited to act as quickly as possible, while maintaining high accuracy. The first 10 trials acted as a practice sequence. They were followed by 72 experimental trials.

At the end of task 1, participants were invited to execute the memory test, by on-screen instructions. This followed immediately the evaluation task without prior notification. For each word, the participant had to indicate if the word had been or had not been previously presented in task 1. At the end of the experiment participants were thanked and debriefed.

5.1.4 Design

Word valence (2: positive versus negative) and animation valence (3: positive, negative and absent) were manipulated in a within-subjects design. The order of the trials was randomised between participants.

5.2 Results

The average error rate across all stimulus types was approximately 7%. The mean error rates for each condition are reported in Table 6, as percentage values computed across experimental conditions.

		Word	
		Positive	Negative
Animation	Positive	6%	10%
	Negative	10%	6%
	No animation	5%	3%

Table 6. Mean error rates in the 6 experimental conditions

By comparing the error occurrence in the 4 experimental conditions composed of words and animations, a significant effect of consistency emerge (χ^2=6.54, $p < .05$). Participants exposed to positive animations were more likely to commit an error when evaluating a negative word rather than a positive one. Vice versa, participants exposed to negative words were more likely to commit errors when evaluating a positive word than a negative one.

Before the correct reaction times (RT) were analysed, the outliers in each cell were removed, using the recursive procedure explained in Experiment 1. A total of 4% of the trials were deleted (maximum iterations = 7). The remaining RT data were averaged across participants and experimental conditions and entered as dependent

variables into a 2-way repeated-measure ANOVA, with word (2) and animation (3) as factors. The analysis returned a strong effect of word $F_{(2,21)} = 23.80$ $p < .001$. Animation and the interaction were not significant.

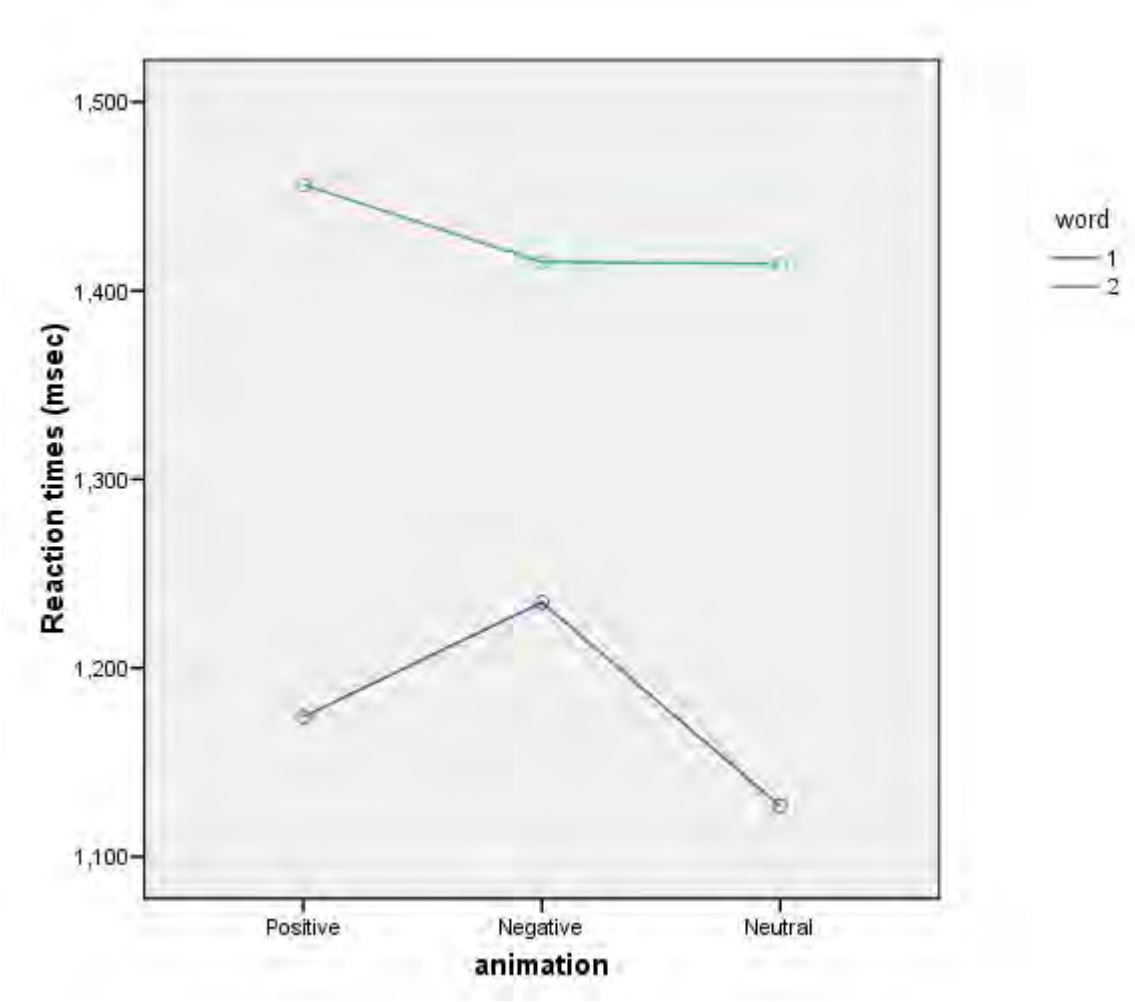


Figure 7. Mean reaction times as a function of animation and word (1=positive; 2= negative)

Figure 6 illustrates the mean values of reaction times as a function of experimental condition. The effect of word is immediately evident: positive words are processed faster than negative ones. Planned contrasts revealed a significant effect for the comparison between positive and negative words in combination with a positive or a neutral animation ($p < .001$), but not in combination to a negative animation.

The memory data were analysed selecting only those words which were correctly responded to in the first part of the experiment, and the distracters. A total of 19% of the trials in the memory test resulted in error. More errors occurred when participants had to evaluate experimental words (23%), rather than distracters (14%), revealing a conservative response bias (participants tended to answer no). No differences across experimental conditions emerged: participants made the same number of mistakes when recognising words which were originally presented as part of consistent combinations (video and word conveying the same meaning, 21%), inconsistent combinations (video and word conveying opposite meanings, 25%) or alone (24%). No difference in recognition was found between positive and negative words.

5.3 Discussion

Overall, experiment 2 showed a set of results highly consistent with those of experiment 2. These results provided mixed support to the persona effect. The analysis of the errors showed a strong conflict between consistent and inconsistent conditions, confirming that non-verbal cues from the agent were processed and interfered with word naming. On the average, inconsistent conditions induces some 4% of errors more than consistent conditions.

The analysis of the reaction times showed a more complex framework. The only reliable effect was the positive valence advantage showing that negative words were systematically processed slower than positive words. This effect is even stronger than in Experiment 1, probably because of the different timing between words and animations.

The memory task showed no differences in recognition between words presented in consistent and inconsistent conditions, or words presented alone. This suggests that non-verbal messages coming from the agent have little effect on learning.

6. Conclusion

In a general sense, support was provided for the persona-effect hypothesis, claiming that users pay attention to non-verbal communication provided by embodied conversational agents. A strong interference between words and gestures was evinced in both experiments in word-recognition accuracy. Positive words which were displayed with negative emotions tended to induce more errors that when they were displayed with positive emotions; and vice versa negative words displayed with positive emotions tended to induce more errors than negative words displayed with negative emotions.

The two experiments are also consistent in showing that there is a speed advantage for the reaction to positive valenced words over negative valenced words. These findings are consistent with psychological literature on the positive valence advantage (Stenberg et al., 1998). Yet, in contrast to previous studies we did not find any modification to this effect induced by the fact that words were presented in consistent or inconsistent conditions. Several procedural differences can explain this difference. Stenberg and his colleagues tested pictures of real human beings, rather than animations of embodied conversational agents; the display duration for the stimuli was significantly longer in our experiments, and the two stimuli were not superimposed as in the original study. The longer inspection time available to participants in our experiments may have allowed more controlled processing to develop, and hence decreased the effect of consistency/inconsistency in reaction times. This assumption would need independent verification. The difference in stimuli (face with a word superimposed in the study by Stenberg et al., 1998 vs. animation with a word on a corner in our study) is also a well-known predictor of the occurrence of interference effects (MacLeod, 1991).

The lack of consistency between RT and error data in both experiments can be explained by the notion of automatic processing. As discussed in section 2.2, both words and faces are automatically processed, meaning that they run their course to completion without controlled processing. What seems to happen in our study is a conflict between orientation of attention and executive attention. In inconsistent

conditions, gestural animations can sometimes capture the user attention and run their course to completion independently of the task at hand (word recognition). When the attention was oriented towards the agent, evaluation errors occurred. On the other hand, when the participant succeeded in keeping their attention focussed on the words (correct answers), then the animation did not seem to have been particularly disruptive in processing time. This explanation would need independent verification, but was confirmed by many comments spontaneously provided by participants, such as '*You know… I have tried not to look at her, but then I could'nt*' (participants 6, experiment 1).

Our results, suggesting that non-verbal cues provided by embodied conversational agents can affect verbal processing, are important to the design and evaluation of embodied conversational agents. Indeed, they warn designers of the fundamental importance of consistent communication codes between what the agent does and what the agent says, not to hamper the user performance. This deliverable provide a methodological framework and a research direction to test consistency in communication, which can be used by designer to evaluate that their 'intended meaning' is actually the meaning perceived by the user. The large scale evaluation of Colette, reported in this deliverable, has demonstrated the difficulty of designing for embodied communication and the need for a user-centred design approach in the design of embodied conversational agents.

7. References

ARGYLE, M. (1988) *Bodily Communication,* London Methuen & Co. Ltd. .

BERRY, D. C., BUTLER, L. T. & DE ROSIS, F. (2005) Evaluating a realistic agent in an advice-giving task. *International Journal of Human-Computer Studies,* 63**,** 304-327.

CASSELL, J., SULLIVAN, J., PREVOST, S. & CHURCHILL, E. (2000) *Embodied Conversational Agents,* Cambridge, The MIT Press.

CLARK, R. E. & CHOI, S. (2005) Five design principles for experiments on the effects of animated pedagogical agents. *Journal of Educational Computing Research,* 32**,** 209-222.

DEHN, D. M. & VAN MULKEN, S. (2000) The impact of animated interface agents:a review of empirical research. *International Journal of Human-Computer Studies,* 52**,** 1-22.

EASTWOOD, J. D., SMILEK, D. & MEROKLE, P. M. (2001) Differential attentional guidance by unattended faces expressing positive and negative emotion. *Perception & Psychophysics,* 63**,** 1004-1013.

EKMAN, P. & FRIESEN, W. V. (1975) *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues,* Englewood Cliffs, New Jersey, Prentice-Hall.

ENGLE, R. W. (2002) Working memory capacity as executive attention. *Current directions in psychological science,* 11**,** 19-23.

GULZ, A. (2004) Benefits of virtual characters in computer based learning environments: Claims and evidence. *International Journal of Artificial Intelligence in Education,* 14**,** 313-334.

HANSEN, C. H. & HANSEN, R. D. (1988) Finding the face in the crowd: An anger superiority effect. *Journal of Personality and Social Psychology,* 54**,** 917-924.

HONGPAISANWIWAT, C. & LEWIS, M. (2003) Attentional Effect of Animated Character. IN RAUTERBERG, G. W. M., MENOZZI, M. & WESSON, J. (Eds.) *Human-Computer Interaction -- INTERACT'03.* Zürich, Switzerland, IOS Press.

KODA, T. & MAES, P. (1996) Agents with Faces: The Effects of Personification of Agents. *HCI 96 People and Computers XI.* London, Springer-Verlag.

LARSEN, R. J., MERCER, K. A. & BALOTA., D. A. (2006) Lexical Characteristics of Words Used in Emotional Stroop Experiments. *Emotion,* 6**,** 62-72.

LESTER, J. C., CONVERSE, S. A., KAHLER, S. E., BARLOW, S. T., STONE, B. A. & BHOGAL, R. S. (1997) The persona effect: affective impact of animated pedagogical agents. *CHI 1997 Human factors in computing systems.* Atlanta, Georgia, United States, ACM Press.

MACLEOD, C. M. (1991) Half a century of research on the stroop effect: An integrative review. *Psychological Bulletin,* 109**,** 163-203.

MORENO, R., MAYER, R. E., SPIRES, H. A. & LESTER, J. C. (2001) The case for social agency in computer-based teaching: Do students learn more deeply when the interact with animated pedagogical agents? *Cognition & Instruction,* 19**,** 177-213.

POSNER, M. I. & ROTHBART, M. K. (2007) Research on attention networks as a model for the integration of psychological science. *Annual Review of Psychology,* 58**,** 1-23.

PRENDINGER, H., MA, C. & ISHIZUKA, M. (2007) Eye movements as indices for the utility of life-like interface agents: A pilot study. *Interacting with Computers,* 19**,** 281-292.

RICKENBERG, R. & REEVES, B. (2000) The effects of animated characters on anxiety, task performance, and evaluations of user interfaces. *SIGCHI conference on Human factors in computing systems.* The Hague, The Netherlands, ACM press.

RODA, C. & THOMAS, J. (2006) Attention aware systems: Theories, applications, and research agenda. *Computers in Human Behavior,* 22**,** 557-587.

SEYEDARABI, H., LEE, W.-S., AGHAGOLZADEH, A. & KHANMOHAMMADI, S. (2006) Facial Expressions Recognition in a Single Static as well as Dynamic Facial Images Using Tracking and Probabilistic Neural Networks. IN CHANG, L.-W., LIE, W.-N. & CHIANG, R. (Eds.) *PSIVT 2006.* Berlin Heidelberg Springer-Verlag.

SHNEIDERMAN, B. (1997) Direct manipulation versus agents: Paths to predictable, controllable, and comprehensible interfaces. IN BRADSHAW, J. (Ed.) *Software Agents.* Menlo Park, California, AAAI Press.

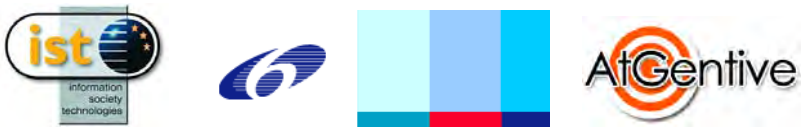
SPROULL, L., SUBRAMANI, M., KIESLER, S., WALKER, J. H. & WATERS, K. (1996) When the interface is a face. *Human-Computer Interaction,* 11**,** 97-124.

STENBERG, G., WIKING, S. & DAHL, M. (1998) Judging words at face value: interference in a word processing task reveals automatic processing of affective facial expressions. *Cognition and Emotion,* 12**,** 755-782.

TAKEUCHI, A. & NAITO, T. (1995) Situated facial displays: towards social interaction. *SIGCHI conference on Human Factors in Computing Systems - CHI'95.* Denver, Colorado, United States, ACM Press/Addison-Wesley Publishing Co.

VAN MULKEN, S., ANDRÉ, E. & MÜLLER, J. (1998) The Persona Effect: How Substantial Is It? *HCI'98 People and Computers XIII* Sheffield, UK, Springer-Verlag.

WITKOWSKI, M., ARAFA, Y. & DE BRUIJN, O. (2001) Evaluating user reaction to character agent mediated displays using eye-tracking technology. *AISB-01 Symposium on Information Agents for Electronic Commerce.*



The effects of a computer agent's gestures in guiding a user's attention on screen

AtGentive Del 4.4: AtGentive Final Evaluation Report Appendix B

Kimmo Vuorinen, Daniel Koskinen, Veikko Surakka, Harri Siirtola and Kari-Jouko Räihä

University of Tampere, Finland

Summary

The present study investigated the effects of a computer agent's gestures in guiding a user's attention on screen. The aim was to find out how an agent character's gestures can be used to attract and direct attention to visual interventions, when using a computer program or environment. To do this, we conducted an experiment where the user was presented with the following stimulus. First, an agent character appeared on the screen. Then, two simultaneous visual interventions were briefly shown and the agent gestured towards one of them. After this, the user was asked to remember the content of the interventions. The user's gaze was tracked with the help of an eye tracker to determine where his or her attention was focused during the tasks. The place of the agent and the direction of the intervention were systematically varied during the tasks. The results showed that the agent's gesture had a significant effect on how well the participants were able to remember the targeted intervention object.

Table of Contents

1. Introduction

The aim of the AtGentive project is to investigate the use of artificial agents for supporting the management of attention in the context of individual and collaborative learning environments. This includes developing educational systems that adapt according to a person's estimated state of attention. The present study aimed to investigate the effects of a virtual agent's gestures on the visual attention of users. To do this, we designed and ran controlled experiments where we used gaze-tracking techniques to determine the focus of users' visual attention. The results are expected to offer us valuable information on the effects that gesture- and expression-based cues of an embodied agent have on a user's attention and, by extension, learning performance. This information can be used to enhance the usability and role of agents in various applications and environment that involve the guidance of attention.

2. Background

Virtual embodied agents have potential to enhance human computer-interaction. They have been proposed as one solution to help users manage their workload, humanize computer interfaces, and provide a social link between the system and the users. In fact, it has been shown that the presence of a virtual character has a strong positive effect on a student's learning experience (Lester et al., 1997; Van Mulken et al., 1998; Moundridou and Virvou, 2002; Prendinger et al., 2003). One of the key strengths of embodied agents is that they have very broad and human-like capabilities for expressing themselves and interacting with computer users. Thus, they are capable of taking advantage of people's natural inclination to interact with computers in a social manner (Nass et al., 1994). Artificial characters can effectively influence several processes that are associated with learning, including memory, problem solving, decision-making, and attention (Schulkin et al., 2003; Matthews and Wells, 1999; Palomba, Angrilli, & Mini, 1997; Bechara et al., 2000).

The human brain is equipped with a wide variety of attentional mechanisms designed to cope with the abundance of information we continuously perceive around us. These mechanisms allow us to select the information we process, either consciously or unconsciously. Attention can help us to select relevant information, disregard irrelevant or interfering information and modulate or enhance the relevant information according to the state and goals of the perceiver (Chun & Wolfe, 2001, Driver, 2001, Lavie & Tsal, 1994).

In fact, attention represents one of the key factors in learning processes (Nissen & Bullemer, 1987; Grossberg, 1999). The most effective learners are not necessary the most intelligent or the brightest ones, but those who are able to organise their time efficiently, concentrate on their key activities, and complete them on time. In online settings, users are generally left on their own without support for attention and guidance from a tutor or peers. It is easy to procrastinate, engage in ineffective activities, or be distracted by something else. Embodied agents have been proposed as a solution to this, as they can potentially help users work more effectively and focus on relevant tasks. It has been suggested that this will also make the system more enjoyable to interact with and thus increase users' motivation as well. For example, there is evidence that embodied agents are effective in reducing the frustration of computer users (Hone, 2006).

It has been shown that artificial characters are efficient in capturing visual attention to themselves. For example, Witkowski and others (2001, 2003) measured gaze direction

while a computer user interacted with an agent. They found out that nearly 20 % of the time was spent looking at the agent and over 50 % of the time reading the agent's speech bubble. However, in a computer program or a virtual environment, an agent's function is often to guide the user to other elements or activities happening on the screen instead of just drawing attention to itself. An example would be a situation where the user is working on a task and the agent will alert him or her to new information (such as received emails or messages). Thus, one challenge for designing agents and their behavior is to support the management of attention effectively. This means guiding and direction an user's attention in a way that is subtle and not too distracting.

In the study by Witkowski and others (2001, 2003) attention was mostly directed to the agent character's face. Other studies have also proven that facial expressions are effective social cues in human-agent interaction (Partala & Surakka, 2004; Partala, Surakka, & Lahti, 2004; Vanhala et al., 2007). Besides facial expression, embodied agents are also capable of using a multitude of gestures, movements and changes in body language to convey emotions and emphasize or clarify what they are communicating. In fact, embodied agents can effectively mimic the same properties and behavior that humans exhibit in face-to-face conversation (Cassell, 2000). Early studies of task-oriented dialogues between a human and an agent by Deutsch (1974) have clearly shown the importance of nonverbal communication in such tasks. Recent studies by Marsi and van Rooden (2007) have shown that users prefer a non-verbal visual indication of an embodied system's internal state to a verbal indication.

Despite the potential benefits of embodied computer agents, there is little empirical evidence to help in designing characters and cues that are effective in guiding attention. Several studies have also had limitations, such as a lack of experimental conditions or different focus of interest, as pointed out by Dehn and Van Mulken (2000). In many cases, the effects of using an embodied agent have been compared to having no agent at all. Thus, the results of those studies can be used to argue for using artificial agents in general, but they are not particularly helpful in designing characteristics and behaviour for agents. Task-oriented studies about agents with respect to learning and collaborating with users have also often focused solely on verbal dialogues, with less emphasis on the potentially effective nonverbal cues that can be provided by the agents (Rickel and Johnson, 2000).

Eye tracking is one established technique for determining the focus of a person's visual attention at any particular time. By analyzing gaze paths, that is, sequences of saccades and fixations, it is possible to determine not only where a person is looking and for how long, but how and when his or her focus of attention changes (e.g. Hyrskykari et al., 2003; Merten & Conati, 2006). By utilizing eye tracking techniques we can, for example, study fixations of gaze to determine the proportion of time a user spends looking at an agent and other visual elements and cues on a computer screen. By analyzing gaze paths, we can study how an agent's actions and gestures influence the direction of users' visual attention.

The aim of the present study was to get insight into the attention-guiding properties of an embodied computer agent. The objective was to get concrete, empirically validated evidence concerning the effects of an agent character's gestures and their potential to guide attention. To do this, we conducted an experiment consisting of tasks where an agent character used gestures to guide a user's attention to visual information shown on the screen. Our aim was to find out how varying the type and direction of the gesture and the place of the agent affected the user's visual attention. Another aim was to find out what effect, if any, this has on the user's ability to remember the information the agent was targeting. Gaze tracking was used to determine the focus of the user's visual attention, and learning performance was investigated by having the user answer statements about information shown on the screen.

3. Methods

3.1 Participants

17 participants completed the experiment. Seven of the participants were male and ten were female, and their ages varied between 20 and 30. All of the participants had normal vision according to their own report.

3.2 Equipment

The experiment was performed in a soundproof laboratory. The experiment was presented in an Internet Explorer browser in full screen mode. The screen resolution was set to 1024 by 768 pixels on a 17" TFT monitor.

Tobii 1750 was used for eye tracking and ClearView software by Tobii was used for recording and analyzing the fixations and the gaze paths.

3.3 Stimuli

An animated agent called Matthew, created by Cantoche (see http://www.livingactor.com), was used in the study. The agent resembled a young boy with a cartoonish look and exaggerated head. We used two kinds of animated gestures that were predetermined for the agent to guide the user's attention. In the first one, the agent used his eyes to glance at the object and in the second one, the glance was also accompanied by a wave of the hand (see Figure 1). Thus, the agent character's head and eyes moved in both of the cases. The place the agent appeared on the screen was varied between the left and the right side of the screen. On the 17" screen, the agent was about 4 cm tall.

Figure 1. The two types of the agent's gestures used in the experiment.

For the intervention objects, we used small colored geometric shapes that were about 70 by 70 pixels in size as well as black letters that were of similar size. The intervention objects were paired so that two objects of the same type appeared together (see Figure 2). The objects were situated ao that they appeared on either side of the agent, at about 5 cm from the agent's approximated middle point. The direction of the interventions was also varied (see Figure 3). The pictures either appeared to the left and right or on the top

and bottom of the agent. This means there were eight possible place-direction combinations for the interventions. The amount of these conditions was balanced and systematically randomized for the participants.

Figure 2. An example of the intervention objects relative to the agent.

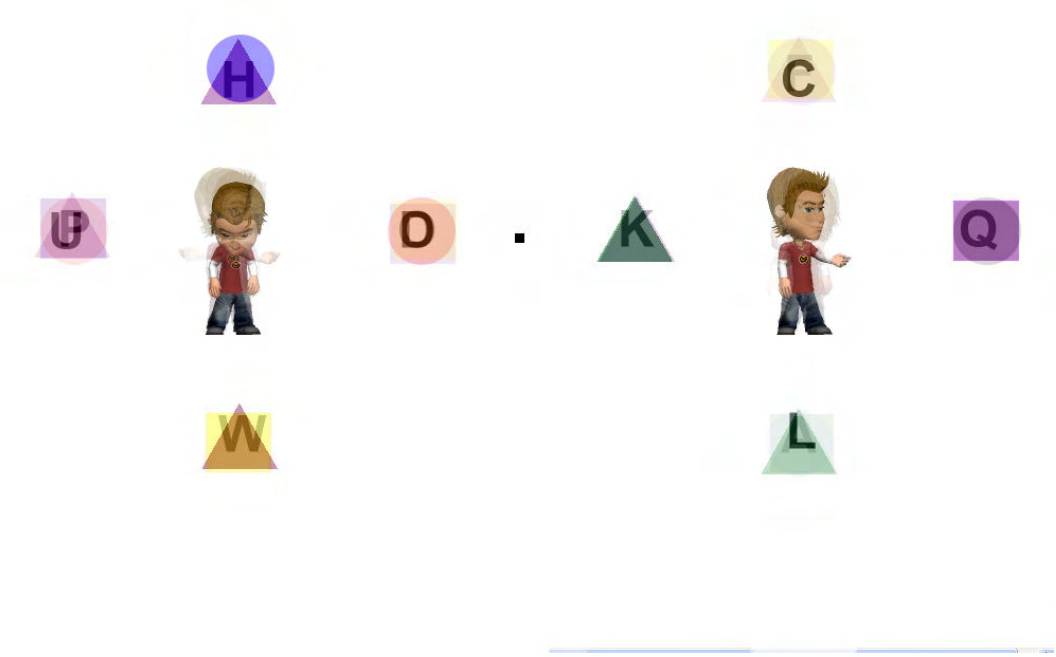


Figure 3. All possible locations for the agent and the intervention objects, superimposed on top of each other.

3.4 Procedure

The laboratory was first introduced to the participant. Then, he or she was instructed to sit comfortably in front of the screen. After this, the participants completed a calibration sequence (during which the coordinates were established on the screen). Before the test, the participants were instructed as follows. First, they were told to focus their gaze to a black dot in the middle of the screen until the agent character appeared. The participants were told to follow what happens on the screen, but they were not explicitly instructed as to what they should look at or what would happen on the screen after the agent character appeared. After the calibration and before the test, the participants were allowed to read the instructions on the screen once more and perform one example task that allowed them to familiarize themselves with how to proceed with the experiment. Then, they were given the chance to ask any questions they might have had.

The participants viewed the 24 stimuli in a systematically randomized order. As described previously, they were instructed to first focus their gaze on a black dot in the middle of the screen. The agent appeared on the screen 3050 milliseconds into the stimulus. After 4900 ms, the agent gestured towards either of two interventions that appeared simultaneously on the screen for 250 ms. The gesture was timed so that it started before the interventions appeared. The agent was hidden 10 seconds into the stimulus and the user was automatically forwarded after 11000 seconds to the statements. The mouse cursor was hidden during the stimulus so that the user's gaze would not be distracted by it. After this, the user was presented with the two true/false statements, for which he or she chose the answer with a mouse. The statements were either about the shape of the object ("e.g. the agent pointed toward a square"), or the color of the object ("e.g. the agent pointed toward a red object"), or the letter ("e.g. the agent pointed toward an R"). The user was moved to the next task after answering and confirming the answer with a "forward" –button.

After the test, the participants were asked if they had noticed the agent character's gestures and if they thought the gestures had had an effect on their attention or performance. Finally, they were debriefed and thanked.

3.5 Data Analysis

The numbers of fixations on defined areas of interest were calculated for each task and participant. The following five areas of interest were used: 1) the agent character 2) the agent character's head 3) the intervention object that was the target of the agent's gesture 4) the intervention object that was not the target of the agent's gesture 5) the middle of the screen where the participant was instructed to focus his attention in the beginning of the task. A 30 pixel margin was used for these areas. Fixations occurring before the appearance of the agent at 3050 ms were discarded.

8×2×2 within-subject analyses of variance (ANOVA) were performed on the numbers of fixations to the intervention objects and the statement answers data. The within-subject factors were the place of the intervention object (location), whether or not the agent was pointing to the object in question (direction), and the type of the gesture (gesture). Pairwise comparisons were performed using Bonferroni corrected two-tailed t-tests for factors with significant effects.

4. Results

Over all the participants and tasks, the numbers of fixations on the interest areas were distributed as follows: the agent character 42 % (n = 2837), center 50 % (n = 3527), targeted intervention object 4 % (n = 248), and non-targeted intervention object 4 % (n = 287). Of the fixations on the agent character, 79 % (n = 2283) were on the character's head.

A general analysis of the gaze paths (i.e. the order in which the fixations occurred) over all participants and tasks, showed that in 37 % of the tasks, there were no fixations on the intervention objects at all. In 12 % of the tasks, there were fixations on both intervention objects. In 76% of these cases, the targeted object was fixated upon before the non-targeted object. In other words, the participants mostly first looked at the targeted object and then moved their gaze to the non-targeted object.

The numbers of fixations on the targeted and non-targeted intervention objects were analyzed further. The 8x2x2 ANOVA for these fixations showed a statistically significant main effect for location $F(7, 10) = 8{,}98$, $p < 0.01$. Post hoc pairwise comparisons between the eight locations showed the following statistically significant differences: Between upper left and bottom left $MD = 0.69$, $p < 0.05$, upper left and bottom right $MD = 0.73$, $p < 0.05$, bottom left and center left $MD = 1.69$, $p < 0.001$, bottom left and center right $MD = 1.19$, $p < 0.01$, far left and center left $MD = 1.43$, $p < 0.001$, far left and center right $MD = 1.03$, $p < 0.001$, center left and upper right $MD = 1.1$, $p < 0.001$, center left and bottom right $MD = 1.63$, $p < 0.001$, center left and far right $MD = 1.46$, $p < 0.001$, upper right and center right $MD = 0.71$, $p < 0.05$, bottom right and center right $MD = 1.24$, $p < 0.01$, and far right and center right $MD = 1.06$, $p < 0.001$. In other words, the interventions in the center left and center right places, or closest to the center, gathered the most fixations.

The ANOVA for fixations the targeted and non-targeted intervention objects also showed a statistically significant main effect for direction $F(1, 16) = 5{,}34$, $p < 0.05$. A post hoc pairwise comparison between the two conditions showed that there were more fixations on the intervention object that the agent was not pointing to when compared to the object the agent was pointing to, $MD = 0.23$, $p < 0.05$.

The main effect of gesture type was not statistically significant for fixations. There were no significant interaction effects.

The 8×2×2 ANOVA for statement answers showed a statistically significant main effect for location, $F(7, 10) = 6{,}19$, $p < 0.01$. Post hoc pairwise comparisons between the eight locations showed the following statistically significant differences: Between upper left and bottom left $MD = 0.1$, $p < 0.01$, upper left and far left $MD = 0.06$, $p < 0.05$, upper left and center left $MD = 0.09$, $p < 0.01$, upper left and bottom right $MD = 0.07$, $p < 0.05$, upper left and center right $MD = 0.06$, $p < 0.05$, and center left and upper right $MD = 0.03$, $p < 0.05$. In other words, the interventions in the upper left and upper right places were identified the most correctly.

The ANOVA for statement answers also showed a stastically significant main effect for direction, $F(1, 16) = 6{,}99$, $p < 0.05$. A post hoc pairwise comparison between the two conditions showed that statements about the intervention object were answered more correctly when the agent was pointing to it when compared to when the agent was not pointing to it, $MD = 0.51$, $p < 0.05$.

The main effect of gesture type was not statistically significant for statement answers. There were no significant interaction effects.

When asked if they had noticed the agent character's gestures (n = 16), 35 % of the participants reported that they had definitely noticed them. 41 % of the participants reported that they had noticed the gestures in the end, in the middle of the test or that they had only partially paid attention to them. 17 % of the participants had not noticed the gestures at all.

5. Discussion

The area of interest that gathered the most fixations was the center of the screen, which was expected because the participants were instructed to focus their gaze on that point at the beginning of each task. The agent character itself also gathered a high number of fixations. Furthermore, most of the fixations were on the agent character's head, which is in line with previous research (Witkowski et al.; 2001, 2003). The intervention objects did not gather that many fixations, perhaps because they were visible on the screen for a very short time during each task.

For the number of fixations on the targeted and non-targeted intervention objects, a statistically significant main effect of location was found. The intervention objects closest to the centre gathered the most fixations, which can be easily attributed to the fact that the participants were instructed to focus their gaze on the center of the screen at the beginning of each task. However, the upper intervention objects, or the ones closest to the agent's head, also gathered many fixations. This further supports the notion that the proximity of the agent's head had a significant effect on attention.

A statistically significant main effect of gesture direction was found on the fixations. Specifically, the intervention object that the agent did not point to gathered more fixations than the intervention object at which the agent pointed. A possible explanation for this is that the users learned to memorize both of the intervention objects as the test progressed. However, because the targeted object might have been easy to memorize instantly, it is possible that they were required to pay active attention to the non-targeted object. This could have resulted in the comparatively higher number of fixations to the non-targeted object.

A statistically significant main effect of location was found on the statement answers. In practice, the intervention objects closest to the agent's head were remembered the most correctly. This suggests that the proximity of the agent's head had a positive effect on how well the users were able to remember the objects closest to it.

A statistically significant main effect of gesture direction was also found on the statement answers. Specifically, the participants answered statements about intervention objects more correctly when the agent pointed to them, compared to when the agent pointed to the opposite direction. This suggests that from a learning standpoint, the agent's gesture was effective in helping the participant remember the intervention picture.

Furthermore, almost a fifth of the participants reported not having noticed the gestures at all and twice as many of them had only noticed them partially. This shows that the agent's gesture helped the participants remember the targeted information, even though they did not necessarily pay attention to or even notice the gestures. In other words, this suggests that the agent's gestures helped the participants also on a subliminal level. Thus, they did not necessarily have to pay active attention to what the agent was doing, but they still benefited from the gestures. This is potentially important because under normal conditions in an online learning environment for example, the agent should

ideally be inconspicuous while the user is focused on doing a task. Thus, the agent's gestures could have the potential to guide the user's attention in a relatively unobtrusive manner, while using the agent's face and expressions would be helpful when the attention needs to be guided more conspicuously. This has also been found out in previous studies (e.g. Hansen & Hansen, 1988; Lundqvist, 2003; Vanhala et al., 2007), where it has been established that negative expressions provoke more immediate and stronger attentional and affective responses than positive ones.

There were no significant effects of gesture type in the present experiment. In other words, varying the type of the gesture had no additional effect to the user's ability to remember the objects or the focus of his or her attention.

In summary, visual attention to the intervention objects was affected both by the proximity of the agent's head and the agent's gestures. The gestures also had a significant effect on how well the participants were able to remember the intervention objects. The finding that the agent's gestures had an effect on the users' learning performance, even on a subconscious level, is an important one. A limitation of the present work was that the agent's animations were predetermined. Thus, the hand gesture could not be isolated from the movement of the agent's head and eyes. It would be interesting to investigate the effect of just using the agent's hand to guide the user's attention, while keeping the head and the agent's expression static. However, the studied cues of gesture and location were mostly independent of the agent character and could be easily varied with other similar embodied characters as well.

6. References

Bechara, A., Damasio, H., Damasio, A. R., (2000). Emotion, decision making and the orbifrontal cortex. Cerebral Cortex 10, 295-307.

Cassell, J. (2000). "More than Just Another Pretty Face: Embodied Conversational Interface Agents." Communications of the ACM 43(4): 70-78

Chun, M. M., & Wolfe, J. (2001). Visual attention. In E. B. Goldstein (Ed.), Blackwell's handbook of perception (pp. 272-310). Oxford, UK: Blackwell.

Dehn, D. M. and Van Mulken, S. (2000). The impact of animated interface agents: a review of empirical research. International Journal of Human-Computer Studies, 52, 1–22.

Deutsch, B. J. (1974). The structure of task oriented dialogs. In Proceedings of the IEEE Symposium on Speech Recognition. Pittsburgh, PA: Carnegie-Mellon University. Also available as Stanford Research Institute Technical Note 90, Menlo Park, CA.

Driver, J. (2001). A selective review of selective attention research from the past century. British Journal of Psychology, 92, 53-78.

Grossberg, S. (1999). The link between brain learning, attention, and consciousness. Consciousness and Cognition, 8(1), 1-44.

Hansen, C. H. and Hansen, R. D. (1988). Finding the face in the crowd: an anger superiority effect. Journal of Personality and Social Psychology, 54, 917–924.

Hone, K. (2006). Empathic agents to reduce user frustration: The effects of varying agent characteristics. Interacting with Computers, 18, 227–245.

Hyrskykari, A. (2006). Utilizing eye movements: Overcoming inaccuracy while tracking the focus of attention during reading. Computers in Human Behavior, 22(4), 657-671.

Lavie, N., & Tsal, Y. (1994). Perceptual load as a major determinant of the locus of selection in visual attention. Perception & Psychophysics, 56(2), 183–197.

Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., and Bhogal, R. S. (1997). The persona effect: Affective impact of animated pedagogical agents. In Proceedings of CHI'97, 359-366.

Lundqvist D., Esteves F., and Öhman A. (1999). The face of wrath: Critical features for conveying facial threat. Cognition and Emotion, 1999; 13: 691-711.

Lundqvist D, Esteves F, and Öhman, A. (2003). The face of wrath: The role of features and configurations in conveying facial threat. Cognition and Emotion, In Print.

Marsi, E., & van Rooden, F. (2007). Expressing uncertainty with a talking head in a multimodal question-answering system, Proceedings of the Workshop on Multimodal Output Generation (MOG 2007), Aberdeen, UK, 105–116

Matthews, G. and Wells, A. (1999). The cognitive science of attention and emotion. In T. Dalgleish (Ed.), Handbook of Cognition and Emotion (pp. 171–192). West Sussex, England: John Wiley & Sons.

Merten C and Conati C. (2006). Eye-Tracking to Model and Adapt to User Meta-cognition in Intelligent Learning Environments. To appear in Proceedings of IUI 06.

Moundridou, M. and Virvou, M. (2002). Evaluating the persona effect of an interface agent in a tutoring system. Journal of Computer Assisted Learning, 18, 253-261.

Nass, S., Steuer, J., and Tauber, E. R. (1994). Computers are Social Actors. In Proceedings of CHI '94, 72-78.

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from perfonnance measures. Cognitive Psychology, 19, 1-32.

Palomba, D., Angrilli, A., and Mini, A. (1997). Visual evoked potentials, heart rate responses and memory to emotional pictorial stimuli. International Journal or Psychophysiology, 27, 55–67.

Partala, T. and Surakka, V. (2004). The effects of affective interventions in human-computer interaction. Interacting with Computers, 16, 2, 295-309.

Partala, T., Surakka, V., and Lahti, J. (2004). Affective effects of agent proximity in conversational systems. In Proceedings of NordiCHI 2004, 353–356. Tampere, Finland.

Prendinger, H., Mayer, S., Mori, J., and Ishizuka, M. (2003). Persona effect revisited. Using bio-signals to measure and reflect the impact of character-based interfaces. Lecture Notes in Computer Science 2792, 283-291.

Rickel, J., Johnson, WL., (2001) Task-oriented collaboration with embodied agents in virtual worlds, Embodied conversational agents, MIT Press, Cambridge, MA.

Schulkin, J., Thompson, B. L., and Rosen, J. B. (2003). Demythologizing the emotions: Adaptation, cognition, and visceral representations of emotion in the nervous system. Brain and Cognition, 52, 15-23.

Surakka, V. and Hietanen, J. K. (1998). Facial and emotional reactions to Duchenne and non-Duchenne smiles. International Journal of Psychophysiology, 29, 1, 23-33.

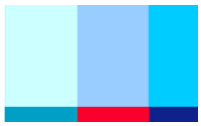
Van Mulken, S., André, E., and Müller, J. (1998). The persona effect: How substantial is it? In Proceedings of HCI Conference on People and Computers XIII, 53-66.

Vanhala, T., Surakka, V., Vuorinen, K., Siirtola, H., & Räihä, KJ. (2007) Virtual proximity and facial expressions as social and emotional cues. Submitted to International journal of human-computer studies.

Witkowski, M., Arafa, Y. and deBruijn, O. (2001) Evaluating User Reaction to Character Agent Mediated Displays using Eye-tracking Equipment. Proc. AISB'01 Symp. on Information Agents for Electronic Commerce, March 2001, pp. 79-87

Witkowski, M., Neville, B., and Pitt, J. (2003). Agent mediated retailing in the connected local community. Interacting with Computers, 15, 5-32.






Additional pedagogical investigation

AtGentive Del 4.4: AtGentive Final Evaluation Report
Appendix C

American University of Paris

Summary

This appendix describes some further work done with the objective of gaining a better understanding of the possible differences in the learning processes of children in the Experimental and Control groups.

We first looked at the data evaluating[1] the results obtained by students on each one of the six tasks they had to complete:

- Introduce themselves (the *intro* variable in the tables below reports the evaluation given to the introduction produced by the students)
- Ask some questions to the experts (the *Questions_asked* variable in the tables below reports the number of questions asked to the expert by the students)
- Specifying a goal (the *Good_goal* variable in the tables below reports the evaluation given to the goal specified by the students)
- Designing a concept map (the *cc* variable in the tables below reports the evaluation given to the concept map specified by the students)
- Write a paper
  - Length (the # *of paragraphs* variable in the tables below reports the number of paragraphs produced by the students)
  - Quality (the *QualityOf_paper* variable in the tables below reports the evaluation given to the paper produced by the students)
- Complete a questionnaire testing the knowledge of the students on the subject of the study, this was similar to a classic class test. Students were tested on the application of the knowledge they acquired in open questions explaining the advantages and disadvantages of the country they studied and the country they live in. (the *Questions* variable in the tables below reports the number of questions answered by the students; the *status* variable simply reports whether the student finished completing the questionnaire).

[1] The evaluation was performed by two of the researchers of the AtGentive project

Table of Contents

1. Overall learning outcomes

A Logistic Regression Analysis on the data relative to the evaluation of the results obtained by students on the six tasks above revealed that **children in the Experimental group asked significantly more questions to the experts (P=0.0491) and produced papers of significantly better quality (P=0.0506) than children in the Control group** (see table 1).

Independent Variable	Parameter Estimates	P-value
Intercept	-4.0116	0.3969
Questions_asked	0.4777	0.0491
Status	0.5967	0.5443
intro	-0.0787	0.6627
Good_goal	1.3687	0.0926
cc	-0.0355	0.7758
# of paragraphs	-0.1831	0.4584
Qualityof_paper	0.9135	0.0506
Questions	-0.0999	0.3033

Table 1 - the Experimental group asked significantly more questions to the experts and produced papers of significantly better quality. Model: log (odds of having the X group being better than the C-group) = -4.0116+0.4777*Questions_asked+0.9135*qualityof_paper

We therefore ran a further test (Logistic Regression Analysis with interaction) in order to verify both the significance of the difference (between X and C) in the performance of the two tasks, and the level of interaction between these two tasks. We confirmed (as reported in table 2 below) that the Experimental group asked significantly more questions (P=0.0131) and produced papers of significantly better quality (P=0.0059) than the students in the Control group. However, we found a significant negative interaction between the number of questions asked to the experts and the quality of the paper produced (P=0.0237) indicating that students who asked more questions were those more likely to produce a paper of lower quality.

Independent Variable	Parameter Estimates	P-value
Intercept	-3.0736	0.0042
Questions_asked	1.8598	0.0131
Qualityof_paper	1.3182	0.0059
Questions_asked*qualityof_paper	-0.6637	0.0237

Table 2 - The Experimental group asked more questions and produced better papers. Also, the number of questions asked and the quality of the paper interact significantly. Model: log (odds of having the X group being better than the C-group)=-3.0736 + 1.8598 * questions_asked + 1.3182 * qualityof_paper-0.6637 * questions_asked*qualityof_paper.

In section 3 we look more carefully at these interactions to gain a better understanding of the children learning processes and how they may differ in the Control and Experimental groups.

2. Effect of the AtgentSchool system interventions

In order to confirm that the better results obtained by the children in the Experimental group were due to the interventions of the AtgentSchool system we have looked at the interaction between the interventions produced by the system and the quality of the paper produced by children in the Experimental group.

Please note that in this appendix we mostly concentrate on analysing the effects of the *interventions received during a task* on the task itself. This is different from the analysis proposed in the main body of this deliverable where the interventions received by the children during the whole pilot study period are considered.

Although we observed a negative correlation (P=0.01) between the number of Cognitive interventions sent during the writing of the paper, and the quality of the paper itself (table 3), we analyzed the relations between the average number of interventions received during the paper writing task and the quality of the paper. Figure 1 shows that children who produced a better paper received, in average, more meta-cognitive (MC) interventions during that task

	Gamma, (ase), (p-value)
N_MC_paper recoded(0-12->0,13-56->1)	0.26, (0.36), (0.49)
N_C_paper recoded(0->0,1-7->1) Qualityof_paper recoded(1->0, 2,3->1)	-0.80, (0.17), (0.01)
N_M_paper recoded(0->0, 16->1) Qualityof_paper recoded(1->0, 2,3->1)	0.10, (0.41), (0.81)

Table 3 – Chi square Association between the quality of paper and the number of interventions sent during the paper writing task

Similar results were found in an analysis without recoding using Spearman's correlation (table 3'):

	Correlation coef (p-value)
N_MC_paper	0.20 (0.31)
N_C_paper	-0.50 (0.007)
N_M_paper	-0.32 (0.10)

Table 3' – Spearman's correlation between the quality of paper and the number of interventions sent during the paper writing task

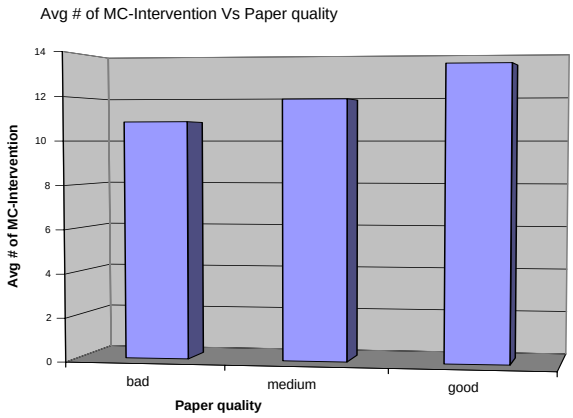


Figure 1 – MC-Interventions received by children who produced papers that were graded as bad, medium, and good.

A detailed analysis of the children's behaviour on the logs has lead us to the conclusion that no significant correlation between the quality of paper and Meta-Cognitive interventions is shown because some of the students who finally produced a bad paper kept swapping between activity windows, which has generated many meta-cognitive interventions. It appears therefore that the system was unable to help appropriately a certain number of students who kept moving around the application windows receiving meta-cognitive interventions that didn't really help them improving their performance.

Figure 2 shows that the same trend cannot be found in the case of Cognitive and Motivational interventions, in fact many children didn't receive any interventions of these types, and others, who have received the most cognitive and motivational interventions, have produced papers of worse quality.

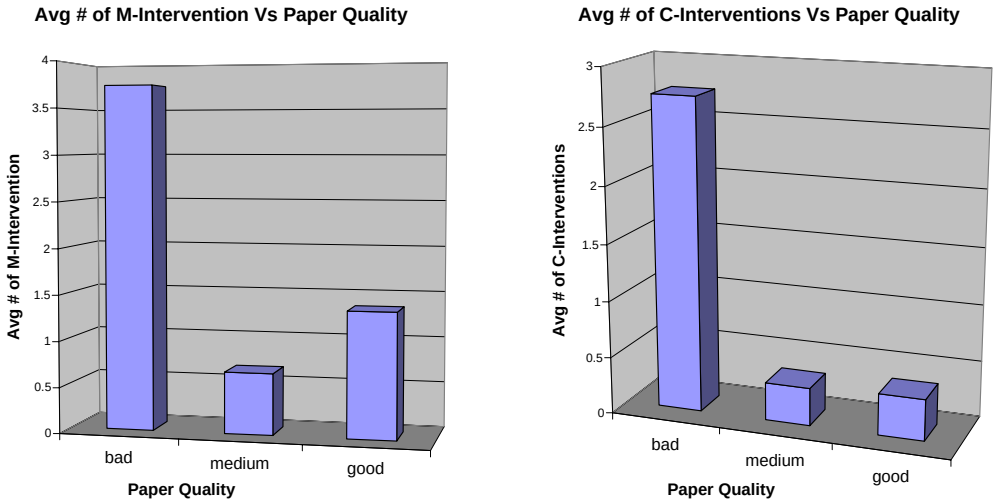


Figure 2 – Motivational (left) and Cognitive (right) Interventions received by children who produced papers that were graded as bad, medium, and good.

The large number of Cognitive and Motivational interventions received by some of the children who produced poor papers (and the negative correlation between the number of Cognitive interventions and the quality of paper) could be explained by the fact that such interventions were addressed to students who were clicking on the *question mark* button and on the *unhappy* button. We hypothesized that children who had been previously

(before the beginning of the pilot) defined by their teachers as *low* performers might had used the *question mark* and *unhappy* buttons more that children who had been defined as *high* performers. The results shown in figure 3 indicate that this was actually happening. Therefore, the students who were asking for help indeed needed the extra help.

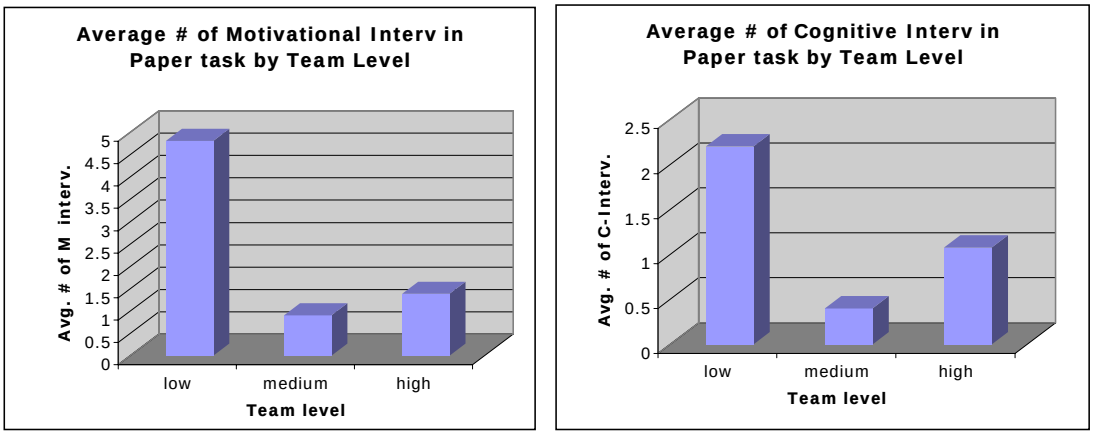


Figure 3 – Average number of interventions in the Paper task by team (children) level. It appears that the majority of Cognitive and Motivational interventions were given to children defined as low or medium performers by their teachers.

The students understood the manner of asking for help and were able to indicate to the system that they needed more help. However the system was unable to bring their performance up to the standards of the successful students. This observation opened the further question of how much help the system was able to supply to children of different levels of ability, we discuss this aspect in section 4.

It is interesting to note that interventions seem to have been particularly effective in stimulating the children to be more active in the forum by asking more questions to the expert. We found a positive correlation between all types of interventions provided during the forum task and the number of questions asked to the expert (see table 4).

	Gamma, (ase), (p-value)
N_MC_Forum recoded(0-12->0,13-218->1)	0.89, (0.12), (0.004)
N_C_Forum recoded(0->0,1-11->1)	0.83, (0.18), (0.02)
N_M_Forum recoded(0->0, 1-5->1)	0.86, (0.12), (0.003)

Table 4 – Chi square Association between the number of questions asked to the expert and the number of interventions sent during the forum activity

The results obtained, without recoding, using a Spearman's correlation (table 4') still reports the positive correlation between the forum activity and the Meta-Cognitive interventions, they don't however report the correlation with Cognitive and Motivational interventions. This confirms the observation made in the analysis of figure 2: there is a clear boundary between teams that didn't use the feedback/help buttons (which elicited those type of interventions) and children who did.

	Correlation coef (p-value)
N_MC_forum	0.67 (0.0001)
N_C_forum	0.20 (0.31)
N_M_forum	-0.004 (0.98)

Table 4' – Spearman's correlation between the number of questions asked to the expert and the number of interventions sent during the forum activity

The system is therefore successful in encouraging the collaboration between the children and the experts.

3. Control of the learning process

In order to gain an understanding of how children in the experimental and control groups might have gained a different level of control over their learning processes we have analysed the interactions between variables indicating the **overall quality of the work** (*QualityOf_paper, # of paragraphs, status*, and *Questions*) and variables indicating the quality of other parts of the learning process (*intro*, *Questions_asked*, *Good_goal, cc*). These interactions, if significant, may indicate how well the children understood the learning process itself (meta-cognitive regulation). In exploring these interactions we wanted to assess whether the system augmented with the attention module provided better support to the children in their meta-cognitive regulation (i.e. if more significant interactions would be found in the Experimental group than in the Control group). It should be noted that, given the relatively small size of the sample, in certain cases we had to recode the variables (i.e. group them in small sets) in order to find significant results, in fact, often we had too many categories and it was difficult to assess the tests performed (e.g. in chi square test, having too many categories increases the number of degrees of freedom). In the tables below we always explicitly indicate recoding specifying how variables were grouped.

3.1. How performance in the learning tasks correlates to the quality of the paper produced

As reported in table 5, when we looked, within the whole sample, at how the quality of the final paper interacted with variables indicating the quality of other parts of the learning process, we found no significant interaction.

Independent Variable	**Chi Square Test (P-value)**
Questions_asked_recod(0->0; 1,2,..,7->1)	0.6700
Intro_recod(0,..9->0,10->1)	0.8500
Good_goal	0.7200
Cc_recod(0,..5->0,6,..10->1)	0.9600

Table 5 - Each independent variable vs Quality of Paper for the Whole Sample. Chi-square test.

However, when we considered the same type of interactions limiting the sample to the Experimental group (table 6), we found that the number of questions asked to the expert is correlated with the paper quality (p=0.0300).

Independent Variable	Chi Square test (P-value)
Questions_asked_recod(0->0; 1,2,..,7->1)	0.0300
intro recod(0,..9->0,10->1)	0.9400
Good_goal	0.5900
cc recod(0,..5->0,6,..10->1)	0.2400

Table 6 - Each independent variable vs Quality of Paper for the Experimental Sample. Significant interaction between the number of questions asked to the expert and the paper quality. Chi-square test.

This effect is not found if the sample is limited to the Control group (see table 7): we found no significant interaction between the quality of the final paper and other parts of the learning process in the Control group.

Independent Variable	Chi Square Test (P-value)
Questions_asked_recod(0->0; 1,2,..,7->1)	0.2100
intro recod(0,..9->0,10->1)	0.7200
Good_goal	0.8400
cc recod(0,..5->0,6,..10->1)	0.4700

Table 7 - Each independent variable vs Quality of Paper for the Control Sample. No significant interaction found for the Control Group as regards to paper quality. Chi Square.

This data reflects the fact that **the interaction between the questions asked and the quality of paper applies only to children in the Experimental group**. From the analysis shown in table 2 we know that such interaction is negative (meaning that asking more questions may actually lower the quality of paper).

In order to gain a better understanding of the interaction between the quality of the paper and the questions asked to the expert, we have first looked at how the average number of questions asked, with respect to a specific paper quality varied between the Experimental and Control group. Figure 4 shows that not only children in the Control group asked significantly less questions on average, but also that the trend to a positive interaction between the number of questions asked and paper quality present in the control group, reverses in the Experimental group.

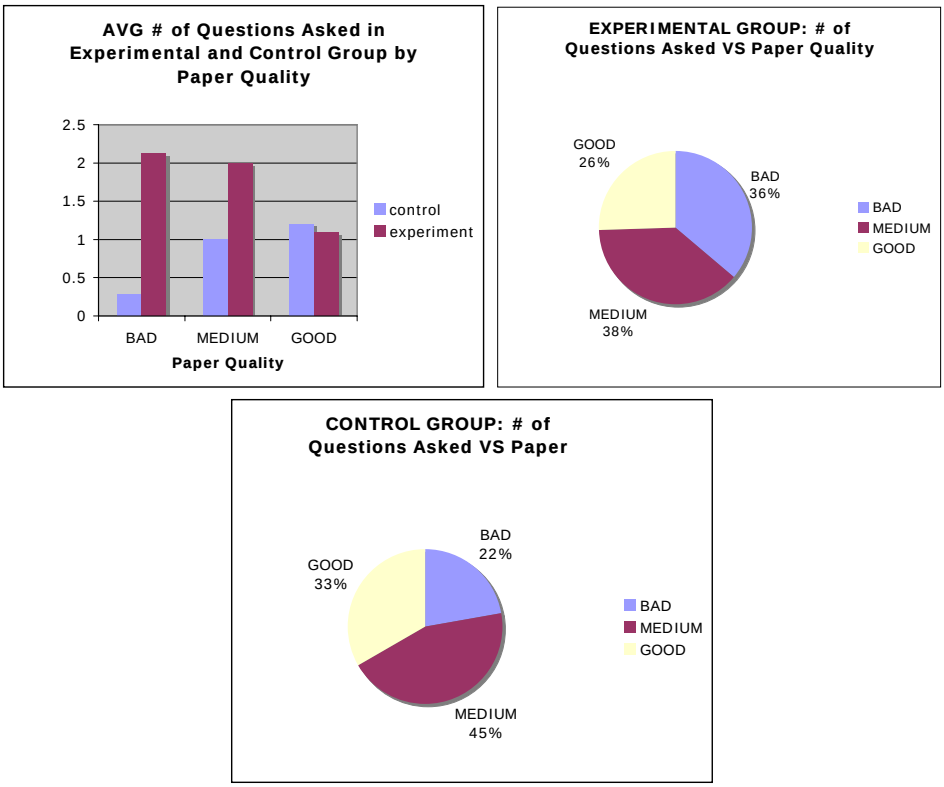


Figure 4 – the first chart shows that in the Control group children who produced better papers asked in average more questions than children who produced poorer quality papers. In the Experimental group children who produced better papers asked in average less questions than children who produced poorer quality papers. The Second and third charts show a questions count in the Control and Experimental groups showing that in both groups the largest percentage of questions were asked by children who produced papers of *medium* quality, in the Experimental group however 36% of the questions were asked by students who produced poor quality papers versus the 22% in the Control group.

Unfortunately the analysis performed so far has not allows us to find the cause for this negative correlation. One hypothesis was that the system had encouraged less capable teams (team level = Low) to ask more questions and that these have produced paper of lesser quality. However no significant correlation was found between the team level and the number of questions asked, therefore we cannot prove this hypothesis. Another possible explanation comes from the observation that, from a record of the dates when children asked the questions to the experts, it appears that most of the children who produced *bad* quality papers asked their questions to the experts very late during the pilot period, which probably didn't allow them to integrate the answers in their paper.

3.2. How performance in the learning tasks correlates to the length of the paper produced

As discussed previously a second, more mechanical, indicator of the strength of the work produced is the length of the final paper. As such we asked whether some relations existed on the quality of tasks that children were asked to perform and the length of the

paper. As shown by the tables[2] below (tables 8-10) no correlations were found, neither in the Experimental, nor in the Control group, nor in the overall sample, between these variables.

Independent Variable	Fisher's Exact Test (P-value)
Questions_asked (recoded:0->0, 1,…7->1)	0.4495
Intro (recoded:0,..,9->0,10->1)	0.7036
Good_goal	1.0000
Cc (recoded:0,..,5->0,6,..10->1)	0.2519

Table 8 - Tasks in the learning sequence vs the Number of Paragraphs (recoded:0,..,3->0; 4,..7->1) for the Experimental sample

Independent Variable	Fisher's Exact Test (P-value)
Questions_asked (recoded:0->0, 1,…7->1)	0.1201
Intro (recoded: 0,..,9->0,10->1)	0.7064
Good_goal	0.2087
Cc (recoded:0,..,5->0,6,..10->1)	0.4495

Table 9 - Tasks in the learning sequence vs the Number of Paragraphs (recoded:0,..,3->0; 4,..7->1) for the Control sample

Independent Variable	Fisher's Exact Test (P-value)
Questions_asked (recoded:0->0, 1,…7->1)	0.7891
Intro (recoded:0,…,9->0, 10->1)	0.4218
Good_goal	0.3290
Cc (recoded:0,…,5->0;6,…10->1)	0.7875

Table 10 – Tasks in the learning sequence vs the Number of Paragraphs (recoded:0,..,3->0; 4,..7->1) for the whole sample

3.3. How performance in the learning tasks correlates to a good score in the questionnaire

A good score in the questionnaire is the final indicator of the strength of the work produced. As such we asked whether some relations existed on the quality of tasks that children were asked to perform and the questionnaire score. As shown by the tables below there was a significant correlation, in the whole sample (table 11), between the number of concepts in the concept map (P=0.0143) and a good score in the questionnaire. Such correlation is still

[2] Tables 6-11 present the results of Chi-square test (with contingency tables) to test the significance of each of the independent variables versus the choice justified by many reasons (questions).Since the sample size was small (n=55), a Fisher's Exact Test was used.

significant (P=0.0235) in the sample including only the Experimental groups (table 12), but not in the sample including only the Control groups (table 13).

Independent Variable	Fisher's Exact Test (P-value)
Questions_asked (recoded:0->0, 1,…7->1)	1.0000
Intro (recoded:	0.0598
Good_goal	0.7458
Cc :(recoded:0,..,5->0,6,..10->1)	0.0143

Table 11 - Tasks in the learning sequence vs the questionnaire score (questions) (recoded: 0,1,2,..,5->0, 6,7,..17->1) for the Whole Sample

Independent Variable	Fisher's Exact Test (P-value)
Questions_asked (recoded:0->0, 1,…7->1)	0.4454
Intro: (recoded:0,…,9->0, 10->1)	0.1358
Good_goal	0.6132
Cc: (recoded:0,..,5->0,6,..10->1)	0.0235

Table 12 - Tasks in the learning sequence vs the questionnaire score (questions) (recoded: 0,1,2,..,5->0, 6,7,..17->1) for the Experimental group

Independent Variable	Fisher's Exact Test (P-value)
Questions_asked: (recoded:0->0, 1,…7->1)	0.4244
Intro: (recoded:0,…,9->0, 10->1)	0.2519
Good_goal	1.0000
Cc: (recoded:0,..,5->0,6,..10->1)	0.2576

Table 13 - Tasks in the learning sequence vs the questionnaire score (questions) (recoded: 0,1,2,..,5->0, 6,7,..17->1) for the Control group

Once again, the correlation holds for the children in the Experimental group but not for the children in the Control group, indicating that the system may have contributed to the creation of these effects between the different parts of the learning process.

3.4. Overall interactions between different parts of the learning process

In order to gain a better understanding of how the main variables describing the learning process interacted, we performed a bivariate correlation analysis over the complete sample (table 14) as well as over the Experimental and Control samples (tables 15, and 16).

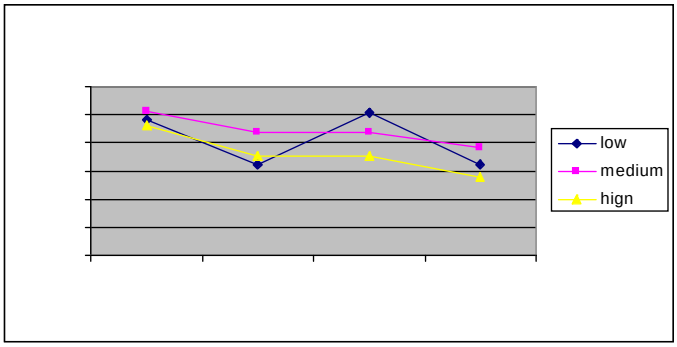
	Questions_asked_r	Intro_rc	Good_goa_	cc_rc	Paraphrase	Qualityof_paper	Questions_rc
Questions asked_rc_	1 (0.000) (0.000)	-0.26 (0.25) (0.32)	-0.12 (0.32) (0.72)	-0.14 (0.27) (0.61)	0.10 (0.27) (0.69)	-0.08 (0.22) (0.67)	-0.04 (0.27) (0.88)
Intro_rc		1 (0.000) (0.000)	0.14 (0.32) (0.66)	0.61 (0.18) (0.01)	-0.25 (0.25) (0.34)	-0.006 (0.22) (0.85)	-0.51 (0.21) (0.04)
Good_goal			1 (0.000) (0.000)	0.27 (0.32) (0.41)	0.39 (0.29) (0.22)	0.18 (0.27) (0.72)	0.19 (0.32) (0.56)
cc_rc				1 (0.000) (0.000)	-0.12 (0.27) (0.67)	-0.03 (0.22) (0.96)	-0.63 (0.17) (0.009)
# of paragraphs _rc					1 (0.000) (0.000)	0.79 (0.10) (<0.0001)	0.10 (0.27) (0.69)
Qualityof_paper						1 (0.000) (0.000)	-0.15 (0.22) (0.78)
Questions_rc							1 (0.000) (0.000)

Table 14 – Bivariate correlations in the Complete Sample. We used the gamma measure of association (ASE –asymptotic standard error) which is an ordinal correlation coefficient (between -1 and 1) based on concordant and discordant pairs using the ordering of the levels of the variables to determine if the association is negative, positive or present at all. The third row of each cell is the p-value obtained from the chi square test.

***Gamma, ASE, p-value**

The most interesting data for the whole sample is the strong negative correlation between the number of concepts in the concept map and the number of justifications provided for the choice made (P=0.009) (questions). When we restrict the analysis to the Experimental group (table 15) such negative correlation is still present (P=0.02), but we did not find it in the Control group (table 16). We haven't been able to prove any clear cause for these negative correlations but a reasonable hypothesis is that children who have mentioned key reasons for accepting or rejecting one country in the concept map may have though that it was unnecessary to report these reasons also in the questionnaire. In particular, a further analysis of the results produced by the students is shown in Figure 5. Students rated as *low* performers were able to name more pro and cons for the Tjech Republic (questions 3 and 4) whereas the teams rated as *high* performer did not mention so many there because they did not find it that important and their answers are more evenly distributed across the four questions.

We believe however that it is important that the correlation between the number of concepts in the concept map and the number of questions answered in the questionnaire exists only in the complete sample and Experimental group (and not in the control group).



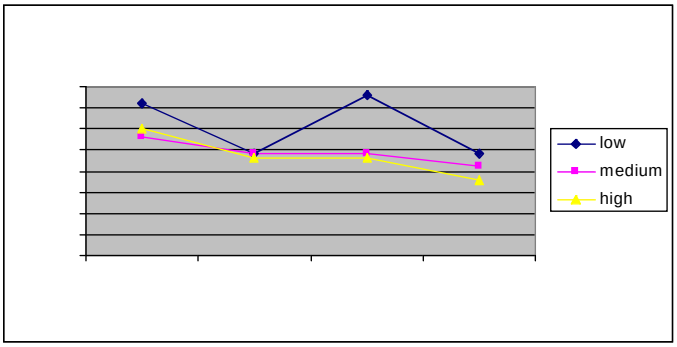


Figure 5 shows the distribution of children answers over the four questions in the questionnaire. Answers of children who have produced papers of high or medium quality are equally distributed across the four questions. Answers of children who have produced papers of low, although much more numerous are concentrated on specific questions, this could explain the inverse correlation between the number of concepts in the concept maps and the number of answers provided to questions in the questionnaire.

Table 15 includes some of the earlier findings such as the negative correlation between the number of questions asked to the expert and the quality of paper (see section 3.1).

The other two relevant effects of the team level on the quality of the goal (P=004) and on the number of questions answered (P=0.03) are discussed in section 4.

	Team Level	Questions asked	Intro	Good _goal	cc	Parag	Qualityof_ paper	Questions
Team_Level	1.000*	-0.22	-0.09	0.95	-0.21	0.15	0.32	0.71
	(0.000)	(0.31)	(0.32)	(0.06)	(0.31)	(0.32)	(0.26)	(0.21)
	(0.000)	(0.76)	(0.85)	(0.004)	(0.69)	(0.87)	(0.30)	(0.03)
Questions_asked		1.000	-0.40	-0.50	-0.13	-0.41	-0.72	-0.40
(recoded:0->0,1…7->1)_		(0.000)	(0.33)	(0.46)	(0.38)	(0.32)	(0.17)	(0.33)
		(0.000)	(0.27)	(0.35)	(0.74)	0.25)	(0.03)	(0.27)
Intro			1.000	0.43	0.51	-0.29	0.10	-0.59
(recoded:0,…,9->0, 10->1)			(0.000	(0.50)	(0.29)	(0.35)	(0.31)	(0.27)
			(0.000	(0.44)	(0.15)	(0.45)	(0.94)	0.10)
Good_goal				1.000	-0.17	0.00	0.40	0.43
				(0.000)	(0.53)	(0.54)	(0.40)	(0.50)
				(0.000)	(0.76)	(1.00)	(0.59)	(0.44)

Cc (recoded:0,..,5->0,6,..10->1)					1.000 (0.000) (0.000)	-0.54 (0.28) (0.13)	-0.16 (0.31) (0.24)	-0.79 (0.18) (0.02)
# of paragraphs_r (recoded:0,..,3->0;4,..7->1						1.000 (0.000 (0.000	0.88 (0.10) (0.0016)	0.54 (0.28) (0.13)
Qualityof_paper							1.000 (0.000) (0.000)	0.26 (0.30) (0.59)
Questions (recoded: 0,..,5->0, 6,..17 >1)								1.000 (0.000) (0.000)

Table 15 – Bivariate correlations in the Experimental Sample. *Gamma, ASE, p-value

	Team Level	Questions asked_	Intro	Good_goa _	cc	Parag	Qualityof_ paper	Questions
Team_Level	1.000* (0.000) (0.000)	0.15 (0.35) (0.91)	-0.52 (0.27) (0.23)	-0.09 (0.38) (0.89) (	-0.43 (0.28) (0.11)	0.24 (0.32) (0.19)	(0.32) (0.27) (0.41)	0.43 (0.28) (0.11)
Questions_asked (recoded:0->0,1…7->1)_		1.000 (0.000) (0.000)	-0.06 (0.40) (0.88)	-0.01 (0.44) (0.97)	-0.14 (0.40) (0.72)	0.62 (0.26) (0.08)	0.55 (0.25) (0.21)	0.45 (0.33) (0.25)
Intro (recoded:0,…,9->0, 10->1)			1.000 (0.000) (0.000)	0.05 (0.42) (0.90)	0.71 (0.21) (0.03)	-0.22 (0.37) (0.57)	-0.02 (0.33) (0.72)	-0.50 (0.30) (0.17)
Good_goal				1.000 (0.000) (0.000)	0.54 (0.33) (0.19)	0.61 (0.30) (0.12)	-0.11 (0.36) (0.84)	0.16 (0.41) (0.71)
Cc (recoded:0,..,5->0,6,..10->					1.000 (0.000 (0.000	0.35 (0.34) (0.34)	0.21 (0.31) (0.47)	-0.47 (0.31) (0.19)
Paraphrase (recoded:0,..,3->0;4,..7->1						1.000 (0.000 (0.000	0.75 (0.17) (0.023)	-0.35 (0.34) (0.34)
Qualityof_paper							1.000 (0.000) (0.000)	-0.43 (0.28) (0.21)
Questions (recoded: 0,..,5->0, 6,..17-								1.000 (0.000) (0.000)

Table 16 – Bivariate correlations in the Experimental Sample. *Gamma, ASE, p-value

3.5. Conclusions about the learning process

The fact that correlations exist (albeit negative), for children in the Experimental group, between the number of questions asked and the quality of paper, and between the number of concepts in the concept map and the ability to answer the questionnaire, could indicate that children in the Experimental group built a different knowledge model of the acquired knowledge during the learning process. This result, together with the fact that students in the Experimental group produced better quality papers and asked more questions to the experts (table 1), could indicate that the attention management system, is influencing the performance of the students effectively, and changing the learning in a profound manner. It could be suggested that with the regulation on the meta-cognitive and cognitive level by the attention management system, students do have a better awareness over their own learning process which supports the effective organization of the knowledge learned, for application of this knowledge in later instances. However it should also be noted that if the system is not able to influence the performance these effects are not found. We did find that *low* performer students were able to indicate their need for help effectively, but it appeared that the system could not support them to obtain better quality performance. This may indicate that for weaker students more support is needed to acquire results, and in particular Cognitive and Motivational interventions could be designed to be more effective. This is further analyzed in the next section.

4. Building on previous knowledge

4.1. Overview

In this section we investigate possible correlations between children's previous knowledge (as indicated by the *team level* variable) and the results obtained in the tasks performed. The *team level* was assigned before the beginning of the pilot by the teachers. Team level can have three values low (L), medium (M), and high (H).

The analysis of the effects of the team level on each one of the tasks that the children had to perform on the whole sample (see table 17) indicated that the team level is significantly correlated with the ability to answer the questionnaire (questions) (P=0.0066). A marginally significant correlation is also detectable between the team level and the number of paragraphs written (P=0.0912).

Independent Variable	**Fisher's Exact Test (P-value)**
Questions_asked	0.1194
Status	0.3171
intro	0.7546
Good_goal	0.1419
cc	0.1918
# of paragraphs	0.0912
Qualityof_paper	0.1524
Questions	0.0066

Table 17 - Effect of *Team Level* on quality of tasks performed for the Whole Sample

This result shows that, in the overall sample, the final outcome of the work (as represented by the two variables reporting the number of paragraphs written in the paper, and the number of questions answered in the questionnaire) is correlated to the starting level of the team, i.e. on the whole sample, better students perform better on the questionnaire and produce longer papers).

Within the Experimental sample we found different significant correlations: the team level is significantly correlated to the ability to generate good goals (P=0.0040), and to the ability to complete the questionnaire (P=0.0300), see table 18.

Independent Variable	**Chi Square Test (P_value)**
Questions_asked(recoded0->0,1..7->1)	0.7600
Status	0.1300
Intro (recoded0,..9->0,10->1)	0.8500
Good_goal	0.0040
Cc(recoded 0..5->0,6..10->1)	0.6900
# of paragraphs (recoded 0..3->0,4…7->1)	0.2500
Qualityof_paper	0.3000
Questions(recoded0..5->0,6..17->1)	0.0300

Table 18 - Effect of *Team Level* on quality of tasks performed for the Experimental Sample

Within the Control group we couldn't observe any correlation between the team level and the quality of tasks performed (see table 19).

Independent Variable	**Chi square Test (P-value)**
Questions_asked(recoded0->0,1..7->1)	0.9100
Status	0.6200
intro(recoded0,..9->0,10->1)	0.2300
Good_goal	0.8900
Cc(recoded 0..5->0,6..10->1)	0.1100
# of paragraphs (recoded 0..3->0,4…7->1)	0.1900
Qualityof_paper	0.4100
Questions (recoded0..5->0,6..17->1)	0.1100

Table 19 - Effect of *Team Level* on quality of tasks performed for the Control Sample

4.2. Conclusions about building on previous knowledge

Whilst these results would support the hypothesis that the children in Experimental group were better assisted in exploiting their previous knowledge, they could also indicate that the enhanced system would increase the gap between *good students* and *bad students*.

If this was the case, good students in the Experimental group would do significantly better than *good students* in the Control group, whilst *bad students* in the Experimental group wouldn't. In order to verify whether this was happening we used a Wilcoxon Rank Sum Test to compare the quality of the task performed for the Experimental and Control groups within specific team levels.

As shown in table 20, there were no significant differences on the quality of task performed between the Experimental and Control groups for teams of levels *low* or *Medium*.

Independent Variable	**Wilcoxon 2-sided Exact Test Statistic, S**	**P-value**
Questions_asked	197.0	0.11
Status	225.0	1.00
intro	257.0	0.33
Good_goal	232.5	1.00
cc	243.5	0.66
# of paragraphs	230.5	0.94
Qualityof_paper	208.5	0.35
Questions	245.5	0.59

Table 20 - Experimental and Control differences for the "L and 'M' group in the Team_Level (*n=30*)

Further, no significant differences were found when considering teams of *low* level only (see table 21).

Independent Variable	**Wilcoxon 2-sided Exact Te Statistic, S**	**P-value**
Questions_asked	11.5	0.75
Status	15.5	1.00
intro	16.0	0.68
Good_goal	15.5	1.00
cc	10.0	0.39
# of paragraphs	9.5	0.32
Qualityof_paper	10.5	0.46
Questions	17.5	0.21

Table 21 - Experimental and Control differences for the "L" group in the Team_Level (*n=8*)

However, we found one significant difference on the ability of setting a good goal (P=0.04) between the Experimental and Control groups for teams of *high* level (see table 22).

Independent Variable	Wilcoxon 2-sided Exact T Statistic, S	P-value
Questions_asked	143.5	0.47
Status	167.0	0.64
intro	146.0	0.59
Good_goal	130.0	0.04
cc	141.0	0.41
# of paragraphs	138.5	0.34
Qualityof_paper	125.5	0.12
Questions	161.0	0.80

Table 22 - Experimental and Control differences for the "H" group in the Team_Level (*n=25)*

These results allow us to conclude that **whilst the enhanced system does seem to support better *good* students in the setting of the goal, in the case of the other tasks, and in particular in the case of the overall quality of work (as described in section 2) the enhanced system supports equally students of different levels**. Figure 6 below, for example, compares the quality of paper of students in the Experimental and control group for children in team levels Low, Medium, and High. First of all, it is possible to observe the finding of table 1 (children in the Experimental group do significantly better than children in the Control group). Also it can be observed that this improvement is distributed across the team levels (L, M, and H). For team levels Low and High children in the Experimental group do significantly better, with some of the children in the Low group obtaining the highest possible grade in the paper (3).

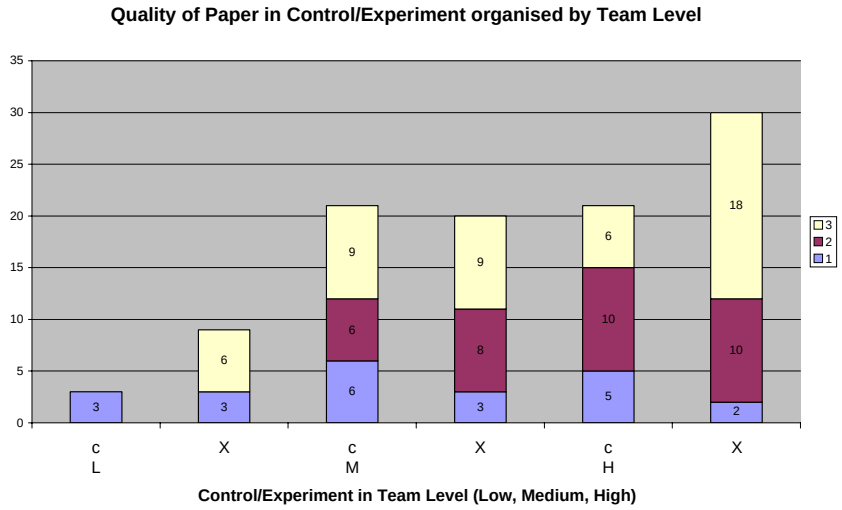


Figure 6 – Comparison between the quality of papers of students in the Experimental and Control groups for children in the teams levels Low, Medium, and High.

Also, Figure 7 shows that the improvement in number of questions asked, of children in the Experimental group with respect to children in the control group, is not limited to children in the high level ability teams, but distributed across the various levels of ability.

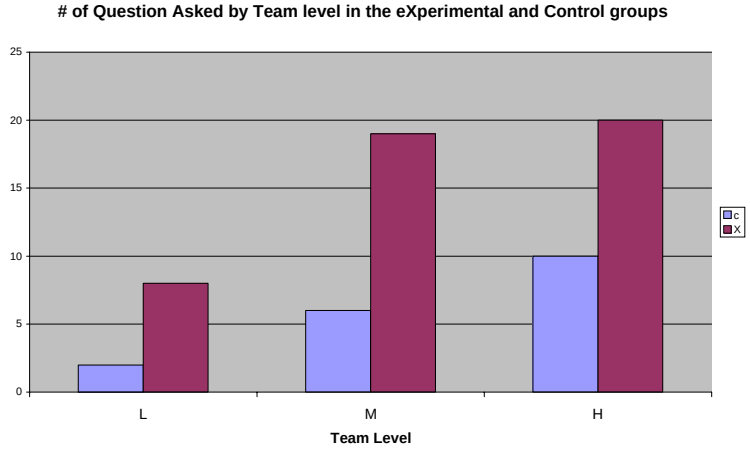


Figure 7 - number of questions asked by team level in the two groups (X and C).

Another important element indicating how well the system supports children across a wide variety of different abilities is described by the distribution of interventions amongst ability levels. The three following graphs show that Meta-Cognitive interventions were mostly received by children in the High performance teams and the children who received them did particularly well in the paper (Figure 8); that Cognitive interventions were mostly received by children in the Low and High performance teams and the children who received them did particularly poorly in the paper (Figure 9); and that Motivational interventions were mostly received by children in the Low performance teams and the children who received them did particularly poorly in the paper (Figure 10).

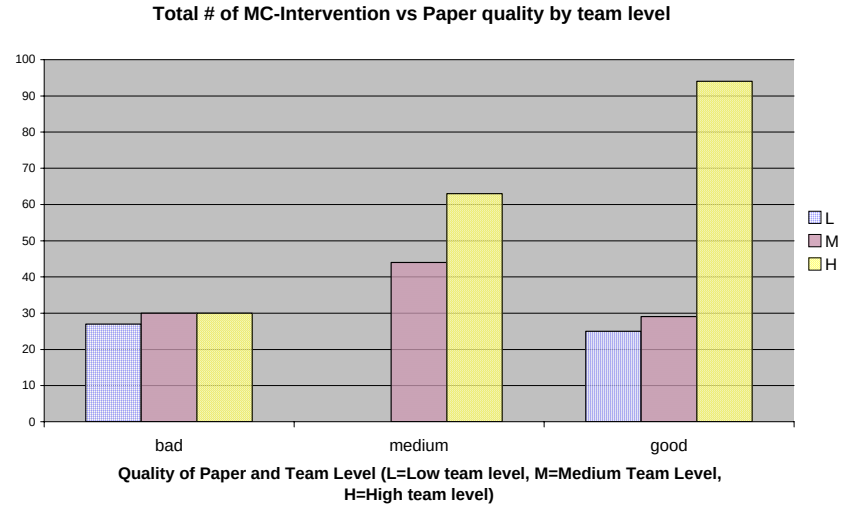


Figure 8 – Distribution of Meta-Cognitive interventions between different team levels with indication of results in paper

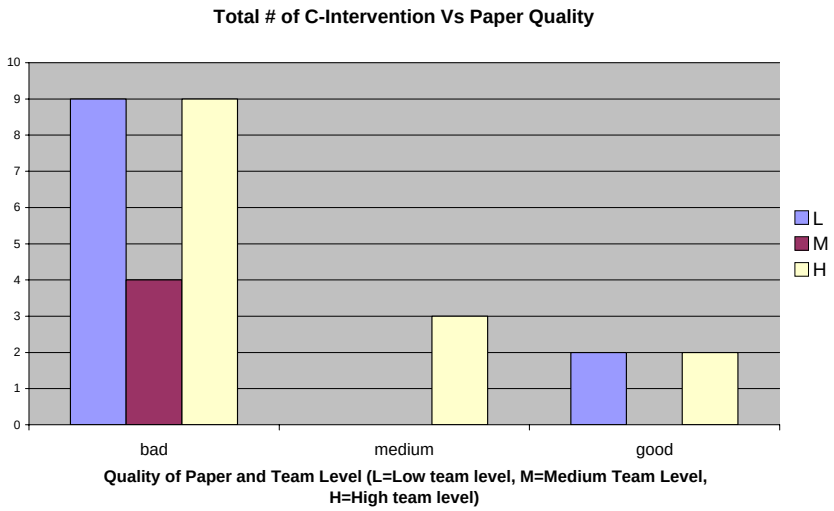


Figure 9 – Distribution of Cognitive interventions between different paper results with indication of the team levels.

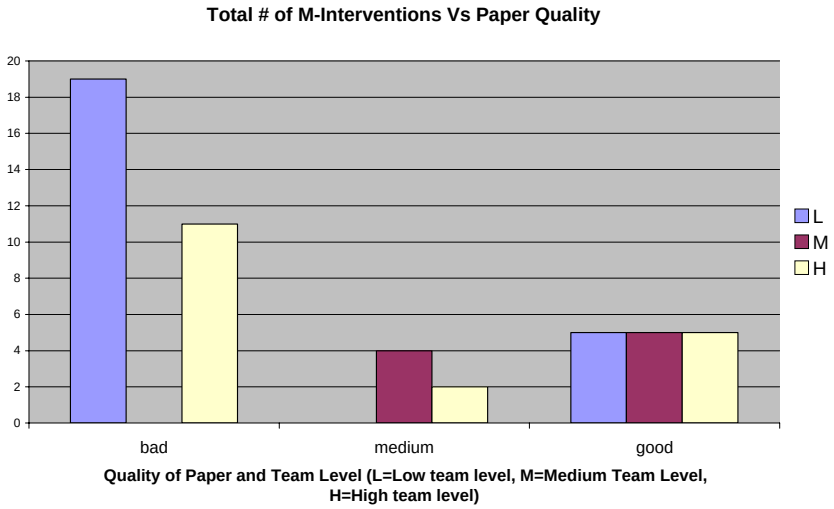


Figure 10 – Distribution of Motivational interventions between different paper results with indication of the team levels.

These results seem to confirm that, at least with respect to the paper task, whilst and improvement in learning results can be observed across the team levels, there is space for improvement with respect to the effectiveness of Cognitive and Motivational interventions.

5. Community effects on learning

It is always the case that the learning process is somehow influenced by the environment in which it takes place, and in particular by the social aspects of the environment. This is reflected in table 23 that analyses the effect that *belonging to a specific class* has on the performance of the learning tasks. Basically the quality of every task in the learning process is significantly affected by the class in which the learning takes place. Table 23 shows that there is a significant *class* effect on the number of questions asked to the expert (P=0.001), having a good grade in the introduction (P=0.02), having many paragraphs in the paper (P=0.0015) and quality of paper (p=0.003) in the total sample..

Independent Variable	Chi Square Test: Gamma (ASE) (P-value)
Questions_asked	0.005 (0.15) (0.001)
Status	0.21 (0.18) (0.09)
intro	-0.39 (0.12) (0.02)
Good_goal	0.008 (0.92) (0.22)
cc	-0.14 (0.11) (0.12)
paraphase	0.28 (0.14) (0.015)
Qualityof_paper	0.44 (0.13) (0.003)
Questions	0.37 (0.13) (0.13)

Table 23 - Analysis of class effect on the quality of each of the learning tasks for the whole Sample

This interaction between the environment and the learning process is also reflected in the Experimental sample as shown in table 24 where are highlighted the *class* effects on the number of questions asked to the expert (P=0.003) and on the quality of the paper (P=0.01) in the experimental sample.

Independent Variable	Chi SquareTest: gamma (ase) (P-value)
Questions_asked	-0.20 (0.19) (0.003)
Status	0.13 (0.25) (0.09)
intro	-
Good_goal	0.23 (0.32) (0.42)
cc	-
paraphase	-
Qualityof_paper	0.51 (0.17)(0.01)
Questions	-

Table 24 - Analysis of class effect on the quality of each of the learning tasks for the Experimental Sample

However, as shown in table 25, the results obtained in the Control sample present a significant correlation to the *class* choice only in relation to finishing the questionnaire (status) (P=0.053).

Independent Variable	Chi Square Test: Gamma (ASE) (P-value)
Questions_asked	-
Status	0.27 (0.25) (0.053)
intro	-
Good_goal	-0.12 (0.28) 0.50
cc	-
paraphase	-
Qualityof_paper	0.39 (0.19)(0.51)
Questions	-

Table 25 - Analysis of class effect on the quality of each of the learning tasks for the Control Sample

This seems to indicate the role of the teacher in combination with the attention management system is more effective than without the attention management system. The strong class effect highlighted in table 23 however could also indicate that other results that we have obtained might have been significantly influenced by the class environment and the teacher in particular.

In order to verify if this was the case we have performed a Logistic regression with Teacher, Team_Level and all the independent variables (see table 26). We found no evidence of a teacher/class effect on the performance of the Experimental and Control group. Again there's borderline significance for questions_asked and the quality of paper.

Independent Variable	Parameter Estimates	P-value
Intercept	-5.03	0.41
Teacher	-0.01	0.97
Team_Level M	-1.47	0.18
Team_Level S	-1.05	0.32
Questions_asked	0.45	0.07
Status	0.95	0.42
intro	-0.09	0.66
Good_goal	1.70	0.06
cc	-0.01	0.93
# of paragraphs	-0.11	0.67
Qualityof_paper	0.92	0.08
Questions	-0.12	0.33

Table 26 - Logistic regression with Teacher, Team_Level and all the independent variables.

6. Attention

6.1. Evaluating the attention indicator

This indicator is probably the most difficult one to evaluate and unfortunately, we will not be able to report on all aspects of the analysis because data is still being analyzed as we write. Furthermore, temporal aspects are particular important in this analysis and extracting detailed temporal data from the log has proven more complex than expected.

Overall we plan to look at three aspects of attention allocation: task fragmentation, task sequencing, and time on task. In this report we briefly report our findings on task fragmentation.

6.2. Task fragmentation

We have initiated an analysis of task fragmentation, which we measure by looking at the number of times children started each task. Our initial hypothesis was that the system would have reduced task fragmentation but in fact, the charts in Figure 11 show that children in the Experimental group have interrupted and restarted tasks somehow more frequently than children in the Control group. It should be noted that however, the average total task fragmentation between the two groups was not significantly different, with an average of 149.33 Start tasks for the Control group, and 150.57 for the Experimental group. Also, table 27 shows that the only statistically significant correlations are relative to the Diary task (where children in the Control group fragmented the task more than children in the Experimental group), and the Forum task (where children in the Experimental group fragmented the task more than children in the Control group).

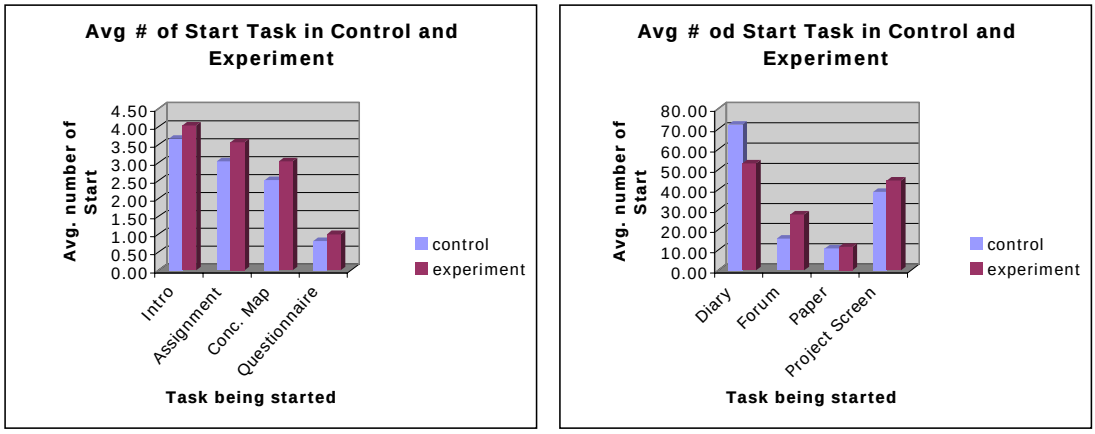


Figure 11 – On average, children in the Experimental group have (re)started more frequently all tasks (with the exception of the Introduction and the Diary tasks) than children in the Control group

Variable	Parameter Estimate	p-value
Intercept	1.0422	0.3530
Total_start_persinfo	-0.0297	0.7549
Total_start_assignmenttarget	0.1366	0.4285
Total_start_conceptmap	-0.0826	0.6810

Total_start_diary	-0.0246	0.0291
Total_start_forum	0.0413	0.0419
Total_start_paper	-0.0324	0.6138
Total_start_question	-0.5782	0.0794
Total_start_projectmanager	0.0183	0.4763

Table 27 - Logistic Regression of the Start Times:

This data can be interpreted in two ways. The first explanation is that the system had no impact on task fragmentation. The second explanation is that children in the Experimental group, after a first period working with the system, realised that the system would have helped them returning to the main task through Meta-Cognitive interventions and for this reason they may have felt more inclined to move freely between tasks. In order to verify this second hypothesis we are currently looking at the times when children started the tasks.